Profiling Report

Team AIML-25

Team Leader:

Divyansh Gupta

Team Members:

Keshav Lakshmi Narasimhan |Rishan Gobse |Kartik Budhani |Krishnam Digga

Mentor:

Harshvardhan Chaudhary

August 2nd, 2025

Contents

1 Introduction

- This project addresses the eighth problem statement of IITI SoC '25 under the AI–ML domain: *Accelerate - WAN – Optimizing Inference and Training Speed of the WAN 1.3B Video Model.*

- We used the following optimization techniques to accelerate the baseline pipeline - **Quantization, LoRA, Operator Fusion, CPU offloading and attention slicing.**
- Overall, we achieved a speed up in inference speed of approximately 7x. Following are the detailed results of baseline model vs the optimization techniques individually and their combinations.
- The inputs used for benchmarking are -
  Prompt: A golden retriever puppy runs joyfully through a field of blooming sunflowers under a bright blue sky, with petals floating in the air and butterflies fluttering around. The camera slowly follows the puppy from a low angle.
  Negative Prompt: default
  Resolution: 720p
  Frames: 24
  FPS: 16
  Seed: 33
  Guidance Scale: 6.0
  Inference Steps: 65
  GPU: H-100 80 GB VRAM

- The output video of all the combinations are available in the github repository `Demo Videos` folder. The link to the repository - https://github.com/DivyanshGupta2006/Optimized-WAN-1.3B-text-to-video-generation-model

2 Optimization Techniques

2.1 Quantization

Table 1: Performance metrics recorded before and after Quantization

Metrics	None (Baseline)	Quantization
Load Time	14.213s	55.270s
Warm up time	83.098s	16.883s
MAX VRAM	56771MB	24727MB
CLIP Score	0.382	0.389
LPIPS Score	0.137	0.163
Throughput	0.020fps	0.143fps
Frame-wise Latency	48.963s	6.817s
Clip-wise Latency	4832.542s	678.138s

