IPA: Inference Pipeline Adaptation to achieve high accuracy, cost-efficiency, and power-efficiency

Misagh Soltani, Sabah Anis, Xeerak Muhammad

Abstract:
Modern ML inference pipelines face the dual challenge of achieving high performance and minimizing energy consumption. Current systems, including IPA, lack the ability to monitor or optimize energy usage, making them unsuitable for real-world applications that demand both efficiency and sustainability. This project attempts to address this limitation by integrating energy profiling into IPA, enabling it to adapt model selection and resource allocation based on real-time energy usage.

Introduction:
The rapid adoption of machine learning (ML) across diverse domains such as autonomous vehicles, healthcare diagnostics, and recommendation systems has brought significant improvements to these fields. These applications often rely on complex ML inference pipelines that demand high accuracy, low latency, and scalability. However, the cost of deploying these models at scale, both in terms of operational expenditure and environmental impact, has become a critical concern.

Current inference systems primarily optimize for accuracy, latency, and cost, but energy consumption, which is a crucial factor for sustainability, is often overlooked. This gap in optimization hinders efforts toward achieving sustainable AI, especially as the carbon footprint of ML systems becomes increasingly evident. The development of energy-aware inference systems is thus essential for aligning the growth of AI with environmental goals.

While existing frameworks like InferLine [3], Loki [4], AutoInfer [5], and Swayam [6] have made strides in optimizing ML inference pipelines for latency and cost, they lack explicit integration of energy efficiency into their optimization objectives. Other studies, such as those by Harty et al. [7], have explored energy-efficient practices in ML training and inference. Systems like Clover [2] have introduced carbon-conscious runtimes but remain limited in their integration of accuracy, latency, cost, and energy trade-offs into a unified framework.

The Inference Pipeline Adapter (IPA) [1] system provides a foundation for optimizing ML inference pipelines. However, it does not yet incorporate energy consumption as a metric for decision-making. Addressing this limitation, our project, titled IPA: Inference Pipeline Adaptation to achieve high accuracy, cost-efficiency, and power-efficiency, extends the IPA to include energy considerations. By integrating energy metrics into IPA, we aim to operationalize energy-efficient AI practices while maintaining high performance and cost-effectiveness.

Therefore, this project makes the following contributions to the IPA system.

Energy-aware optimization: Extends IPA to use the perf tool to measure energy consumption and incorporate these measurements into IPA's optimization processes.

Benchmarking framework: Designed and executed experiments to benchmark IPA's energy performance under varying workloads, model sizes, and configurations.
Performance Analysis: Analyzed the results of energy and performance benchmarks and provided meaningful insights.

By addressing the critical challenge of energy efficiency, this project contributes to the broader goal of sustainable AI, bridging the gap between performance optimization and environmental responsibility.

Related Work:

The pressing need to incorporate energy consumption and sustainability considerations into ML inference pipelines has recently gained attention, as evidenced by emerging research in carbon-aware and energy-efficient systems. While established frameworks like InferLine [3], Loki [4], AutoInfer [5], and Swayam [6] focus on optimizing inference pipelines primarily for performance and cost, they do not explicitly integrate energy or carbon objectives into their optimization decisions. In contrast, the following works have made significant strides in directly addressing the energy and environmental implications of ML inference, albeit through distinct mechanisms and scopes.

Clover: Carbon-Aware ML Inference Services
Clover [2] represents an early yet robust effort to design a carbon-friendly ML inference service runtime system that jointly considers carbon emissions, accuracy, and performance. Building on the premise that data centers are key contributors to global carbon footprints, Clover adopts an optimization approach that weighs latency, accuracy, and carbon intensity. Its primary innovation lies in combining mixed-quality models—where certain requests are served by computationally lighter, low-quality models—and GPU partitioning to dynamically balance competing objectives. Clover's runtime system can thus opportunistically switch between models of varying complexity to reduce carbon emissions when the energy powering the data center is derived from carbon-intensive sources, without severely compromising accuracy or meeting service-level agreements (SLAs).

By employing real-world carbon intensity traces and production-grade ML models, Clover demonstrates substantial reductions in carbon emissions. However, Clover's approach predominantly centers on the interplay of model quality variants and GPU resource allocation to influence carbon outcomes. While this strategy effectively reveals trade-offs between accuracy, latency, and carbon intensity, it does not generalize easily to other dimensions of inference pipeline optimization, such as dynamic model selection across heterogeneous frameworks or integration with tools that measure and factor energy usage directly into decision-making. Consequently, Clover lays the groundwork for carbon-aware inference but remains limited in its scope, focusing more on carbon intensity than holistic energy consumption metrics.

SPROUT: Sustainable Generative LLM Inference

SPROUT [9] extends the paradigm of sustainable inference to the domain of Generative AI (GenAI), with a particular focus on large language models (LLMs) that have become prevalent and compute-intensive. Recognizing that generative LLM inference often involves iterative token generation, SPROUT introduces “generation directives” to control the autoregressive generation process. By strategically guiding the length and complexity of the generated content, SPROUT reduces the carbon footprint of LLM inference services without severely diminishing output quality. This approach is grounded in the observation that not all user queries necessitate maximal generation complexity, and that adjusting output length and format can yield carbon savings.

