Evaluating Generational Improvements in NVIDIA x60-Class GPUs Under Academic AI Workloads

Cedric John C. Cangco*
Mapua University
Makati, Philippines
cccangco@mymail.mapua.edu.ph

Angel Michael Lu.
Mapua University
Makati, Philippines
amalu@mymail.mapua.edu.ph

Vincent Angelo S. Macaraeg
Mapua University
Makati, Philippines
vasmacaraeg@mymail.mapua.edu.ph

Jefferson A. Costales
Mapua University
Makati, Philippines
jacostales@mapua.edu.ph

Abstract— Consumer-grade x60-class GPUs are commonly used for academic AI training due to their balance of cost and performance. As newer GPU generations become available, it is important to understand whether upgrading within the same tier provides meaningful benefits in practical academic workloads. This study compares the NVIDIA RTX 3060 Ti (Ampere) and RTX 5060 Ti (Blackwell) using representative AI training tasks under controlled and identical system conditions. To ensure fairness, the RTX 5060 Ti was constrained to the same VRAM capacity as the RTX 3060 Ti. Results show that performance differences are minimal for lighter workloads, while the newer RTX 5060 Ti demonstrates improved efficiency through consistently lower power consumption. For more demanding workloads, the RTX 5060 Ti exhibits clearer performance advantages, highlighting the impact of architectural improvements. Overall, the findings indicate that older x60-class GPUs remain viable for basic academic training, while newer generations offer greater efficiency and scalability for more demanding models and parameter-heavy tasks.

Keywords—AI Training, Consumer GPUs, NVIDIA, Ampere, Blackwell

I. Introduction

Many students and academic researchers rely on consumer-grade GPUs for local model training. This has been made possible because of AI education resources which are easily accessible and increased availability. Tighter budgets have made the x60-class GPUs the main choice for those choosing affordability and computational capability. As such, x60-class GPUs remain in use for coursework, prototyping, and small-scale research. Steam Hardware Survey's market data supports this claim, stating that x60-tier chips remain among the most common choice for desktop users for gaming and general purposes [1].

Due to the continuous release of newer generations, consumers have questioned the practicality of upgrading older models for newer designs within the same product tier. A report has shown that the NVIDIA Blackwell architecture offers major performance improvements on the standard benchmark suite for AI training when compared to the previous GPU architecture generations. This provides an emphasis on the need for real-world workloads as evaluation [2, 3]. As such, understanding whether newer generations provide substantial advantages beyond their raw specifications would help enable consumers to have informed hardware decisions in resource-constrained academic environments.

To represent academic AI workloads, the generational improvements between the NVIDIA RTX 3060 Ti (Ampere) and the RTX 5060 Ti (Blackwell) are measured for demonstration. The convolutional, transformer-based, and object detection models are evaluated through controlled training conditions and equal VRAM constraints in order to isolate architectural and efficiency gains. The results provide practical insights into the true-world impact of GPU generation upgrade on student and entry-level AI training.

II. Theoretical Background

A. GPU Acceleration in Academic AI Workloads

Modern AI training has heavily relied on GPU acceleration because of the highly parallel nature of operations such as matrix multiplications, convolutions, and attention mechanisms. As such, deep learning frameworks translate these operations into optimized GPU kernels. These kernels enable thousands of threads to execute simultaneously. Also, it reduces the training time compared to CPU-based execution [5]. In an academic setting, the commonly used workloads include convolutional neural networks (CNNS) for image classification, transformer-based models for natural language processing, and object detection architectures for applied coursework and research projects [6]. The workload performance is influenced by raw compute throughput, memory bandwidth, kernel scheduling efficiency, and execution stability. This makes the GPU architecture an important factor in real-world training behavior.

B. NVIDIA GPU Architectural Evolution

NVIDIA GPU architectures have upgraded and evolved throughout their generations in order to improve the GPU's computation efficiency, power efficiency, and AI-specific acceleration. The Ampere architecture has introduced enhanced Tensor Cores and improved mixed-precision support, which enables faster execution of deep learning workloads compared to earlier designs [7]. The following architectural developments emphasized the improved kernel scheduling, data movement efficiency, and tighter integration with modern deep learning frameworks. With the Blackwell architecture, it continues this direction by focusing more on efficient tensor execution and improving the utilization of modern AI kernels. This aims to bring performance and efficiency improvements even when it is under constrained memory conditions [2,3]. These improvements make controlled generational comparisons important in assessing academic AI training workloads.