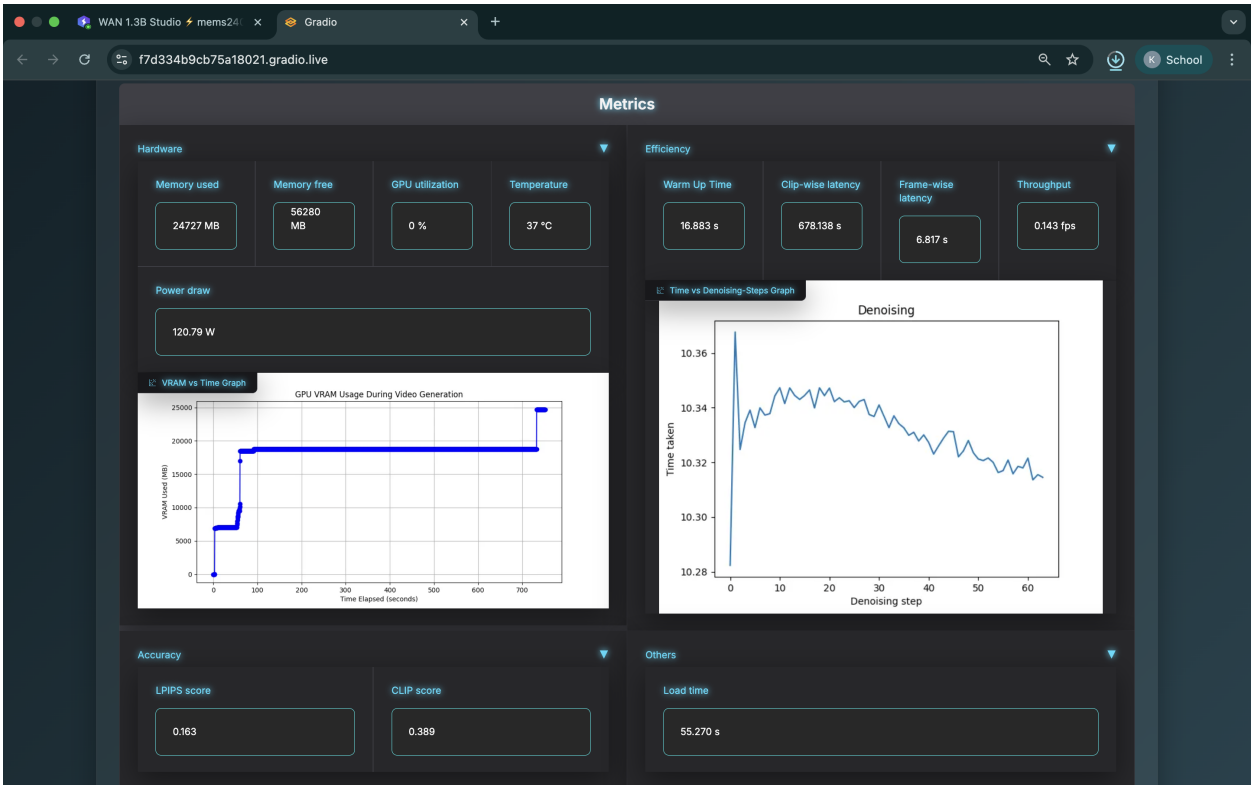

Figure 1: Quantization Performance Metrics - UI Screenshot

2.2 Operator Fusion

Metrics	None (Baseline)	Operator Fusion
Load Time	14.213s	20.142s
Warm up time	83.098s	127.648s
MAX VRAM	56771MB	57123MB
CLIP Score	0.382	0.381
LPIPS Score	0.137	0.137
Throughput	0.020fps	0.020fps
Frame-wise Latency	48.963s	48.452s
Clip-wise Latency	4832.542s	4827.515s

Figure 2: Operator Fusion Performance Metrics - UI Screenshot

Note: The operator fusion fuses the kernels so that the inference is fastened. It does so in the first forward pass, hence the increased warm-up time. Although this operation is costly, but over a batch of prompts, the cost is distributed while the acceleration is stacked, causing overall increase in inference speed. Since we are not generating prompts in batches, this technique is redundant and hence is excluded from further profiling and in the best option.

2.3 CPU Offloading

Metrics	None (Baseline)	CPU Offloading
Load Time	14.213s	83.959s
Warm up time	83.098s	77.839s
MAX VRAM	56771MB	41265MB
CLIP Score	0.382	0.366
LPIPS Score	0.137	0.285
Throughput	0.020fps	0.020fps
Frame-wise Latency	48.963s	49.035s
Clip-wise Latency	4832.542s	4834.197s