Moreover, SPROUT employs a linear programming approach to balance carbon reduction with generation quality, leveraging real-world electricity grid carbon intensity data. Its automatic offline quality assessment mechanism ensures informed decision-making that continuously adapts to changing energy conditions. However, similar to Clover, SPROUT is predominantly tailored to a specific model class (LLMs) and to a particular sustainability lever (manipulating generation complexity). While it offers a powerful technique for carbon-aware inference in generative modeling tasks, its applicability to other model modalities and general inference pipeline architectures is not fully explored.

Our Contribution

Where Clover and SPROUT have made essential progress in introducing sustainability considerations into ML inference, they primarily operate within specialized contexts: Clover focuses on GPU partitioning and mixed-quality models for conventional ML tasks, and SPROUT tackles the unique challenges of generative LLM inference. Both methods emphasize carbon intensity and incorporate strategies to reduce carbon emissions by adjusting model quality or output complexity.

Our work extends this line of research by integrating energy profiling directly into an established inference pipeline adaptation framework (IPA [1]), thereby moving beyond the carbon-intensity lens to encompass a more comprehensive view of energy consumption. Rather than focusing solely on one dimension (e.g., model quality or generation length), we incorporate fine-grained energy metrics gathered through the perf tool into the optimization loop of IPA. This integration allows for dynamic adaptation of model selection, pipeline configuration, and resource allocation grounded in real-time energy usage data. Consequently, our approach complements the specialized strategies of Clover and SPROUT by offering a more generalizable energy-aware inference optimization system that can adapt to various model architectures, workload characteristics, and operational environments.

Overall, while existing studies have broken ground in carbon awareness and energy-conscious inference, our work aims to generalize energy profiling into a flexible, end-to-end inference pipeline optimization framework. By doing so, we strive to address a broader set of energy efficiency challenges, fostering sustainable AI practices that align operational efficiency with environmental responsibility.

Data:
We logged energy consumption in Joules (J) for the cpu package per second for the twitter trace workload. Each line of the data log shows the energy consumption per second of the cpu package using the perf utility in linux as shown in Figure 1. The per second data was very noisy when first graphed so we visualized our data by using a per minute average as well as smoothing the graphs to make the trends in our energy consumption more interpretable.

```
Starting CPU energy monitoring for video-mul-1...
     1.000096633              71.65 Joules power/energy-pkg/
     2.000322608              72.91 Joules power/energy-pkg/
     3.000516939              73.71 Joules power/energy-pkg/
     4.000755138              71.90 Joules power/energy-pkg/
     5.000993450              71.10 Joules power/energy-pkg/
     6.001219742              71.32 Joules power/energy-pkg/
     7.001451818              70.88 Joules power/energy-pkg/
     8.001678273              72.04 Joules power/energy-pkg/
     9.001848192              74.87 Joules power/energy-pkg/
    10.001999723             134.09 Joules power/energy-pkg/
    11.002119003              86.83 Joules power/energy-pkg/
    12.002278526              95.87 Joules power/energy-pkg/
    13.002396434              83.72 Joules power/energy-pkg/
    14.002558695              91.32 Joules power/energy-pkg/
    15.002682422              88.60 Joules power/energy-pkg/
    16.002849467              81.99 Joules power/energy-pkg/
    17.003013235              86.58 Joules power/energy-pkg/
    18.003181390              82.35 Joules power/energy-pkg/
    19.003315290              83.20 Joules power/energy-pkg/
    20.003504849              78.84 Joules power/energy-pkg/
    21.003688958              73.41 Joules power/energy-pkg/
    22.003932934              80.47 Joules power/energy-pkg/
    23.004118598              78.65 Joules power/energy-pkg/
    24.004303940              89.84 Joules power/energy-pkg/
    25.004484499              85.32 Joules power/energy-pkg/
    26.004725341              74.23 Joules power/energy-pkg/
```

Figure 1. Sample Energy Consumption Output

Methods:

Experiment Replication

Our first step included a replication of the IPA pipeline where we ran the pipeline on the twitter trace workload which included images and text. The inference pipeline included IPA, FA2-low, FA2-high, and RIM-high. The throughput of the workloads varied from bursty, steady low, steady high, and fluctuating. We ran the four pipelines across the twitter trace workload for four different workloads. This means that there were a total of 4 different workloads across 4 different inference pipelines for a total of 20 workloads.

Overview of Energy Profiling in ML Systems

Energy profiling in machine learning inference systems is crucial for understanding and optimizing the energy consumption of computational pipelines. Several tools are available for this purpose, ranging from hardware-integrated solutions like Intel's Running Average Power Limit (RAPL) to external power meters. The selection of an appropriate tool requires careful consideration of its compatibility, measurement granularity, integration ease, and operational overhead. In this project, perf was selected as the primary tool for energy measurement due to its alignment with these requirements and its ability to integrate seamlessly into the existing Inference Pipeline Adapter (IPA) framework.