C. Controlled Memory Constraints

In AI training, the role of Video Random Access Memory (VRAM) is to store model parameters, intermediate activations, gradients, and input data branches. With larger VRAM capacities, larger batch sizes or higher input resolutions are made possible, which can improve throughput. However, these large capacities also introduce other variables when comparing different GPUs [8]. Due to how batch sizes can directly affect memory usage and iteration behavior, uncontrolled VRAM difference would result in unclear architectural efficiency improvements. In order to isolate execution efficiency away from memory-driven advantages, applying a fixed VRAM constraint and through the use of identical batch sizes across GPUs is a common experimental strategy. This approach help enable performance differences to be credited to architectural design and kernel efficiency instead of batch scaling behavior to produce a fair assessment in academic environments.

III. Conceptual Framework

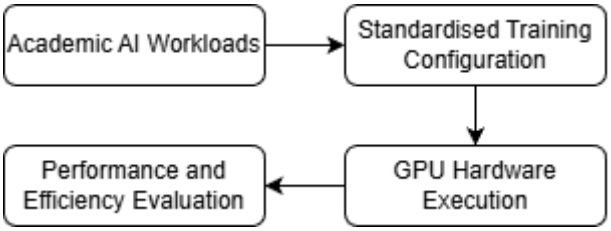

Figure I illustrates the Conceptual Framework for the comparative analysis of the NVIDIA RTX 5060 Ti and NVIDIA RTX 3060 Ti.

The figure shows the conceptual framework used by the researchers in this study. The representative academic AI workloads are standardized through fixed preprocessing and batch-size configurations. Afterwards, training is conducted within a controlled environment, by applying an 8 GB VRAM constraint to eliminate memory capacity as a confounding variable. The workloads are then executed on two NVIDIA GPU generations, Ampere and Blackwell, to observe their differences in execution behavior. During training, the performance and efficiency data are collected to assess the improvements between the generations under identical workload and memory conditions.

IV. GPU Specifications

A. Architectural and Hardware Differences

NVIDIA RTX 5060 Ti and RTX 3060 Ti differ in memory design, compute capability, and acceleration features. These architectural variations influence how each GPU handles AI training workloads.

Architecture/Hardware	RTX 5060Ti	RTX 3060Ti
Memory Type	GDDR7	GDDR6
Memory Bandwidth	448 GB/s	448 GB/s
PCIE Interface	PCIE 5.0 x 8	PCIE 4.0 x 16
FP32 Compute	19 TFLOPS	16.2 TFLOPS
Boost Clock	2572 Mhz	1665 Mhz
Architecture	Blackwell	Ampere
Accelerators	Tensor Cores, RT Cores	Tensor Cores, RT Cores
Driver/Backend	CUDA+ cuDNN	CUDA+ cuDNN

Table I. illustrates the architecture and hardware of the two GPUs

The table summarizes the architectural and hardware characteristics of both NVIDIA RTX 3060 Ti and RTX 5060 Ti. Although these GPUs belong to the x60-class segment, with similar memory bandwidth, their architectural designs and platform features differ from one another. The RTX 3060 Ti is based on the Ampere architecture, featuring GDDR6 memory and a PCIe 4.0 x16 interface. On the other hand, the RTX 5060 Ti utilizes the newer Blackwell architecture with GDDR7 memory and a PCIe 5.0 x8 interface. These differences will help the assessment be generational architectural efficiency focused instead of memory throughput or capacity

B. Rationale for GPU Selection

The RTX 3060 Ti and RTX 5060 Ti were selected as representative x60-class GPUs commonly used in academic and student settings due to their balance of affordability and performance. To serve as the baseline, the researchers chose the RTX 3060 Ti. The RTX 3060 series remains one of the most widely used desktop GPUs. In contrast, the RTX 5060 Ti will represent the latest generation of the x60-class GPUs, making it a suitable choice for existing users as an upgrade. By comparing these two GPUs under identical system and memory constraints, the data will focus on the differences in generational architecture and efficiency improvements, rather than capacity and system configuration. This selection enables a market-relevant assessment of how newer GPU generations translate into practical benefits for academic AI training workloads.

V. Literature Review

A. Consumer GPUs in Academic AI Training

Consumer-grade GPUs are commonly used in AI training for accessibility, affordability, and decent performance in coursework and small-scale research. Mid-range consumer GPUs can assist with typical academic workloads, such as image classification, natural language processing, and object detection, without the need for specialized hardware [7,8]. Although cloud platforms provide access to high-end accelerators, they provide limited runtime, memory, and reproducibility leading to students and researchers to rely on local GPUs for consistent experimentation [9].