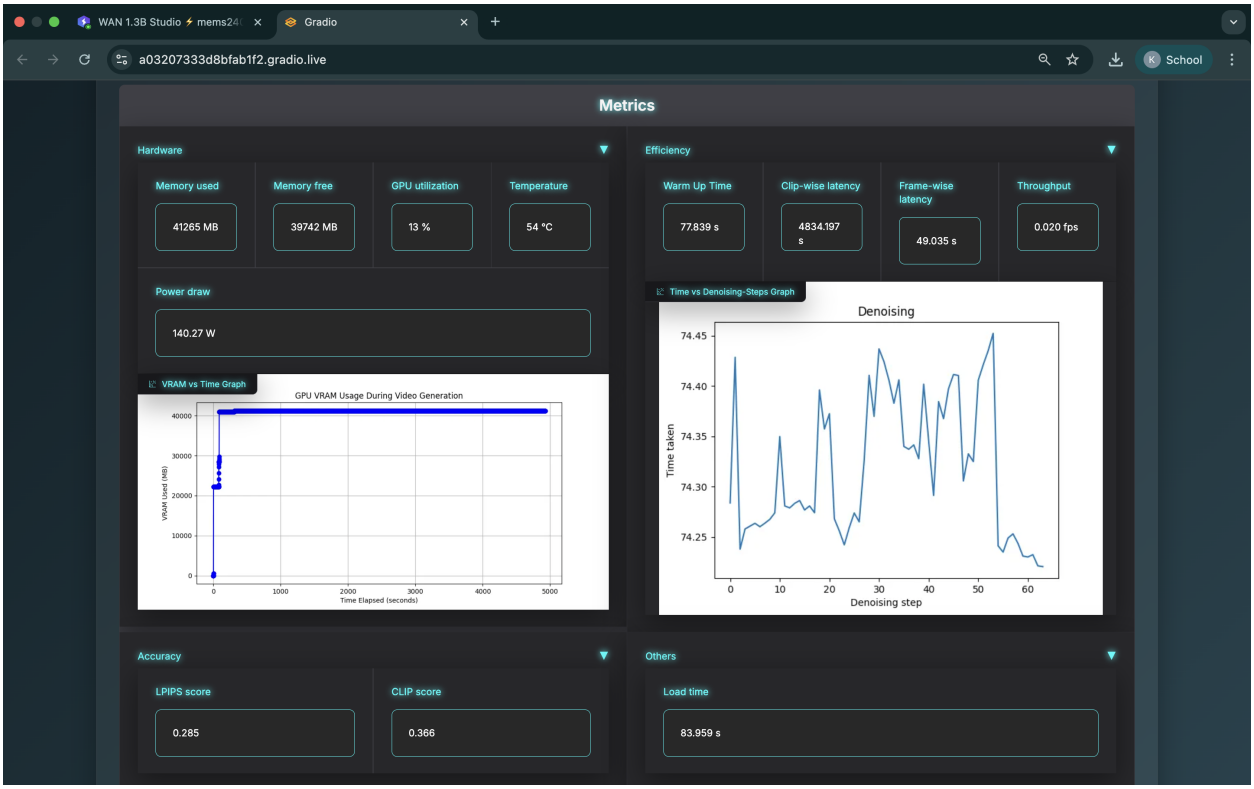

Figure 3: CPU Offloading Performance Metrics - UI Screenshot

Note: Here we had enough VRAM in GPU, so most of the part gotloaded in GPU, but on low resource devices, more components are offloaded to CPU automatically. Clearly, when we have enough VRAM, doing CPU offloading is redundant and will only increase the inference time, hence it is excluded from further profiling and in the best option.

2.4 LoRA

Metrics	None (Baseline)	LoRA
Load Time	14.213s	6.443s
Warm up time	83.098s	82.765s
MAX VRAM	56771MB	56867MB
CLIP Score	0.382	0.361
LPIPS Score	0.137	0.055
Throughput	0.020fps	0.020fps
Frame-wise Latency	48.963s	49.037s
Clip-wise Latency	4832.542s	4839.361s

Figure 4: LoRA Performance Metrics - UI Screenshot

2.5 Attention Slicing

Metrics	None (Baseline)	Attention Slicing
Load Time	14.213s	41.071s
Warm up time	83.098s	82.845s
MAX VRAM	56771MB	56771MB
CLIP Score	0.382	0.382
LPIPS Score	0.137	0.137
Throughput	0.020fps	0.020fps
Frame-wise Latency	48.963s	48.873s
Clip-wise Latency	4832.542s	4823.492s

Figure 5: Attention Slicing Performance Metrics - UI Screenshot

2.6 Schedulers

Metrics	None (Baseline)	Schedulers
Load Time	14.213s	40.788s
Warm up time	83.098s	102.585s
MAX VRAM	56771MB	56479MB
CLIP Score	0.382	0.384
LPIPS Score	0.137	0.151
Throughput	0.020fps	0.020fps
Frame-wise Latency	48.963s	48.686s
Clip-wise Latency	4832.542s	4825.097s