Criteria for Tool Selection

The choice of an energy measurement tool was guided by several critical factors. First, compatibility with the IPA system and the broader experimental environment was essential. The tool needed to support existing hardware and software configurations while providing sufficient flexibility to accommodate diverse workloads. Second, the tool's ability to deliver fine-grained energy consumption data was critical for capturing the nuances of energy usage across different model configurations, batch sizes, and workloads. Third, ease of use was prioritized to ensure that the tool could be efficiently integrated into automated workflows for benchmarking. Finally, the profiling tool needed to impose minimal computational overhead to avoid distorting the actual energy usage patterns of the inference pipelines.

Rationale for Choosing perf

Perf was chosen as the energy profiling tool for this project due to its strong alignment with the aforementioned criteria. As a widely adopted performance monitoring tool, perf integrates directly with Intel's RAPL interface, enabling detailed energy measurements for CPUs. This compatibility makes it particularly well-suited for analyzing energy consumption in CPU-intensive inference tasks. Furthermore, perf provides granular energy consumption metrics, including per-core and package-level data, which are essential for fine-tuned energy optimization in ML inference systems.

The integration of perf into the IPA framework was straightforward due to its support for command-line operation and scripting. This feature facilitated its use in automated experiments, where energy consumption needed to be measured across a wide range of configurations and workloads. Additionally, perf introduces minimal overhead during profiling, ensuring that the energy measurements accurately reflect real-world performance. The tool's extensibility, allowing for the measurement of additional performance metrics such as CPU cycles and cache usage, further supports its role as a comprehensive performance analysis solution.

Another key factor supporting the choice of perf was its robust community support and extensive documentation. As a standard tool in Linux environments, perf benefits from regular updates and well-maintained resources, which facilitated its adoption and use in this project. This reliability ensured that potential issues could be resolved quickly, allowing the focus to remain on the primary research objectives.

Addressing Limitations
While perf provides significant advantages, it has certain limitations. One notable constraint is its reliance on hardware support for RAPL, which restricts its usage to compatible processors. To address this, the experiments were conducted on hardware platforms in the Chameleon server with verified RAPL compatibility. Moreover, perf primarily measures CPU energy consumption and does not directly account for GPU energy usage. Acknowledging this limitation, we only performed the experiments on CPUs.

Segmentation Workload:
Segmentation is a fundamental pre-processing step in many video tracking and classification tasks. We aimed to benchmark one of the foundational one-shot segmentation models in the computer vision space known as Segment Anything Model (SAM) [8]. SAM can generate high quality segmentation masks in an unsupervised mode and a mode that takes in a seed point for segmenting specific regions in an image. The unsupervised mode returns an individual mask label for each object in the scene. The seed point mode takes in a point prompt or bounding box and returns the respective mask for said region. It contains three different variants which include ViT-B, ViT-H, and ViT-L. ViT-B contains 91M parameters, ViT-L contains 308M parameters and ViT-L contains 636M parameters. We used 59 samples from the CityScapes dataset to benchmark the three different model variants because it is a good dataset that can eb sued for downstream task such as object detection for autonomous vehicles [10]. CityScapes contains RGB images taken from a moving car with a stereo camera setup. It contains input images and ground truth segmentation mask pairs for images of resolution 2048 x 1024. Segmentation can give us different contours of objects such as civilians, cars, and roads in real-world driving scenarios. The contours can be used as candidate regions for object detectors in autonomous systems.

Integration of Q-learning Optimizer into the IPA Framework:
This part of the project aims to test Q-Learning as an approach to tackle optimization scenarios involving dynamic workloads, nonlinear trade-offs, and environments where traditional approaches like Gurobi might struggle. The key motivations included:
- Dynamic Workloads: Optimizing systems under changing input rates and system conditions.
- NonLinear Trade-offs: Addressing non-convex and non-linear relationships among accuracy, resource usage, latency, and throughput.
- RL experimentation: study the adaptability and decision-making by reinforcement learning and determine whether it's a viable alternative or complementary method to classical deterministic approaches.

Formulation of Q-Learning for Integrating into IPA
Q-Learning is formulated to solve the pipeline optimization problem by mapping the configuration options of the pipeline into a state-action-reward framework. The implementation is meant to interact directly with the pipeline's key parameters while adhering to its constraints.

Here is how the Q-Learning algorithm is adapted for this use case:

State Space
The state space was designed to the model through the current configuration of the pipeline. Each state is a combination of:

- **Model Variants:** The active variant of each stage in the pipeline.
- **Replica Counts:** The number of replicas allocated to each stage.
- **Batch Sizes:** The batch size used by each stage.

These dimensions define a multi-stage state space, where each state is a complete pipeline configuration.

Action Space
The action space consists of all possible adjustments that can be made to the pipeline configuration, including the following:

- Switching model variants for any stage.
- Increasing or decreasing the number of replicas for any stage within a prespecified range.
- Adjusting the batch size for any stage, with constraints from hardware and performance limitations.

Actions are designed such that the optimizer can easily explore neighbors of good configurations to balance exploration and exploitation.

Reward Function
The reward function is designed to reflect the objectives of optimization of the pipeline:
1. Maximizing Accuracy: The optimizer receives a reward for the settings improving the overall accuracy of the pipeline.
2. Minimizing Resource Usage: Higher utilization of CPU or GPU results in negative rewards.
3. Meeting SLAs: The settings that violate latency requirements or fail to sustain the arrival rate are heavily penalized.
4. Balancing Throughput and Latency: Rewards account for configurations that maintain throughput while staying within latency bounds.
This reward function directly incorporates the *objective* and *constraints* methods from the pre-defined *Optimizer* class.