B. Generational Improvements in NVIDIA GPUs

The evolution of NVIDIA GPU architectures highlighted the improvements in efficiency, scheduling, and AI-specific accelerations across generations. Consecutive NVIDIA architectures also emphasized that performance improvements are caused by architectural refinements and optimized execution pipelines rather than raw improvements in compute throughput [10,11]. Meanwhile, newer NVIDIA architectures have demonstrated improved tensor execution, enhanced kernel fusion, and better utilization of modern deep learning frameworks, resulting in measurable gains in training efficiency for convolutional and transformer-based models [2,3]. However, most existing literature relies on vendor benchmarks or synthetic workloads, resulting in a gap in empirical evaluations that measure generational improvements under a controlled, workload-driven academic training setting.

C. GPU Benchmarking Practices for AI Training

Effective GPU benchmarking for AI training requires controlled experimental design to ensure fair and reproducible comparisons. Prior benchmarking studies emphasize the importance of using identical training configurations, consistent datasets, and meaningful performance metrics such as throughput, total training time, power consumption, and performance per watt [12], [13]. Reliance on theoretical specifications alone has been shown to be insufficient for predicting real-world training behavior, as factors such as kernel maturity, memory handling, and execution stability significantly influence outcomes [14]. Consequently, workload-based benchmarking under controlled conditions is widely regarded as the most reliable approach for evaluating GPU performance in academic AI training contexts.

VI. Methodology

A. Models and Datasets

This study examines four representative AI workloads frequently used in undergraduate coursework and entry-level research. Each model is evaluated using a commonly adopted benchmark dataset, as summarized in Table II.

Resnet-152	30 Epochs	Oxford-IIIT Pet dataset
DistilBERT	30 Epochs	AG NEWS 2K
YOLOv8n	30 Epochs	COCO Subset 5K
YOLOv8m	30 Epochs	COCO Subset 5K

Table II. shows the chosen datasets used to test the models

Together, these models represent a balanced set of computer vision, natural language processing, and object detection workloads commonly encountered in academic exercises and small-scale research projects.

B. Hardware and GPU Configuration

CPU	**Intel Core i7-12700K**
Memory	32 GB DDR4 ECC
Storage	2 x 1 TB WD Black SN850x
Power Supply	1000W Gold + Rating
NVIDIA Driver	581.57

Table III. This represents hardware and driver specifications on the setup.

At any given time, only one GPU and its corresponding NVMe drive were installed in the system to ensure an isolated and conflict-free testing environment. All other system components remained unchanged across experiments. Each NVMe drive contained the vendor-specific drivers and software stack required for its respective GPU.

C. Training Settings and Benchmark Procedure

All models were trained for 30 epochs using identical code, preprocessing, and hyperparameter settings. Default PyTorch and Ultralytics configurations were used, with batch size fixed under the imposed VRAM constraint. Experiments were conducted on Windows 11 using Visual Studio Code with CUDA 12.9, with throughput and timing obtained from training logs and supplementary metrics monitored via HWInfo and Task Manager.

D. Performance Metrics

The following metrics were collected during all training runs to evaluate GPU behavior under academic AI workloads:

- **Training Throughput** – measured in samples per second, depending on the workload.
- **Total Training Time** – the total duration required to complete all training epochs for each model.
- **Average VRAM Usage** – peak and average memory consumption observed during training, used to assess memory efficiency.
- **Average GPU Utilization** – the proportion of active compute usage throughout the training process.
- **Average Power Draw** – mean electrical power consumption during training, used to evaluate energy-related behavior.
- **Performance per Watt** – throughput normalized by average power draw to assess energy efficiency.
- **Final Validation Performance** – accuracy or validation score obtained at the end of training to confirm correct and comparable model behavior.

Collectively, these metrics provide a comprehensive view of GPU performance by capturing training speed, memory utilization, power consumption, and model correctness. This enables a balanced assessment of both computational efficiency and practical usability across different academic AI workloads.

