LLM-Driven Descriptive Analysis of Metaphase Cell Images Containing ecDNA

Mohit Sridhar
Halıcıoğlu Data Science Institute
University of California, San Diego
msridhar@ucsd.edu

Archit Pimple
Halıcıoğlu Data Science Institute
University of California, San Diego
apimple@ucsd.edu

Andrew Yin
Halıcıoğlu Data Science Institute
University of California, San Diego
anyin@ucsd.edu

Utkrisht Rajkumar
Amazon Inc.
utkrisht96@gmail.com

Thiago Mosquiero
Amazon Inc.
tmosquiero@gmail.com

1 Abstract

Extrachromasomal DNA (ecDNA) is a key factor in tumor evolution and resistance to treatment, making its precise identification crucial for cancer diagnostics and prognosis. In our study, we explore the potential of large language models (LLMs) to replace or supplement pathologists in ecDNA identification using DAPI-stained cell images. We integrate MiniCPM, Qwen, and Pixtral models into our workflow, leveraging their multimodal capabilities to process and interpret fluorescence microscopy data. Our approach involves N-shot learning and multi-layered prompts to refine model responses and enhance diagnostic accuracy. We evaluate the effectiveness of these models in recognizing ecDNA patterns and assessing their performance against expert pathological annotations. Our findings provide insight into the feasibility of LLM-assisted pathology and highlight the challenges and advantages of using AI-driven approaches in medical imaging.

2 Introduction

Our project focuses on analyzing and understanding extra chromosomal DNA (ecDNA). ecDNA exists outside the chromosomes in a cell. Unlike typical chromosomal DNA, which is organized within the chromosomes in the nucleus, ecDNA is typically found as circular pieces of DNA. It can exist in various forms, such as plasmids in bacteria or circular DNA in cancer cells.

In humans, ecDNA is often associated with cancer because it can carry genes that promote tumor growth, like oncogenes. Since it is separate from chromosomes, ecDNA can replicate independently, leading to an increased copy number of certain genes, which can make cancer more aggressive. Researchers study ecDNA to understand its role in genetic diseases, cancer progression, and drug resistance.

There aren't many tools available for analyzing images with ecDNA in detail. Current tools mostly focus on identifying ecDNA and doing basic measurements, but they don't give information about the structure of ecDNA or analyze metaphase spreads to detect chromosomal abnormalities or genetic diseases. This makes it hard for researchers to fully understand their data and the role of ecDNA.

There are also models that generate text descriptions for general images, but they haven't been specifically applied to biological images, creating an opportunity for new developments. Our solution aims to use large language models (LLMs) as virtual pathologists. These models will be informed by tools like ecSeg to provide more meaningful and detailed insights which will assist researchers.

2.1 Problem Statement

There is a lack of specialized tools used to conduct descriptive analysis on metaphase images containing extra chromosomal DNA (ecDNA). Current tools work on identification of ecDNA and quantitative analysis, however there is no insight provided into ecDNA structure

or metaphase spreads in order to identify chromosomal abnormalities or presence of genetic diseases. This limits the ability of researchers to make meaning out of their data and properly understand the role of ecDNA.

One existing solution on such analysis is ecSeg, which performs semantic segmentation on metaphase images to quantify ecDNA. However, one limitation of this is that there is no information provided on cell state in order to prepare these images for any sort of diagnosis by a pathologist.

There are also text captioning models implemented for general analysis of images, but they have not been explicitly applied on bio-related topics and provide a new avenue for exploration.

2.2 Output Format

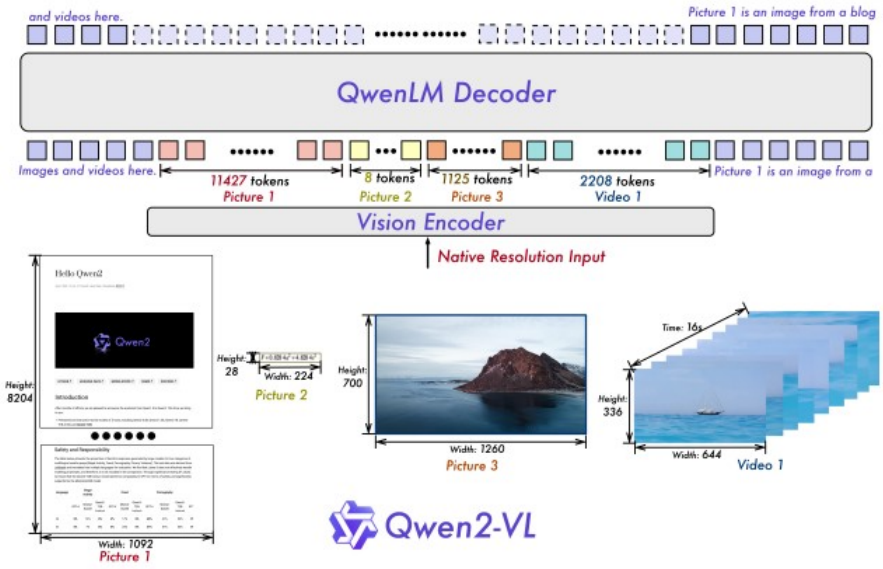
Our primary output is going to be represented by a count of ecDNA generated by the prompt-engineered LLMs on various metaphase images including ecDNA, and the results will be compared to the ground truth counts collected from the connected component analysis performed on the cell images from the ecSeg repository. Our findings will be presented in a paper along with a demo of these LLMs where example images will be passed in as input. We will communicate our findings through the use of scoring metrics measuring the performance of the LLMs such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

3 Models

3.1 Qwen2-VL

Qwen2-VL is a multimodal model used for visual processing. It can take in various inputs of images and videos and provide context surrounding general chat, video understanding, grounding, multilingual OCR, live chat, UI interaction, and more. It has three currently available models with parameters of 1.5 billion (Qwen2-VL-2B), 7.6 billion (Qwen2-VL-7B), and 72 billion (Qwen-2-VL-72B).