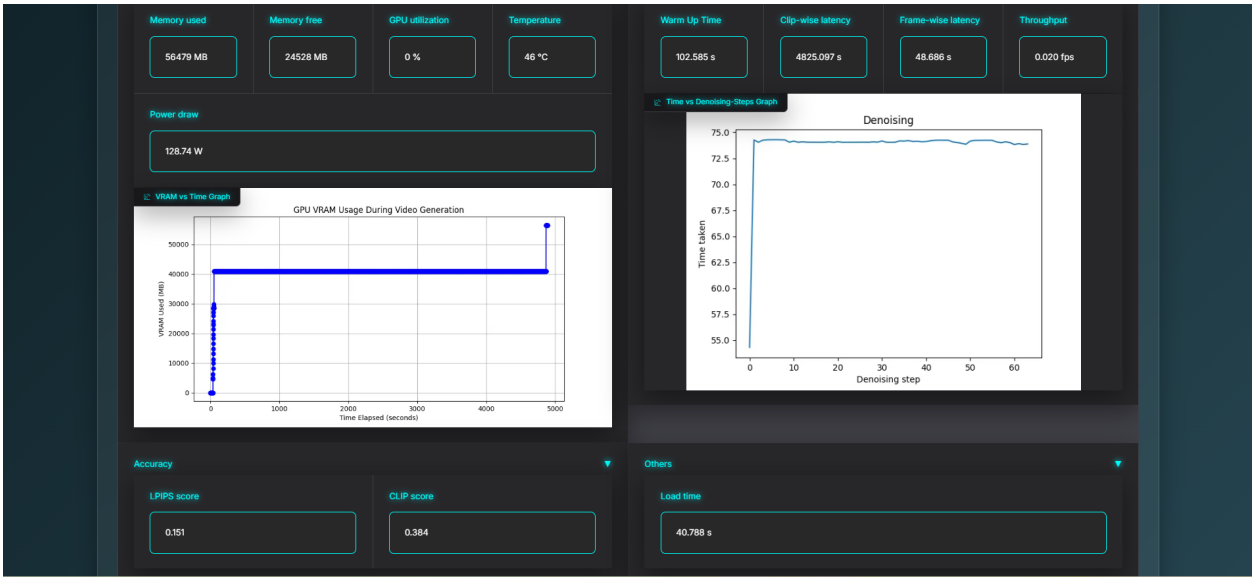

Figure 6: Schedulers Performance Metrics - UI Screenshot

2.7 Quantization + DPM + LoRA

Metrics	None (Baseline)	Quantization + DPM + LoRA
Load Time	14.213s	11.118s
Warm up time	83.098s	18.066s
MAX VRAM	56771MB	24543MB
CLIP Score	0.382	0.391
LPIPS Score	0.137	0.141
Throughput	0.020fps	0.144fps
Frame-wise Latency	48.963s	6.779s
Clip-wise Latency	4832.542s	675.629s

