The video of these benchmarks:

https://www.youtube.com/watch?v=lgP1LNnaUaQ

The gist file of the video:

https://gist.github.com/FurkanGozukara/63e1ef499b5f0c16882a5a202c37f33f

1. RTX 3090 compared to RTX 3060 at every test.
2. Installation of GPU and PSU takes about 1 hour.
3. If you encounter a booting problem, remove every removable part and try 1 by 1. I figured out that 1 of the older SATA HDD was preventing motherboard booting – took 2 hours to find it.
4. Stable Diffusion text to image, OpenAI Whisper speech to text, FFmpeg and Davinci Resolve video rendering are tested.
5. Testing PC features: Core i7 10700F CPU, 16 cores, working at 4.5GHZ, 64 GB RAM memory, NMVe SSD.
6. RTX 3060 was working on PCI-E version 3 but only x4 – 4 lanes. It didn't cause any negative performance issues. Verified with GPU-Z bus usage.
7. Whisper tutorial video: https://youtu.be/msj3wuYf3d8
8. Used Python version for benchmarks is 3.10.8.
9. How to install Python and Automatic1111 Web UI tutorial: https://youtu.be/AZg6vzWHOTA
10. Whisper GitHub: https://github.com/openai/whisper
11. At minute 13:13 how to install latest cuDNN DLL files shown
12. The OBS studio screen recording with Nvidia Broadcast also uses a lot of GPU. So actual speed tests are done when both application were closed.

Whisper Speech to Text Transcription Benchmarks Summaries

1. Torch 1.13 or Torch 2 doesn't make significant difference in terms of speed or the latest cuDNN DLL file.
2. RTX 3090 can process 2.7 seconds speech data in 1 second. So speed is 2.7
3. RTX 3060 can process 1.78 seconds speech data in 1 second. So speed is 1.78
4. Remember that I have used large model v1, best of 10 and beam size 10. Best of 10 and beam size 10 makes significant difference in speed.
5. When RTX 3090 used alone the PC case used about 500 watts per hour. The board power draw was 400 watts for RTX 3090
6. RTX 3060 board power draw was about 140 watts.
7. When both cards did transcription at the same time, peak watt usage was 680 watts.

Davinci Resolve 4k Video Rendering Benchmarks Summaries

1. When only RTX 3090 were working entire system used about 250 watts
2. RTX 3090 rendered 4k 40 minutes clip with 68 minutes. Original video was 42 Mbps so it was a huge video. Speed is 0.58 that means it processed 0.58 seconds data in 1 second.
3. Unfortunately, there is a bug in Davinci Resolve software. So even though I have set RTX 3060 to be used, it still used RTX 3090 for video rendering task. Therefore, I am were not able to compare both cards in this benchmark.

4. No effects were used for Davinci Resolve clip rendering.

FFmpeg 8k Video Rendering Benchmarks Summaries

1. Check video minute 23 for how to install and set as default FFmpeg version that supports GPU – CUDA video processing.
2. RTX 3060 almost performed as RTX 3090. This is probably due to video encoding software capabilities of the FFmpeg and perhaps due to video encoding hardware of the both cards. I also didn't notice such speed increase in Davinci Resolve when compared to my previous experience after installing the new RTX 3090.
3. RTX 3060 board power draw is about 48 watts
4. The speed of RTX 3060 was 0.259. So, for every 1 second, 0.259 second of the video file was rendered. The process was upscaling 4K video into 8K with bitrate of 46 Mbps.
5. RTX 3090 board power draw is about 128 watts
6. The speed of RTX 3060 was 0.278. So, for every 1 second, 0.278 second of the video file was rendered. The process was upscaling 4K video into 8K with bitrate of 46 Mbps.