Qwen2-VL's architecture consists of the Vision Transformer (ViT), with approximately 675 million parameters. The model makes use of image resolution enhancement methods such as Naive Dynamic Resolution, which involves 2D Rotary Position Embedding (RoPE), which replaces absolute position embeddings of ViT and allows models to capture details across various image resolutions.



Naive Dynamic Resolution works by condensing multiple image resolutions into a single sequence to reduce GPU memory usage. Next, an MLP Layer is used to compress the temporary 2 by 2 tokens into single tokens.

Another method used to enhance model architecture is M-RoPE, or Multimodal Rotary Position Embedding (M-RoPE). This method models the positional features of multimodal inputs. As in our case, the original rotary embeddings are divided into three features: temporal, height, and width. The temporal IDs remain constant for images while unique IDs are assigned to the height and width based on a token's position in the image. This allows the model to extrapolate to longer sequences for longer inference due to the reduction in value of position IDs.

The training of this model is a three step process:

Step 1: Model training is initially focused on the ViT component on image-text pairs in order to improve the semantic understanding of the model. The Vision Transformer is trained on various tasks including OCR, image-text articles, video understanding and more, exposed to a training corpus of about 600 billion tokens.

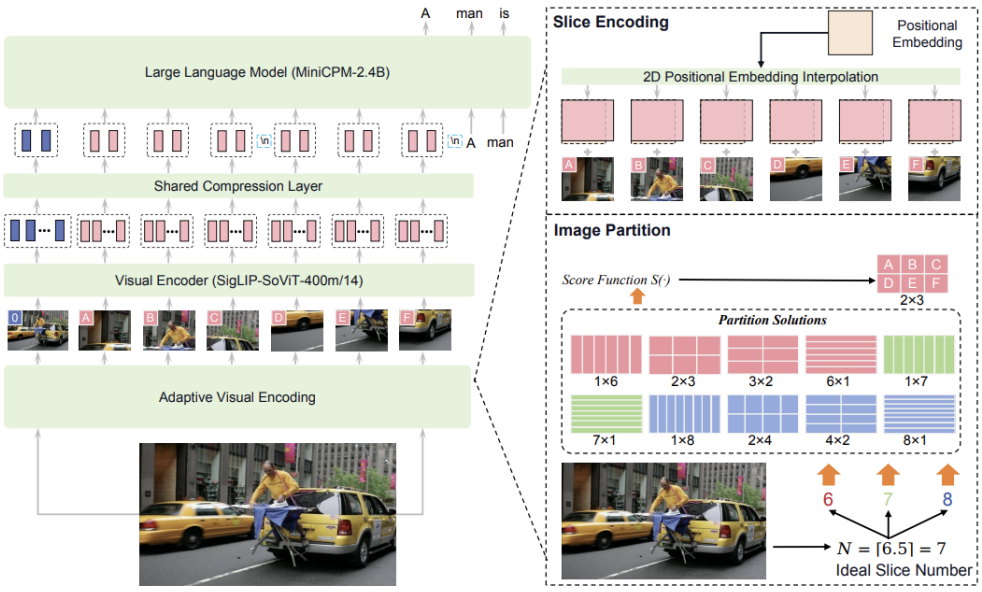
Step 2: All parameters are unfreezed and the model is trained with a more diverse range of data to improve comprehensive understanding.More image-related datasets are introduced in order to help the model gain more competency in visual question answering tasks and guides the model towards multitasking capabilities. In total, the model is introduced to almost 1.4 trillion tokens of training.

Step 3: The model is fine-tuned on instructional datasets using the ChatML format is used to help the model understand instruction-following data. This includes multimodal components such as document parsing, video comprehension, agent-based instructions, and more. This allows the model to understand complex multimodal tasks in addition to text-based interactions.

3.2 MiniCPMv

MiniCPMv is a large language model designed to integrate seamlessly with vision models for enhanced multimodal understanding. It extends the capabilities of traditional language models by incorporating image processing directly into the language modeling process. MiniCPMv has been utilized in various tasks that require a sophisticated blend of textual and visual information processing.

The architecture of MiniCPMv consists of 3 major components. The first is a Visual Encoder which transforms raw image data into a compact representation that captures the essential visual features. This encoder is crucial for the model's ability to understand and process visual content in conjunction with textual data. The next is a compression layer that that further refines the visual tokens. This layer uses cross-attention mechanisms to distill the most relevant information from the visual encoder, optimizing it for subsequent processing by the language model. The core of MiniCPMv is its language model, which takes the compressed visual tokens and integrates them with textual input. This integration allows the model to generate contextually relevant text based on both the text and image inputs.