VII. Results

A. ResNet-152 Results

Metric	RTX 5060 Ti	RTX 3060 Ti
Throughput	56.7 imgs/s	43.1 img/s
Training Time	67.03 min	81.30 min
Avg VRAM Usage	4.5 GB	4.3 GB
Avg GPU Utilization	74.60%	81.9%
Avg Power Draw	96.6 W	171.66 W
Performance per Watt	0.587 img/s/W	0.251 img/s/W
Best Validation Accuracy	22.08%	24.01%

Table IV. presents the results of training with ResNet-152

B. DistilBERT Results

Metric	RTX 5060 Ti	RTX 3060 Ti
Throughput	283 Samples/s	224 Samples/s
Training Time	18 mins	22.9 mins
Avg VRAM Usage	6.7 GB	6.7 GB
Avg GPU Utilization	99%	98%
Avg Power Draw	162.6 W	182.3 W
Performance per Watt	1.74 S/s/watt	1.23 S/s/watt
Best Validation Accuracy	91.25%	91.24%

Table V. presents the results of training with DistilBERT

C. Yolo V8n Results

Metric	RTX 5060 Ti	RTX 3060 Ti
Throughput	158.19 img/s	151.52 img/s
Training Time	12.64 mins	13.9 mins
Avg VRAM Usage	3.7 GB	3.6 GB
Avg GPU Utilization	66%	97%
Avg Power Draw	83.65 W	145.28 W
Performance per Watt	1.89 S/s/watt	1.04 S/s/watt
Best mAP50-95	32.84%	32.84%

Table VI. presents the results of training with Yolo V8n

D. Yolo V8m Results

Metric	RTX 5060 Ti	RTX 3060 Ti
Throughput	50.73 imgs/s	37.33 imgs/s
Training Time	39.42 mins	53.57 mins
Avg VRAM Usage	7.81 GB	6.7 GB
Avg GPU Utilization	91%	90.30%
Avg Power Draw	123.25 W	183.9 W
Performance per Watt	0.41 img/s/W	0.20 img/s/W
Best mAP50-95	32.84%	32.84%

Table VII. presents the results of training with Yolo V8m

E. Summary of Observed Performance Trends

Across all evaluated workloads, the RTX 5060 Ti and RTX 3060 Ti exhibit similar training performance under identical system configurations and an enforced 8 GB VRAM constraint. For ResNet-152, DistilBERT, and YOLOv8n, differences in throughput and total training time are relatively small, with both GPUs completing training successfully and achieving comparable validation metrics. More pronounced differences are observed in the YOLOv8m workload, where the RTX 5060 Ti achieves higher throughput and shorter training time despite increased VRAM usage. Across all experiments, validation accuracy and mAP values remain consistent between GPUs, indicating that observed differences are attributable to performance and efficiency characteristics rather than model convergence or correctness.

VIII. Discussion

A. Performance Across Workload Complexity

The performance impact of generational GPU improvements varies with workload complexity. For lighter to moderate workloads such as ResNet-152, DistilBERT, and YOLOv8n, differences in throughput and training time between the RTX 3060 Ti and RTX 5060 Ti remain relatively small, with the RTX 5060 Ti achieving roughly 15–20% faster training times under the 8 GB VRAM constraint. In these scenarios, both GPUs are able to sustain stable execution and high utilization, indicating that computational and memory demands do not strongly stress either architecture. As workload complexity increases, performance differences become more pronounced. This is most evident in the YOLOv8m results, where the RTX 5060 Ti achieves approximately 25–30% faster training times alongside higher throughput, indicating that generational performance benefits increase with model size and computational demand.

B. *Power Efficiency and Architectural Gains*

The consistently lower power draw observed on the RTX 5060 Ti across all workloads indicates that generational improvements are primarily driven by architectural efficiency rather than raw performance scaling. Compared to the RTX 3060 Ti, the RTX 5060 Ti reduces average power consumption by approximately **44% for ResNet-152**, **11% for DistilBERT**, **42% for YOLOv8n**, and **33% for YOLOv8m**, while maintaining comparable or improved training performance. These results show that the newer architecture converts a larger fraction of consumed power into useful computation, particularly in lighter workloads where similar performance is achieved at substantially lower utilization.

An additional observation is that the RTX 5060 Ti tends to reserve more total VRAM despite similar active memory usage, likely due to differences in memory allocation strategies. In earlier high batch-size experiments, this behavior led to out-of-memory conditions on the RTX 5060 Ti in scenarios where the RTX 3060 Ti remained stable. While not directly reflective of compute performance, this highlights generational differences in memory management that may affect workload scaling under strict VRAM constraints and may improve as software support matures.

C. *Practical Implications for Academic Users*

From a cost perspective, the RTX 3060 Ti remains attractive for academic users, as it is commonly available at roughly half the cost of the RTX 5060 Ti on the second-hand market. It provides sufficient performance for lighter to moderate workloads, albeit with higher power consumption. The RTX 5060 Ti is better suited for users prioritizing efficiency and heavier models, while multi-GPU setups using lower-cost cards may offer competitive throughput at the expense of increased power and system complexity.