Q-Table and Learning Process
The Q-Table keeps track of the expected return for every state-action combination. The learning process involves a step-by-step exploration of the state-action space where the optimizer executes different configurations, gets rewards with values as per the reward function, and updates the *Q-Table* accordingly. The optimizer learns to give more importance to actions yielding higher rewards over time, converging on optimal or near-optimal configurations.

Exploration-Exploitation Trade-off
The implementation uses an ε-greedy policy as described here. With probability ε, the optimizer explores new actions to find better configurations. On the other hand, it exploits the best-known

actions with probability 1 - ε, based on the current *Q-Table*. However, the exploration rate ε decays with time to balance initial exploration w.r.t. final exploitation.

Migration and Integration Process

The Q-Learning optimizer was implemented in the current *Optimizer* class without changing the structure of the class, and without breaking any functionality present:

1. Conformity with the Current Framework: The newly added method is designed with the same design patterns followed by the class, reusing methods like *latency_parameters*, *throughput_parameters*, and *accuracy_parameters* for compatibility.

2. Preservation of Signatures: This approach has the same input parameters and return types as other optimizers so it can be called interchangeably.

3. Constraints Handling: SLA and throughput constraints are baked right into the reward function to ensure the optimizer narrows down to feasible configurations only.

Pipeline Interaction

The Q-Learning optimizer interacts with the pipeline in the following ways. It dynamically selects configuration through exploration of states and actions, and then applies the configurations on the pipeline through methods like *model_switch*, *re_scale*, and *change_batch* and evaluates the pipeline performance using throughput, latency, accuracy, and resource utilization.

Experiments using Q-Learning:

The Q-learning component of the project is still an ongoing effort, and due to the technical issues and pivots that took place while formulating the problem, as a well-fitting problem for Q-learning, as well as the high configuration space of the whole project and the complexity of the different technologies used in the original project, we do not have appealing results for the current version of the document.

It is noteworthy that, as of now, we have done different experiments with different versions of the implementation for integration of Q-learning. Although all of these ended up in failed experiments, the latest formulation and version of the code was the most promising one based on the group decision, and we are putting all of our efforts into getting the results. We hope to see better results by the final version of this submission (and the video presentation).

However, the version of the code we have prepared so far is available in the project repository. We believe that these failures have had an important role in our learning process, throughout the course and working on the project.

Experiments:

This section presents the experimental results obtained from replicating the Inference Pipeline Adapter (IPA), monitoring energy consumption, creating diverse workload pipelines, and developing a Reinforcement Learning (RL)-based optimizer for the IPA. We see similar results to the original paper by Saeid et al. from the Inference Pipeline Adapter paper when looking at Figures 2 and 3. Some variations in pipeline accuracy score (PAS) and cost in the steady high graphs can be seen in Figures 2 and 3. Upon further inspection we believe that some errors with assigning minion nodes may lead to some differences such as higher cost and PAS in some workloads such as the steady high workload. We also see a variation in the pipeline

accuracy score when looking at the bursty workload which could be due to all the load being provisioned on a single master node for kubernetes instead of 2 other minion nodes.

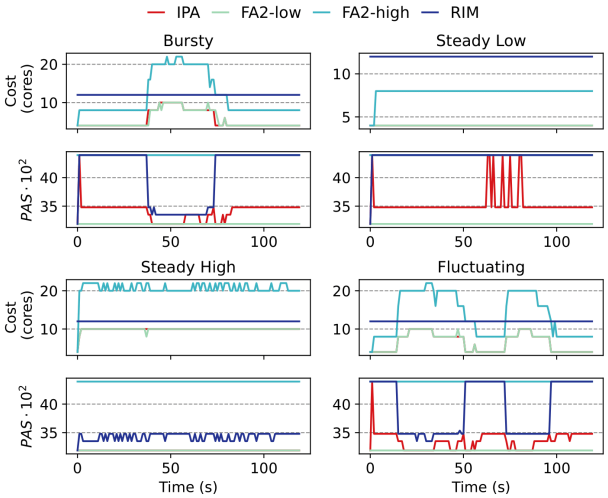


Figure 2. Original Paper Results

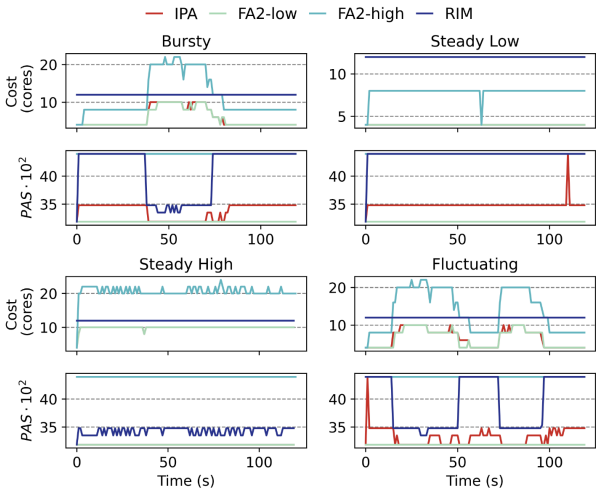


Figure 3. Replication Results

Energy Consumption of CPU per workload:

The following figures show the cpu energy consumption per minute across four workloads (bursty, steady low, steady high, and fluctuating) across four inference pipeline models (IPA, FA2-low, FA2-high, RIM-high). We noticed that IPA had the lowest energy consumption (J) throughout all four workloads as seen in Figures 4-6. RIM-high had the highest overall energy consumption with the exception of the steady low workload in Figure 5.