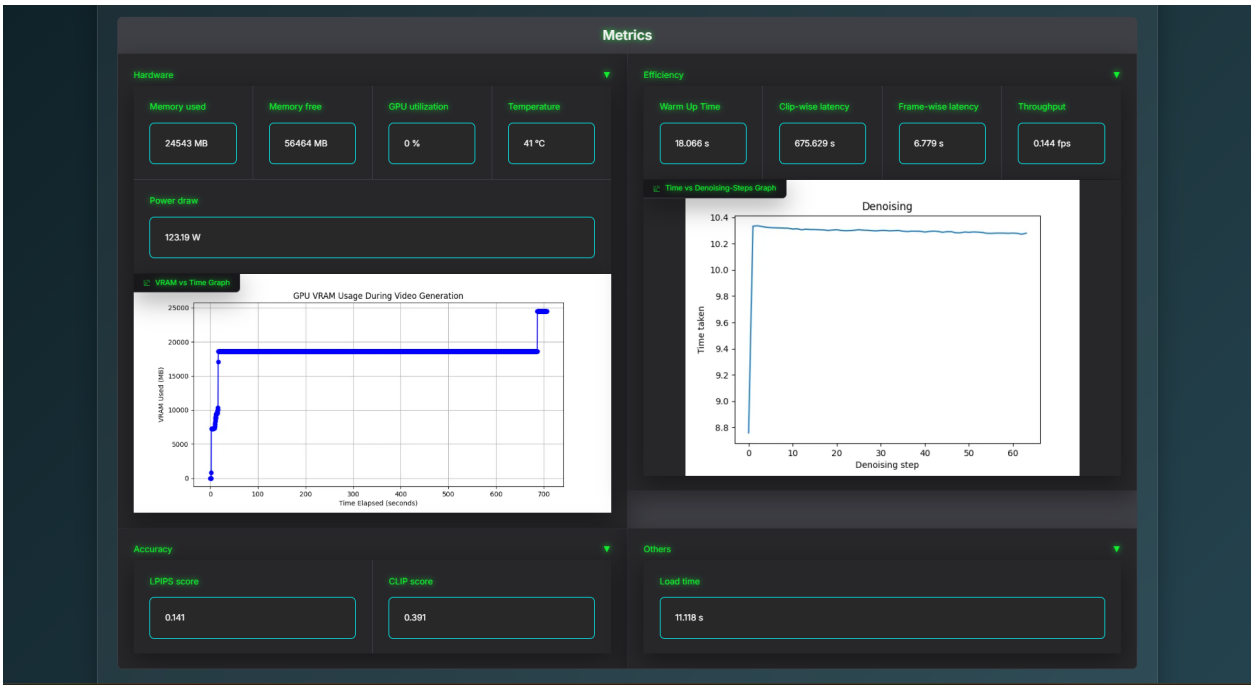

Figure 7: Optimizations Performance Metrics - UI Screenshot

2.8 Quantization + DPM + Attention Slicing

Metrics	None (Baseline)	Quantization + DPM + Attention Slicing
Load Time	14.213s	12.763s
Warm up time	83.098s	18.716s
MAX VRAM	56771MB	24453MB
CLIP Score	0.382	0.385
LPIPS Score	0.137	0.155
Throughput	0.020fps	0.143fps
Frame-wise Latency	48.963s	6.815s
Clip-wise Latency	4832.542s	679.725s

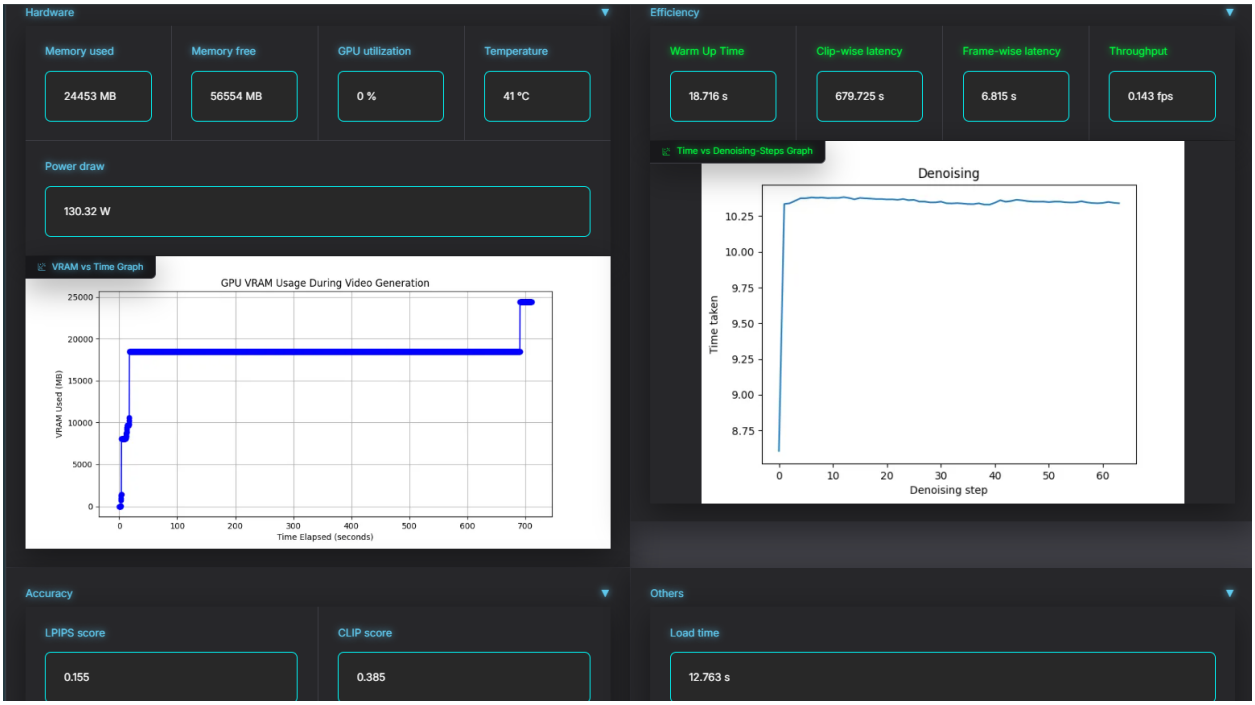

Figure 8: Optimizations Performance Metrics - UI Screenshot

2.9 Quant + DPM + Attention Slicing + Operator Fusion

Metrics	None (Baseline)	Quantization + DPM + Attention Slicing + Operator Fusion
Load Time	14.213s	25.392s
Warm up time	83.098s	182.683s
MAX VRAM	56771MB	31449MB
CLIP Score	0.382	0.384
LPIPS Score	0.137	0.156
Throughput	0.020fps	0.124fps
Frame-wise Latency	48.963s	6.187s
Clip-wise Latency	4832.542s	782.836s