(a) overall structure (b) adaptive visual encoding

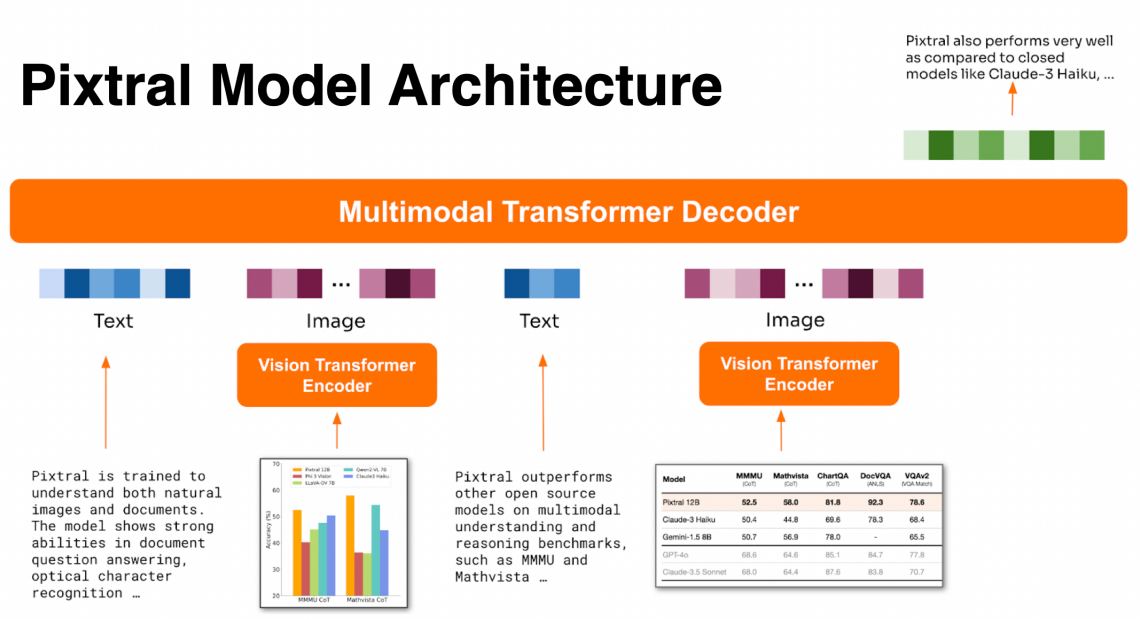
MiniCPMv is trained using a comprehensive array of datasets designed to enhance its multimodal capabilities. Part-1 of its training data focuses on fundamental recognition and response skills using datasets like Flickr-30K for image captioning and various VQA datasets for visual question answering. Part-2 extends these capabilities by integrating more complex datasets such as SVIT and UniMM-Chat, which are tailored for instruction-based tasks and sophisticated text-only interactions. This rigorous training regimen enables MiniCPMv to excel in tasks that require a deep understanding of both textual and visual inputs, making it adept at processing and responding to complex multimodal scenarios. The combination of these diverse training datasets ensures that MiniCPMv not only understands visual and textual data but can also gen-

erate contextually relevant responses in a variety of applications, bridging the gap between traditional language models and advanced image processing techniques.

3.3 Pixtral 12B

Pixtral 12B is a multimodal AI model developed by Mistral AI, designed to proficiently handle both image and text inputs. Building upon the Mistral Nemo 12B architecture, Pixtral 12B integrates a newly developed 400-million-parameter vision encoder, enabling it to process visual data alongside textual information. The original research paper can be found here. (1)

The Pixtral 12B model consists of a 12B parameter multimodal decoder and a 400M parameter vision encoder, designed to interpret and process visual inputs. Additionally, specialized cross-attention layers are incorporated to facilitate seamless integration and interaction between textual and visual data, enhancing the model's multimodal understanding. The model also boasts a context window of 128,000 tokens, allowing it to process up to 30 high-resolution images per input or the equivalent of a 300-page book.



Trained with interleaved image and text data, Pixtral 12B excels in tasks that require understanding and reasoning across both modalities. It demonstrates strong performance in areas such as chart and figure interpretation, document question answering, and general multimodal reasoning. The model supports variable image sizes and aspect ratios, and processes multiple images with a 128,000 token context window, providing users with flexibility in the number of tokens allocated for image processing. Despite its advanced multimodal capabilities, Pixtral 12B achievs state-of-the-art performance on text-only benchmarks, excelling in instruction following, coding, and mathematical tasks without compromising its textual understanding.

The feedforward network (FFN) within the attention blocks is modified by introducing gating in the hidden layer, replacing the standard FFN structure to enhance information flow and

improve learning efficiency. To optimize batch processing, sequence packing is used by flattening images along the sequence dimension and concatenating them, while a block-diagonal attention mask ensures that patches from different images do not interfere with each other.

Finally, instead of traditional learned or absolute position embeddings, the model utilizes RoPE-2D (Rotary Position Encodings in 2D) within self-attention layers. This approach allows for a more natural handling of variable image sizes without requiring interpolation, which can otherwise degrade performance. These enhancements collectively enable the vision decoder to efficiently handle complex multimodal inputs while preserving spatial relationships across images.

4 Methods

We began by collecting our dataset from the ecSeg repository, which contains high-resolution DAPI-stained cell images annotated for extrachromosomal DNA (ecDNA) detection. Using this dataset, we applied various machine learning techniques to optimize the performance of MiniCPM, Qwen, and PixTral models in analyzing ecDNA patterns. Our approach involved the following techniques:

4.1 N-Shot Learning

N-shot learning allows models to generalize from a limited set of examples. We experimented with different values of N to determine the optimal number of examples needed for effective ecDNA recognition. By providing a few labeled DAPI-stained cell images along with their corresponding ecDNA annotations, we aimed to improve model accuracy and reduce the need for extensive labeled datasets.

4.2 Multi-Layered Prompting

We designed a multi-layered prompting strategy to refine model responses and guide them through a structured reasoning process. Instead of a single-step input-output approach, we broke down the analysis into multiple stages, such as: Step 1: Identifying cell structures and segmenting the nucleus. Step 2: Recognizing ecDNA patterns within the segmented regions. Step 3: Validating and refining predictions based on confidence scores. This hierarchical approach helped improve the interpretability and robustness of the model's predictions.