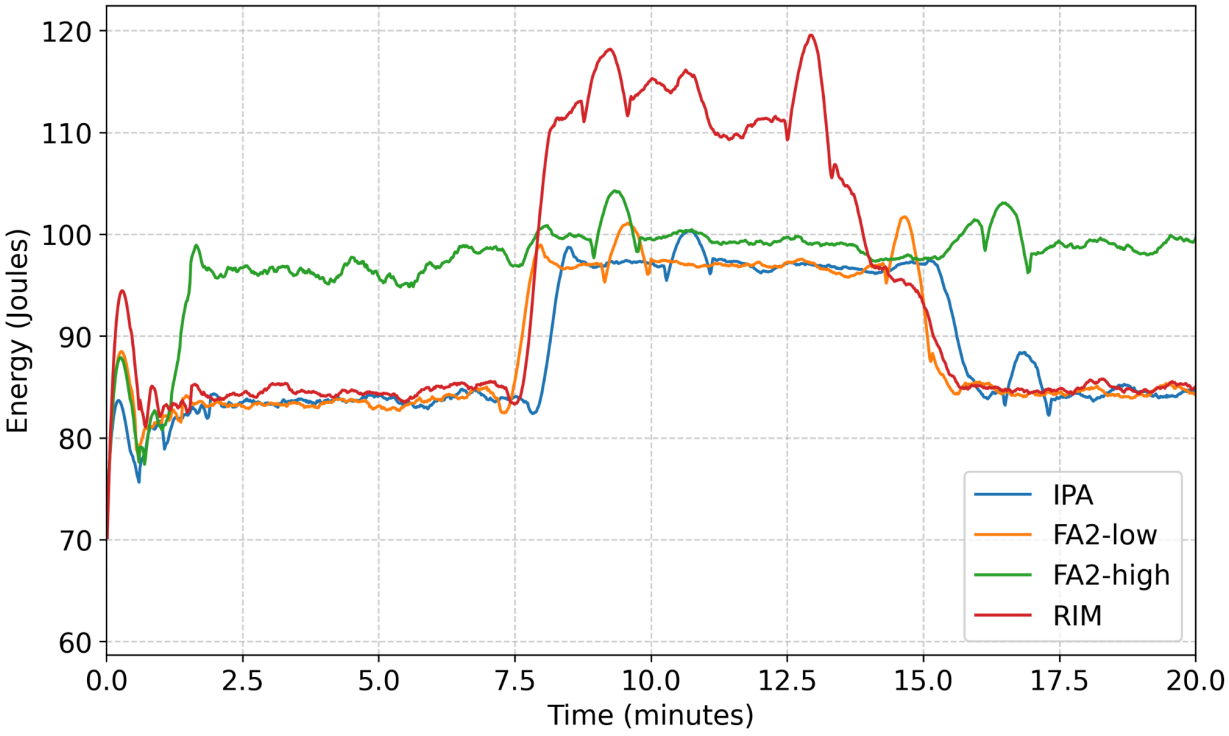


Figure 4: Energy measurement of bursty workload across four inference pipelines

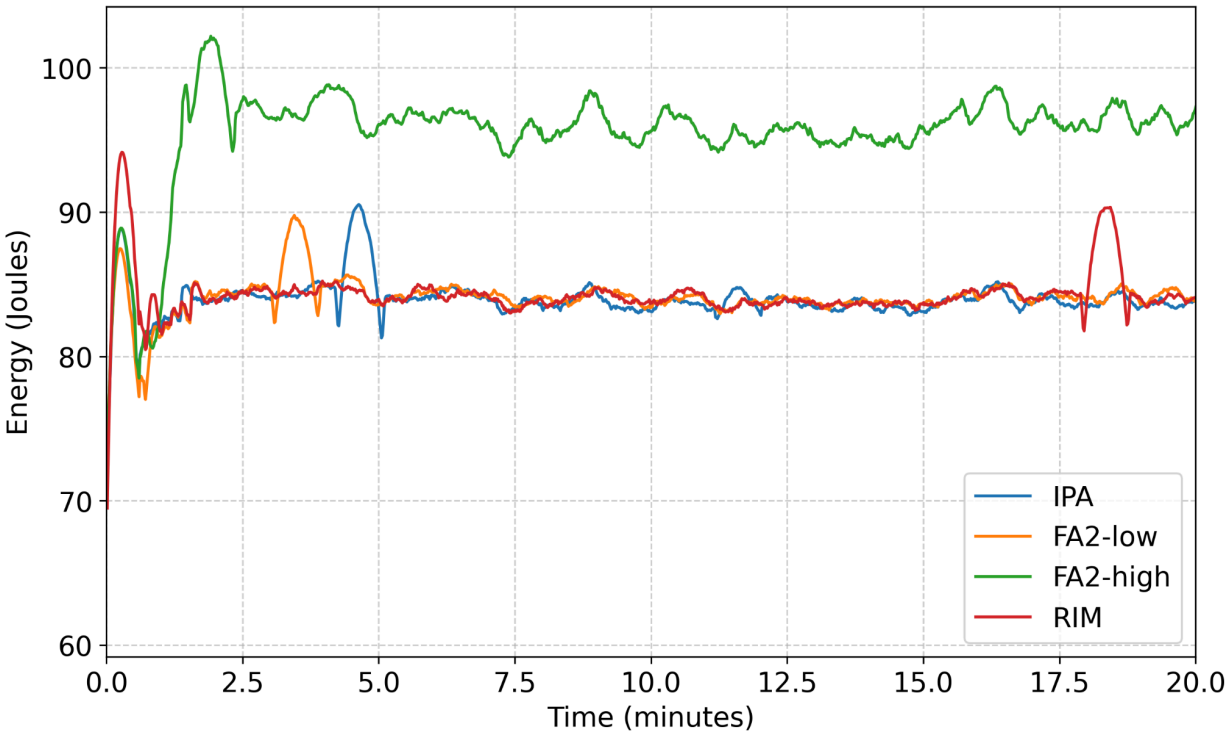