Figure 9: Optimizations Performance Metrics - UI Screenshot

2.10 Quantization + DPM + LoRA + Operator Fusion

Metrics	None (Baseline)	Quantization + DPM + LoRA + Operator Fusion
Load Time	14.213s	12.978s
Warm up time	83.098s	221.212s
MAX VRAM	56771MB	31547MB
CLIP Score	0.382	0.389
LPIPS Score	0.137	0.140
Throughput	0.020fps	0.119fps
Frame-wise Latency	48.963s	6.143s
Clip-wise Latency	4832.542s	817.085s

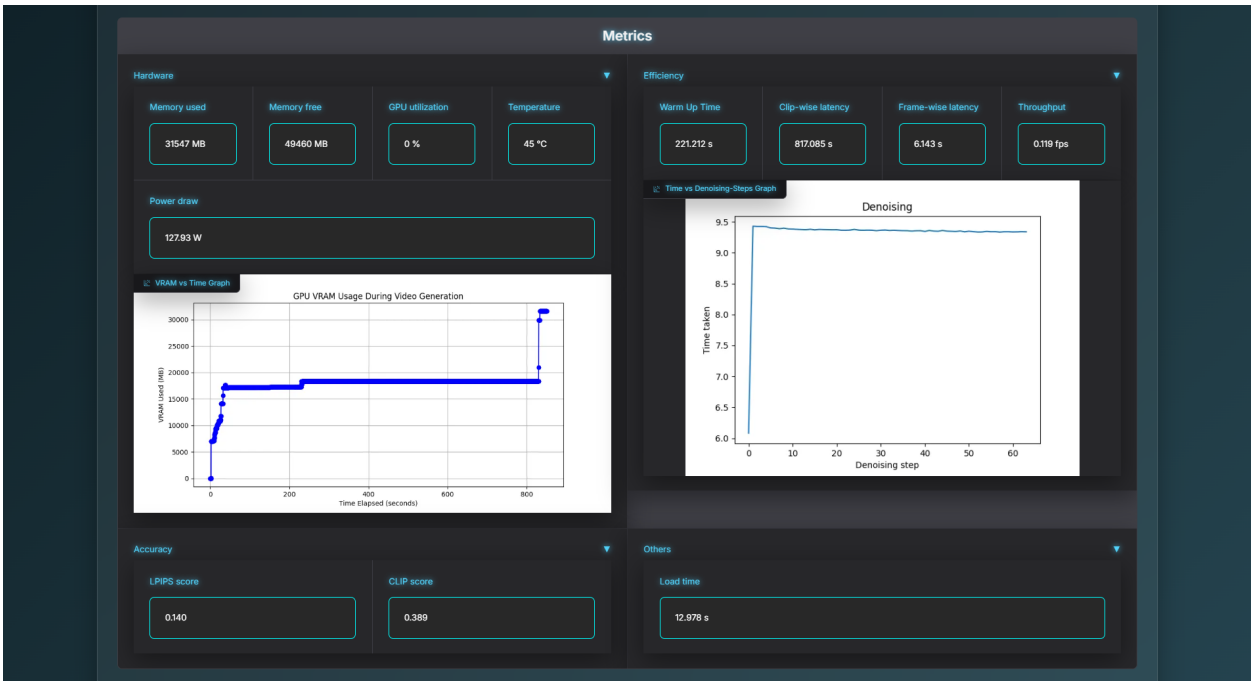

Figure 10: Optimizations Performance Metrics - UI Screenshot

2.11 Quantization + DPM + Operator Fusion

Metrics	None (Baseline)	Quantization + DPM + Operator Fusion
Load Time	14.213s	50.720s
Warm up time	83.098s	209.551s
MAX VRAM	56771MB	31435MB
CLIP Score	0.382	0.384
LPIPS Score	0.137	0.156
Throughput	0.020fps	0.120fps
Frame-wise Latency	48.963s	6.187s
Clip-wise Latency	4832.542s	809.729s