4.3 Temperature Adjustment

To control the randomness and certainty of the model's predictions, we fine-tuned the temperature parameter during inference. A lower temperature (e.g., 0.2) encouraged more deterministic outputs, making the model confident in its responses, while a higher temperature (e.g.,

0.8) allowed for more variability and exploration of different interpretations. By adjusting the temperature dynamically, we balanced precision and diversity in model-generated insights.

4.4 Aggregating Results Over Multiple Trials

Since single-run predictions may be inconsistent or subject to noise, we aggregated results over multiple trials to enhance reliability. Each model was run several times on the same input data, and the final prediction was determined by majority voting or averaging confidence scores. This approach reduced the impact of outliers and ensured a more stable and accurate detection of ecDNA.

By combining these techniques, we aim to improve the effectiveness of LLMs in supplementing or replacing pathologists for ecDNA analysis. Our methodology provides a structured and adaptive framework for using AI in biomedical imaging tasks.

4.5 Context Messages

In order to modify the behavior of the models, different context or system messages were provided to the models describing the way in which the model should assess the images. These descriptions consisted of acting as a pathologist, an AI agent, and a researcher.

Using the baseline prompt, we aimed to utilize these messages to understand the behavior of the multimodal models and understand the effectiveness of this approach towards assisting these models in gaining a more comprehensive understanding of the cell images.

5 Results

5.1 Qwen2-VL

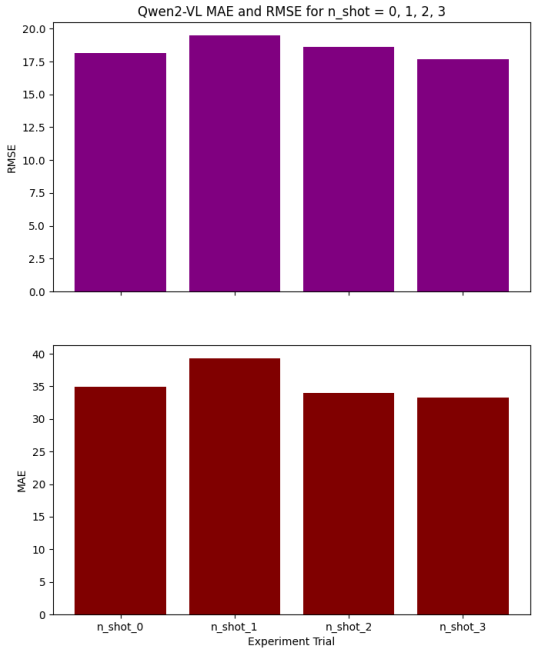


Figure 1: N-Shot Learning Experiment MSE and RMSE

As shown in Figure 1, the 3-Shot experiment yielded the lowest MAE and RMSE score, and the 1-Shot experiment resulted in the highest MAE and RMSE scores. This implies that providing Qwen2-VL with context on train images paired with counts allows the model to gain a better understanding of the task. In the 3-shot experiment, Qwen2-VL has exposure to images with 0, 1, and many counts of ecDNA, covering three possible cases of counts in these cell images.

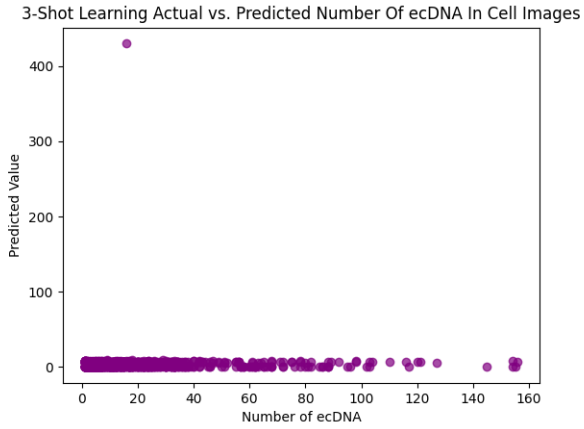


Figure 2: 3-Shot Learning Experiment Correlation Between Actual and Predicted ecDNA Counts

In contrast, Figure 2 illustrates that there is no sort of correlation between the predicted and actual number of ecDNA. he predictions made by Qwen2-VL are in the range of 0 to 20, with one outlier of about 40,000 ecDNA, implying that the Qwen2-VL predictions are based on the ranges provided during the learning process, and the model lacks the ability to onduct additional in-depth analysis. This indicates that Qwen2-VL requires more fine-tuning on the metaphase cell images in order to comprehend the task of counting these ecDNA structures.

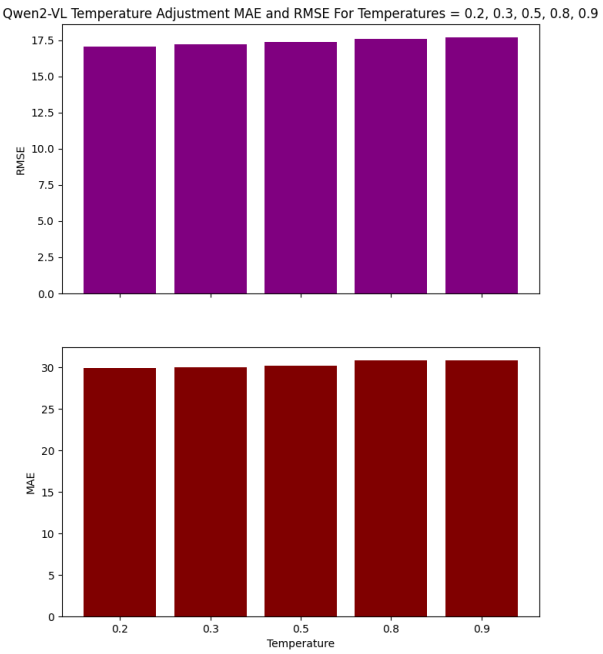


Figure 3: Temperature Adjustment Experiment MSE and RMSE

When tested on various temperature values, the temperature value of 0.2 resulted in the lowest MAE and RMSE, which performs better than all other temperature-value experiments, including the baseline MAE and RMSE recorded in the 0-shot experiment referenced in Figure 1 (first bar from the left). The initial implication is that making Qwen2-VL more deterministic results in better model performance.