Figure 5: Energy measurement of steady low workload across four inference pipelines

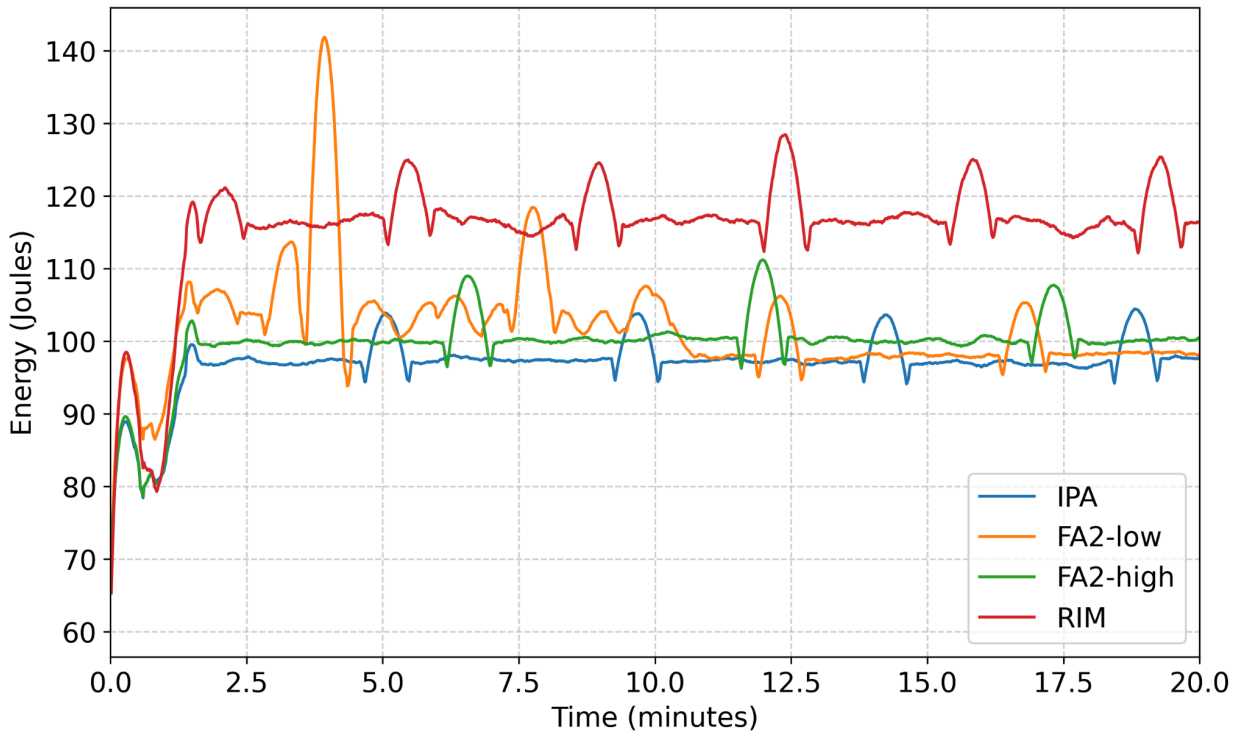


Figure 6: Energy measurement of steady high workload across four inference pipelines