IX. Conclusion

This study evaluated the generational improvements between the NVIDIA RTX 3060 Ti (Ampere) and RTX 5060 Ti (Blackwell) under representative academic AI workloads using controlled training conditions and an enforced 8 GB VRAM constraint. The results show that performance differences are relatively small for lighter to moderate workloads, indicating that older x60-class GPUs remain viable for common academic training tasks. However, the RTX 5060 Ti consistently demonstrated substantially lower power consumption across all workloads, resulting in significantly improved performance-per-watt.

For more demanding models, particularly YOLOv8m, the RTX 5060 Ti exhibited clearer advantages in throughput and training time, highlighting the growing importance of architectural efficiency as workload complexity increases. These findings suggest that while upgrading within the same GPU tier may offer limited performance gains for basic workloads, newer generations provide meaningful benefits in efficiency and scalability. Overall, the results emphasize that GPU selection for academic AI training should consider not only raw performance, but also power efficiency, workload complexity, and long-term usability.

X. Recommendations for Future Research

Future work may incorporate kernel-level profiling and per-layer analysis to better attribute observed efficiency gains to specific architectural behaviors. Evaluating larger or attention-heavy models and multi-GPU training scenarios would further clarify how generational improvements scale with workload complexity and cost. Repeating experiments under Linux-based software stacks and newer driver releases may also help assess the impact of ongoing software optimizations on emerging GPU architectures.

XI. Data and Code Availability

All code, configuration files, and raw benchmarking results used in this study are publicly available in the GitHub repository referenced in [16].

References

[1] Valve Corporation, “Steam Hardware & Software Survey: Video Card Statistics,”2025.[Online].Available:https://store.steampowered.com/hwsurvey/videocard/.

[2] NVIDIA Corporation, “NVIDIA Blackwell Enables 3× Faster Training and Nearly 2× Training Performance per Dollar than Previous-Generation Architecture,” *NVIDIA Developer Blog*, Dec. 2025. [Online]. Available: https://developer.nvidia.com/blog/.

[3] NVIDIA Corporation, “NVIDIA Blackwell Architecture Sweeps MLPerf Training v5.1 Benchmarks,” *NVIDIA Developer Blog*, Nov. 2025. [Online]. Available: https://developer.nvidia.com/blog/.

[4] A. Jarmusch and S. Chandrasekaran, “Microbenchmarking NVIDIA’s Blackwell Architecture: An In-Depth Architectural Analysis,” *arXiv preprint arXiv:2512.02189*, Dec. 2025.

[5] PyTorch Foundation, “CUDA Semantics,” *PyTorch Documentation*, 2024. [Online]. Available: https://pytorch.org/docs/stable/notes/cuda.html.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] NVIDIA Corporation, “NVIDIA Ampere Architecture In-Depth,” *NVIDIA Technical Blog*, 2021. [Online]. Available: https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth.

[8] S. Mittal, “Understanding the Impact of Batch Size on Deep Learning Performance,” *Journal of Systems Architecture*, vol. 138, 2023.

[9] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, “Deep Learning for Computer Vision: A Brief Review,” *Computational Intelligence and Neuroscience*, 2022.

[10] Google, “Google Colab Documentation,” 2024. [Online]. Available: https://colab.research.google.com.

[11] M. Ivanov, “Cross-Vendor GPU Benchmarking for Deep Learning Workloads,” *arXiv preprint arXiv:2205.01950*, 2022.

[12] NVIDIA Corporation, “NVIDIA Ada Lovelace Architecture Whitepaper,” 2022. [Online]. Available: https://www.nvidia.com.

[13] MLCommons, “MLPerf Training Benchmark Results,” 2024. [Online]. Available: https://mlcommons.org.

[14] S. Dong, J. Kosaian, and N. Shah, “GPU Workload Benchmarking for Deep Learning Applications,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 9, 2021.

[15] P. Jin, H. Zhang, and M. Ma, “Evaluation of Deep Learning Frameworks over Different GPUs,” in *Proc. IEEE Int. Conf. on Big Data*, 2021.

[16] C. J. C. Cangco, “GPU Benchmarking Code for Academic AI Workloads,” GitHub repository, 2025. [Online].Available:https://github.com/Krashedd/GPU-AI-Training-Evaluation-RTX-5060ti-16GB-RX-9060XT-16GB-on-Windows