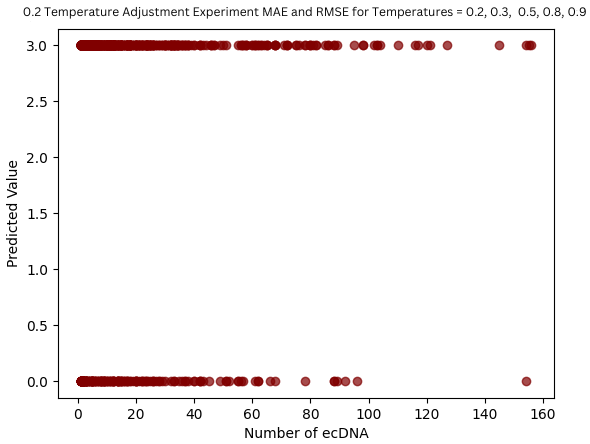


Figure 4: 0.2 Temperature Adjustment Experiment Correlation Between Actual and Predicted ecDNA Counts

There is a similar situation as with the N-shot learning experiment, where we failed to find any correlation between the actual and predicted counts of ecDNA. We can see in Figure 4 that providing Qwen2-VL with a baseline prompt that includes context and instructions on the task of finding the count results in an almost binary prediction, where the counts are either 0 or 3. This further implies that fine-tuning is required for this model, particularly on labeled metaphase cell images paired with their respective counts.

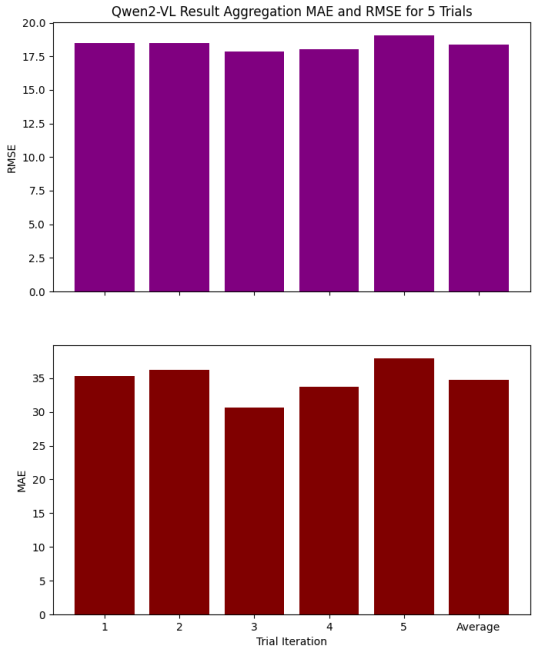


Figure 5: Result Aggregation Experiment MAE and RMSE

In the result aggregation experiment, we can observe that the MAE and RMSE metrics are generally lower than those of the individual trails. Figure 5 shows that the third trial of the same prompt resulted in a lower MSE and RMSE compared to the other trials. It also shows that averaging multiple iterations does not necessarily indicate an improvement in performance, and that their is not much variability produced by the model after multiple attempts at a default temperature value of 0.7.

Table 1: Performance metrics (MAE and RMSE) for Qwen2-VL on various experimental setups.

Experiment	**MAE**	**RMSE**
Baseline	34.94	18.12
3-shot Learning	33.27	17.69
Multi-Layer Prompting	37.03	18.77
0.2 Temperature	29.93	17.05
Result Aggregation	34.80	18.38

Table 1 shows the MAE and RMSE metrics for each experiment, where the best metrics are selected from multiple trials. The "**0.2 Temperature**" experiment had the best performing metric for MAE, and also the lowest RMSE metric. This implies that lowering the temperature provides better performance by making the model more deterministic when making its predictions.

The experiment with the highest MAE and RMSE metrics was the "**Multi-Layer Prompting**" experiment. This indicates that creating a chain of thought by asking the model multiple context questions leads the model to lose the original goal which is asked after prompting the descriptive questions.

Overall, Qwen2-VL was the least performant model out of the three models that were tested. While metrics indicate that the model has decent performance before fine-tuning, correlation assessment in each experiment indicates that the model has not yet grasped the concept of counting in such images, and that further training is required to yield improved results and contextual understanding for this model.

5.2 MiniCPMv2

MiniCPM stood as the smallest model and the second best performing one. The biggest issues with MiniCPM remained it's inability to generalize when given specific prompts.