Stable Diffusion Automatic1111 Web UI Benchmarks Summaries

1. PCI-E x4 was not a limiting factor again for RTX 3060
2. Both cards were working 100% even when simultaneously used
3. Watt usage was about 750-800 watts when both cards were fully used

The results of the tests are like below

* pytorch 1.13 , xformers: 0.0.18.dev494 - cudnn 8500 *

512x512 image generation

1. rtx 3090 - 14.61 it/s - 453 watt system usage
2. rtx 3060 - 7.07 it/s - 279 watt system usage

1216x1216 with hires fix image generation

1. when both cards working system watt usage 750
2. rtx 3090 - 6.21 s/it - second pass when doing high res - 546 watt system usage
3. rtx 3060 - 18.28 s/it - second pass when doing high res - 281 watt system usage

DreamBooth Training

1. rtx 3090 - 3.03 it/s - 400 watt system usage
2. rtx 3060 - 2.20 it/s - 288 watt system usage

* pytorch 2 , xformers: 0.0.18 - cudnn 8700 *

512x512 image generation

1. rtx 3090 - 15.83 it/s - 495 watt system usage
2. rtx 3060 - 6.8 it/s - 275 watt system usage

1216x1216 with hires fix image generation

1. rtx 3090 - 6.04 s/it - second pass when doing high res
2. rtx 3060 - 17.37 s/it - second pass when doing high res

Dreambooth training

1. rtx 3090 - 3.24 it/s
2. rtx 3060 - 2.27 it/s

* pytorch 2 , --opt-sdp-attention - cudnn 8700 *

512x512 image generation

1. rtx 3090 - 16.39 it/s
2. rtx 3060 - 6.8 it/s

1216x1216 with hires fix image generation

1. throws out of memory error after final step of image generation when producing the image

Dreambooth training

1. rtx 3090 - 2.47 it/s
2. rtx 3060 - 1.82 it/s

* pytorch 2 , --xformers - cudnn 8801 *

1. same speed

* pytorch 2 , --opt-sdp-attention - cudnn 8801 *

1. image generation same speed

1216x1216 with hires fix image generation - still out of memory error

dream booth still same speed

if you want to generate very high resolution you have to not use xformers or --opt-sdp-attention

without xformers and --opt-sdp-attention

1. both cards able to generate 2048x2048 pixel resolution image with
2. rtx 3090 : 6.04 second per it
3. rtx 3060 : 16.57 second per it
4. with xformers max resolution is 1448 x 1448
5. in 1456 pixel you get this error

NansException

A tensor with all NaNs was produced in Unet. This could be either because there's not enough precision to represent the picture, or because your video card does not support half type. Try setting the "Upcast cross attention layer to float32" option in Settings > Stable Diffusion or using the --no-half commandline argument to fix this. Use --disable-nan-check commandline argument to disable this check

RTX 3060 can generate 2048x2048 pixels image without xFormers or opt-sdp-attention.

Conclusions

1. CPU was never a limiting factor during my tests. Core i7 10700F performed very well even when both cards were working simultaneously.
2. It doesn’t worth to upgrade from RTX 3060 to RTX 3090 for only Whisper transcription or video rendering tasks. However, with Stable Diffusion and

high resolution, while RTX 3060 gets 1x speed, RTX 3090 gets 3x speed. So the card shines truly when working with most up to date machine learning algorithms and models. Moreover, 24 GB will help you tremendously in many of the high VRAM demanding machine learning applications and models.

3. However Stable Diffusion Automatic1111 web UI is extremely well programmed at back end and able to 100% utilize both of the cards with 450 watt for RTX 3090 and 150 watt for RTX 3060 – 100% GPU load.
4. For Stable Diffusion
5. RTX 3090 was able to reach 16.39 it/s and RTX 3060 7.07 it/s
6. To be able to use second card make your command line arguments like below
7. set VENV_DIR=
8. set CUDA_VISIBLE_DEVICES=1
9. set COMMANDLINE_ARGS=--xformers
10. Solution for NansException : A tensor with all NaNs was produced in Unet When using Stable Diffusion
11. The best solution is not using any optimizer such as xFormers or opt-sdp-attention of Torch 2.
12. Other than that, using full model instead of pruned, using ckpt instead of safetensors and using the best VAE file 840000 may help you to get rid of this annoying error along with --no-half command line argument.
13. RTX 3090 is able to draw up to 450 watts from your PSU. So I suggest to get at least 700 watts good PSU for using RTX 3090 alone. For using both RTX 3060 and RTX 3090 simultaneously you probably should get 850-900 watt good PSU.
14. PCI express 4x lanes is not a problem. It doesn't reduce your GPU speed.

15. Torch version 2 definitely makes improvement so should be used.
16. --opt-sdp-attention increases image generation speed for RTX 3090 however it is slowing down training speed. xFormers is still looking like the best optimizer overall.