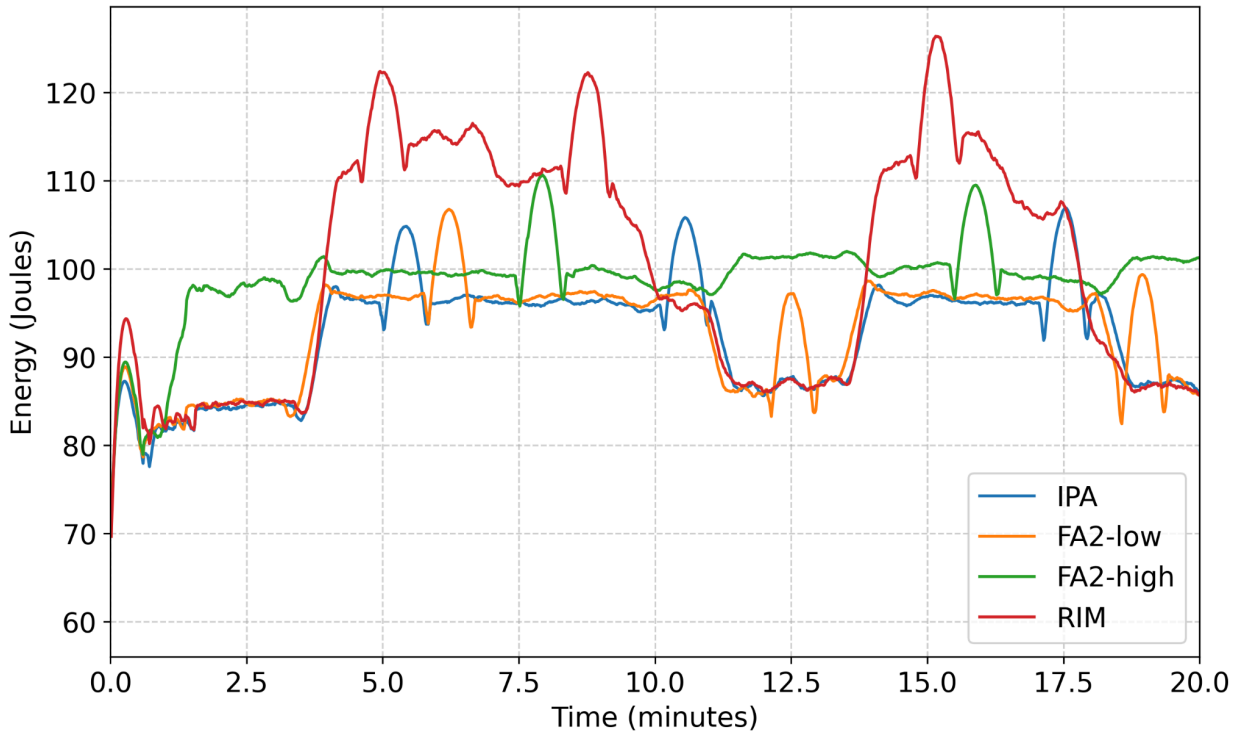


Figure 7: Energy measurement of fluctuating workload across four inference pipelines

Furthermore, we were able to benchmark the cpu energy consumption of the 3 different variants of the segment anything model as shown in Figure 8. We see that ViT-L consumed less energy

and finished the segmentation workload in less time than ViT-B and ViT-H. We see that ViT-H and ViT-L had the highest energy profiles between the 0-120 minute time interval. This shows us that model parameter size and energy consumption are inversely related. We calculated the average Intersection over Union (IoU) metric for the three variants of our model as seen in Table 1. IoU is calculated using IoU=(area of union) / (area of intersection) between ground truth mask and predicted mask. The IoU score from the three models were identical (0.5257) and this is due to the performance of the unsupervised segmentation in the SAM model. The mask outputs from SAM are much more granular than the ground truth masks thus a more appropriate unsupervised segmentation dataset may be better for assessing segmentation performance with SAM. A side-by-side comparison of the ViT-L predicted mask and ground truth CityScapes mask can be seen in Figure 9 below. In summary the tradeoff between model parameter size and energy consumption is not linear in these zero-shot segmentation models.