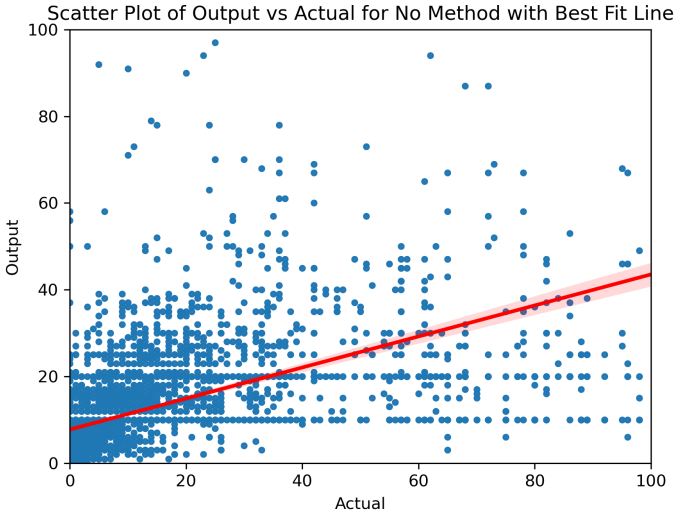


Figure 6: Scatter plot of predicted vs. actual ecDNA counts with no method

From the above figure we see that MiniCPM out of the box does a decent job at understanding its task. While not perfect, it understands some correlation between higher ecDNA vs lower ecDNA counts up to a certain threshold.

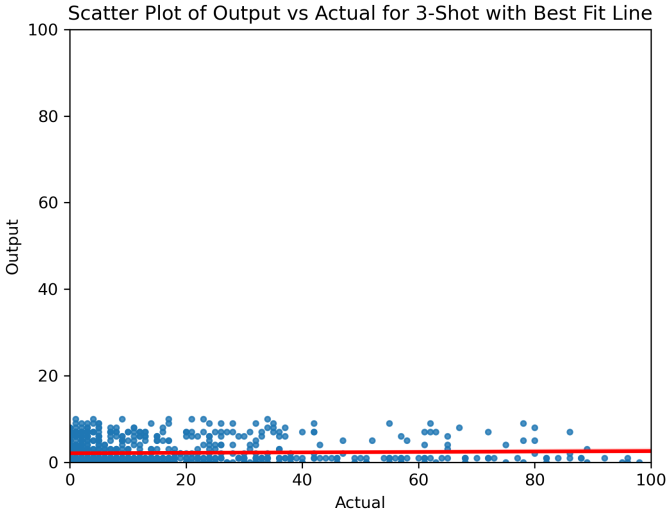


Figure 7: Scatter plot of predicted vs. actual ecDNA counts with 3-shot learning

When introducing N-Shot learning, minicpm fails to generalize and continually picks from the n-samples it is learning from, making certain prompt engineering techniques difficult. This also applied to chain of thought as the model ultimately chose from a select few data points when given too much information.

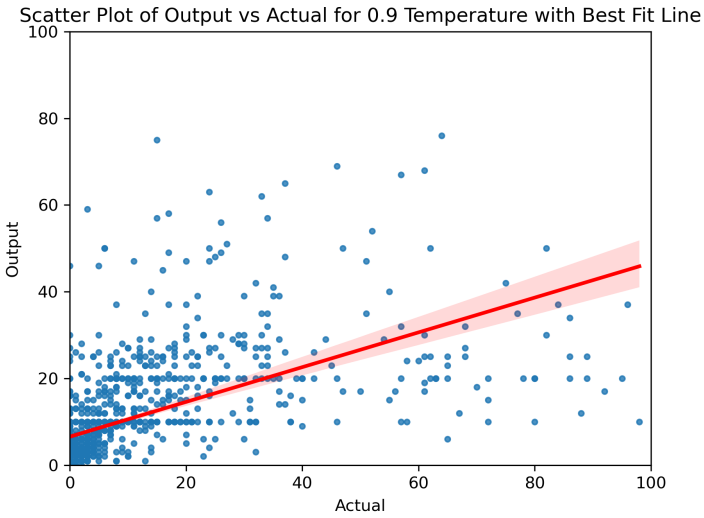


Figure 8: Scatter plot of predicted vs. actual ecDNA counts with 0.9 temperature

When testing for different temperatures however, certain quantities helped the model perform better. The best of our tests was a temperature of 0.9, implying that a higher temperature ultimately correlates with a more accurate MiniCPM.

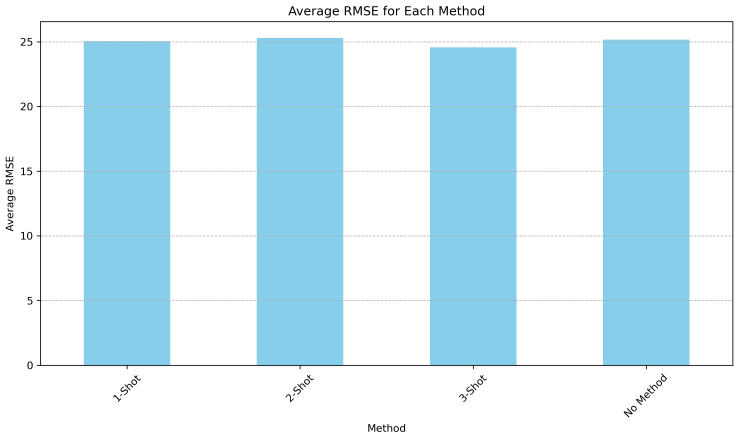


Figure 9: Scatter plot of predicted vs. actual ecDNA counts with 0.9 temperature

When evaluating the RMSE of the predictions between different methods, we notice that between n-shot and no method, the rmse remains similar. This can be attributed to the average disparity of a random guess being similar to the variation of an accurate prediction. Further experiments were conducted using different contexts, but these all proved inferior to temperature changes. The most plausible reason for this is the small size of MiniCPM not allowing it to generalize from new information and instead cause it to fixate on any context given.

Table 2: Performance metrics (MAE and RMSE) for MiniCPMv2 on various setups.