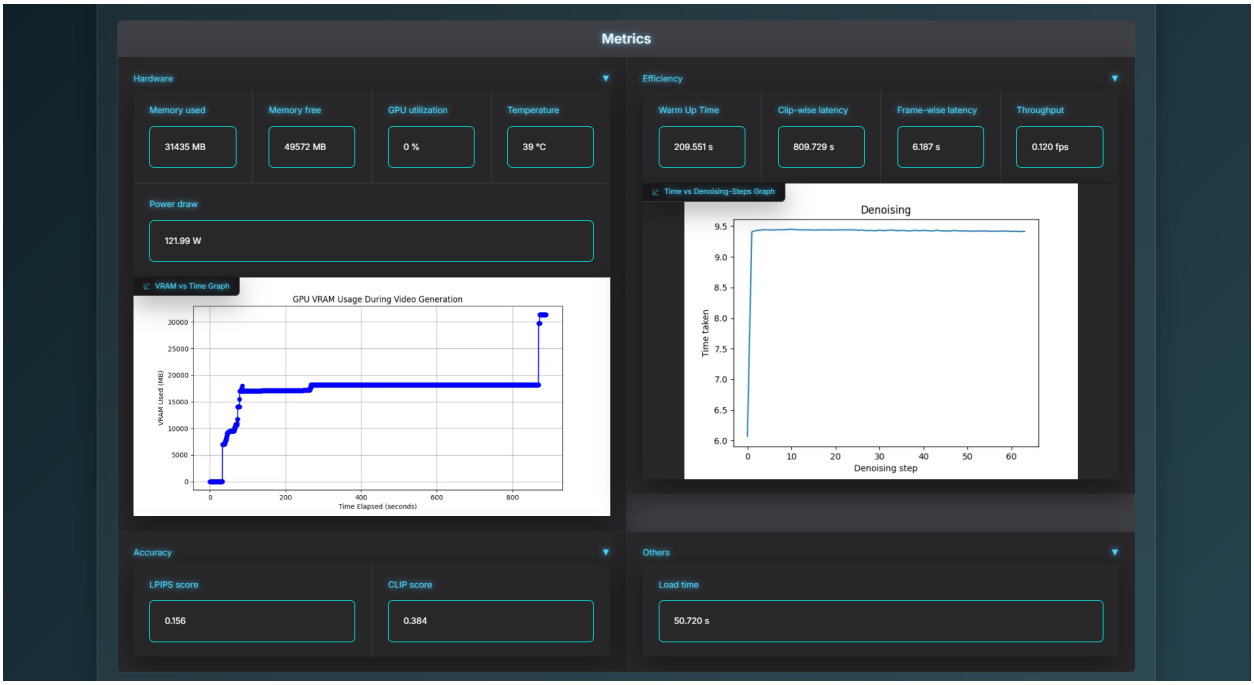

Figure 11: Optimizations Performance Metrics - UI Screenshot

2.12 Best

Metrics	None (Baseline)	Best
Load Time	14.213s	11.217s
Warm up time	83.098s	18.072s
MAX VRAM	56771MB	24553MB
CLIP Score	0.382	0.391
LPIPS Score	0.137	0.141
Throughput	0.020fps	0.144fps
Frame-wise Latency	48.963s	6.779s
Clip-wise Latency	4832.542s	675.647s

Table 2: Performance metrics recorded before and after optimizations

Figure 12: Best Performance Metrics - UI Screenshot

3 Conclusion

- After evaluating all the benchmarks, we concluded that provided enough vram (25-30 GB), it is best to use a combination of Quantization, LoRA, and Attention Slicing when generating videos from individual prompts.
- If the GPU VRAM is limiting factor, using CPU offloading will enable the model to generate output, although the inference speed is a bit slower.
- If the prompts are required to be generated in batches, using Operator Fusion will fasten the inference speed.
- Overall, we achieved approximately 7x speed boost over the baseline model in the best mode, with similar quality (similar CLIP and LPIPS scores).