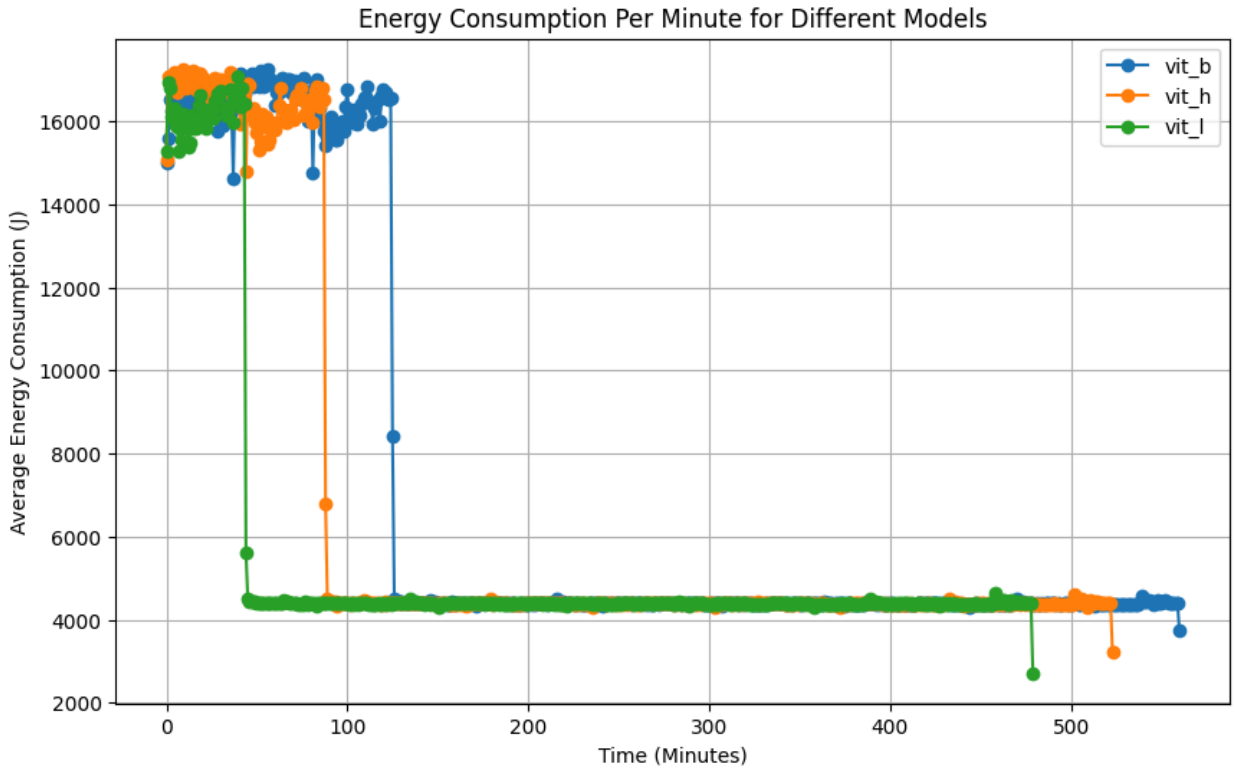
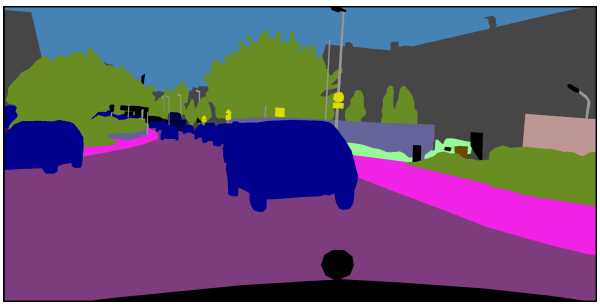


Figure 8. Energy measurement of the three different segment anything model variants

Table 1. IoU scores for the three variant of SAM for the CityScapes Workload

	ViT-B	ViT-H	ViT-L
Average IoU	0.5257	0.5257	0.5257



Ground Truth Mask

Predicted Mask from SAM (ViT-L)

Figure 9. Comparison of Ground Truth Mask to SAM (ViT-L) mask

Conclusion:
In summary, we were able to monitor the energy consumption of four workloads across four inference models and understand some trends in energy consumption. IPA has the lowest power consumption when compared to other models. Some future steps would be to develop a more robust segmentation workload to integrate into IPA, quantify Q-learning results when compared to Gurobi, and integrate power consumption optimization into the tradeoff between energy consumption, cost, and pipeline accuracy score.

Supplementary Materials:
Final presentation video link: https://youtu.be/aiqI5TcQTSs

References:
[1] Ghafouri, S., Razavi, K., Salmani, M., Sanaee, A., Lorido-Botran, T., Wang, L., ... & Jamshidi, P. (2023). IPA: Inference Pipeline Adaptation to Achieve High Accuracy and Cost-Efficiency. arXiv preprint arXiv:2308.12871.

[2] Li, B., Samsi, S., Gadepally, V., & Tiwari, D. (2023, November). Clover: Toward sustainable ai with carbon-aware machine learning inference service. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-15).

[3] Crankshaw, D., Wang, G., Zhou, X., Franklin, M. J., Gonzalez, J. E., & Stoica, I. (2020). InferLine: Latency-aware provisioning and scaling for prediction serving pipelines. Proceedings of the ACM Symposium on Cloud Computing (SoCC), 477–491.

[4] Du, Z., Guo, W., Wang, X., Gao, B., Wang, Z., & Zhang, Y. (2023). Loki: A system for serving ML inference pipelines with hardware and accuracy scaling. Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP), 427–439.

[5] He, C., Sun, L., Li, J., & Li, K. (2021). AutoInfer: Self-driving management for resource-efficient, SLO-aware machine-learning inference in GPU clusters. IEEE Transactions on Parallel and Distributed Systems, 32(1), 92–105.

[6] Harlap, A., Narayanan, D., Phanishayee, A., Seshadri, V., & Ganger, G. R. (2017). Swayam: Distributed autoscaling to meet SLAs of machine learning inference services with resource efficiency. Proceedings of the 2017 ACM Symposium on Cloud Computing (SoCC), 185–197.

[7] Harty, A., Shah, N., Fleming, T., Asuni, N., & Jiang, N. (2023). Computing within limits: An empirical study of energy consumption in ML training and inference. arXiv preprint arXiv:2303.12101.

[8] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015-4026).

[9] Li, B., Jiang, Y., Gadepally, V., & Tiwari, D. (2024). Toward sustainable genai using generation directives for carbon-friendly large language model inference. arXiv preprint arXiv:2403.12900.

[10] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213-3223).