Experiment	**MAE**	**RMSE**
Baseline	10.44	24.67
3-shot Learning	12.21	24.57
Multi-Layer Prompting	154.46	265.65
0.2 Temperature	9.80	22.33
Result Aggregation	10.58	25.17

By confining our results into the form of MAE and RMSE, we can see that a temperature adjustment ultimately makes for the best model. Multi-Layer Prompting and Chaining lead to the worst result by a wide margin due to MiniCPM's tendency towards extreme outliers when given too much context. In the case of N-Shot learning the values ultimately end up being selected from just the values given as inputs. Result aggregation while theoretically promising, ultimately gives less accurate results simply due to the nature of an average prediction being farther off from an actual prediction.

5.3 Pixtral 12B

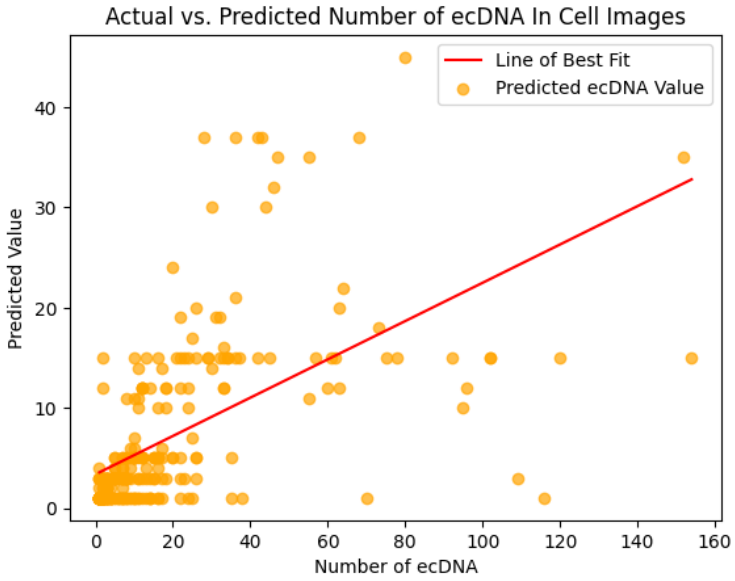


Figure 10: Scatter plot of predicted vs. actual ecDNA counts.

As seen in Figure 10, we can see that Pixtral-12B has better performance compared to the other models when tasked with predicting ecDNA counts in metaphase images. The line of best fit shows that there is a strong, positive correlation between actual and predicted values.

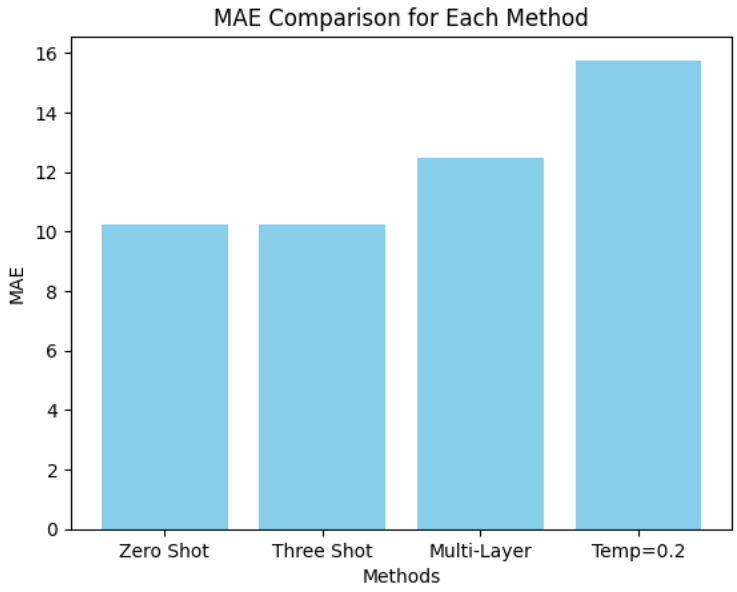


Figure 11: MAE across the various methods tested.

Figure 11 shows that setting the temperature to a value of $T = 0.2$ results in the highest mean absolute error. Since our task requires a single numerical answer to be provided, this is

unexpected because we would want the LLM to be deterministic. However, the model's pre-training and default hyperparameters may be tuned for the best performance, especially for tasks such as this.

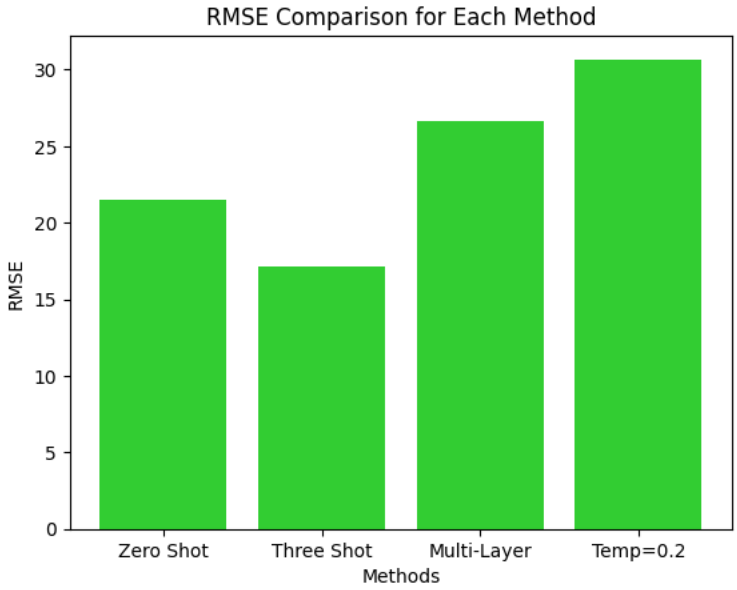


Figure 12: RMSE across the various methods tested.

Figure 12 shows that setting the temperature to a value of $T = 0.2$ also results in the highest root mean squared error (RMSE). This was a surprising result as well because making the model more deterministic should increase performance for this task - which asks for a single numerical answer. However, it seems that changing the model's temperature from the default of $T = 0.7$ actually made it worse. Additionally, the three-shot prompting method had a much better RMSE compared to its MAE.

Table 3: Performance metrics (MAE and RMSE) for Pixtral-12B across various methods.

Experiment	**MAE**	**RMSE**
Baseline	10.71	18.64
3-shot Learning	12.97	23.43
Multi-Layer Prompting	12.37	22.36
0.2 Temperature	12.67	25.30
Result Aggregation	11.43	22.43

As we can see in Table 3, the baseline experiment using zero-shot prompting has the best MAE and RMSE. Although this is a different result than the other models above, Pixtral-12B still outperforms them in the MAE category. However, it has similar values for the RMSE. This implies that Pixtral's default hyper-parameters and pre-training are much better suited to understand complex tasks. The experiments with the highest MAE and RMSE were the "**0.2**

Temperature" followed by "**Multi-Layer Prompting**", which may cause the model to lose sight of the original goal by introducing too much context/complexity.

6 Discussion

The results show that Pixtral-12B is the best-performing model with respect to MAE. Trailed by MiniCPMv2 and then Qwen2-VL, this difference is likely due to the larger context and parameter size of Pixtral allowing it to handle complex queries and produce accurate answers. In the future, we hope to implement fine-tuning to enhance the accuracy of our best-performing model and reduce dependence on pre-training. Additionally, we hope to eventually use datasets annotated by experts to minimize faulty data.

6.1 Limitations

Several limitations impacted the scope and performance of our project.

Dataset Limitations: Our dataset was derived from the ecSeg repository and generated using segmentation techniques. While this provided a highly accurate approximation of cellular structure counts, it remains an estimate rather than a ground truth. A professionally annotated dataset would likely improve the model's performance by allowing it to learn from nuanced biological details that segmentation algorithms may overlook.

Model Behavior and Methodological Constraints: Some of our modeling approaches introduced challenges that limited overall effectiveness. In the case of N-shot learning, models frequently defaulted to copying counts from the provided examples rather than making independent inferences. Additionally, when incorporating contextual information, certain background variations caused the models to fail, often responding with statements such as "this task is too complex." These issues suggest that more refined prompt engineering, better task framing, or the use of more advanced models could significantly improve performance.

Computational Constraints: Our most significant limitation was computational power. With only three personal GPUs available for running three models, training became time-intensive—each experiment often taking several hours. This severely restricted the number of experiments we could conduct and limited our ability to iterate and optimize. Future work would benefit from access to high-performance computing resources, enabling faster experimentation and more comprehensive evaluation.

7 Contributions

In this project, each of our three team members—Archit, Mohit, and Andrew—took responsibility for benchmarking a different model. Archit worked on Qwen2-VL, Mohit focused on Pixtral 12B, and Andrew researched MiniCPMv. This division of work allowed us to explore various LLMs and determine which one was best suited for our task. Moving forward, we

will experiment with fine-tuning the selected model to further enhance its performance and accuracy.

7.1 Archit Pimple

Worked on using the Qwen2-VL model to create descriptions of images and the various structures present. Set up the ecSeg repository in order to generate ground truth images to compare results against and ultimately use these images to calculate performance metrics listed in the "Results" section. Conducted all experiments and implemented chain of thought and N-shot conversations to generate error metrics.

7.2 Mohit Sridhar

Researched and applied the Pixtral-12B model to analyze and generate textual descriptions of metaphase images containing ecDNA, chromosomes, and nuclei. Set up the GitHub website and repository. Worked on a simple GUI to show how images were segmented and passed into the LLMs. Performed various experiments and collected metrics to gauge Pixtral-12B's relative performance.

7.3 Andrew Yin

Worked on MiniCPMv model to count and analyze ecDNA imagery. Compiled prompt engineer methods for evaluating the 3 models. Developed standardized n-shot prompting pipeline in addition to multi-layered prompts for a standardized model. Aggregated accuracy over multiple runs into csv for data analysis. Developed visualizations to determine model accuracy and performance.

References

[1] Agrawal, P., Antoniak, S., Hanna, E., Bout, B., Chaplot, D., Chudnovsky, J., Costa, D., Monicault, B., Garg, S., Gervet, T., Ghosh, S., Héliou, A., Jacob, P., Jiang, A., Khandelwal, K., Lacroix, T., Lample, G., Casas, D., Lavril, T., Scao, T., Lo, A., Marshall, W., Martin, L., Mensch, A., Muddireddy, P., Nemychnikova, V., Pellat, M., Platen, P., Raghuraman, N., Rozière, B., Sablayrolles, A., Saulnier, L., Sauvestre, R., Shang, W., Soletskyi, R., Stewart, L., Stock, P., Studnia, J., Subramanian, S., Vaze, S., Wang, T. & Yang, S. Pixtral 12B. (2024), https://arxiv.org/abs/2410.07073

[2] Rajkumar, U. et al. ecSeg: Semantic Segmentation of Metaphase Images containing Extrachromosomal DNA. iScience. 21, 428-435. (2019)

8 Appendix

1. https://drive.google.com/file/d/1oKCRb0GxTXocFfpkQ50u0Hjym1LJi_rw/view?usp=drive_link

2. https://drive.google.com/file/d/1pXLSZZnu6LVZeLmYZd-A7sv7lPAR7d_g/view?usp=sharing