Github Link :

https://github.com/PrashantNeelwan/GenAI/tree/main/assignment_5

(Model 1: Stable Diffusion XL)

1 Introduction (in SDXL section)

Stable Diffusion XL (SDXL), introduced by Stability AI, is one of the most notable developments of text-to-image diffusion models. SDXL is trained at higher fidelity and with complex prompts, and can be scaled to other resolutions, compared to the previous diffusion architectures.

The SDXL is selected in this report since it represents the state of the art in terms of image synthesis that creates a good contrast to the multimodal reasoning abilities of Qwen2-VL (which will be discussed later).

2 Technical Overview of SDXL

Architecture:

- SDXL, which is based on the diffusion framework, produces images as a result of step-by-step denoising of latent representations.
- Adds a larger context window, a larger UNet backbone and a variational autoencoder (VAE) that is more refined to provide higher quality outputs.
- Enables slice based attention and slice based VAE, making good use of GPU memory.

Innovations:

- Base model which is trained on high quality datasets.
- Ability to use a variety of different schedulers (Euler, DPM, etc.) to trade quality and speed.
- Guidance scaling enhances timely alignment without being unrealistic.

Comparative Advantages:

- Generates resolution images without the need to upscale them.
- Faster latency and schedulers are optimized.

- Expansive creative control by negative prompts and guidance tuning.

A_golden_retriever_puppy_sitting_in_a_field_of_flowers_sunny__dpm_28st_512x512

A_cozy_wooden_cabin_in_snowy_mountains_at_sunset_warm_lights__euler_16st_512x512

3. Industry Impact

SDXL is transforming several artistic and business sectors:

- Advertising & Media: Generating high-quality content on a large scale, automating and saving time and money on the side of the agencies.
- Entertainment/Gaming: Fast conceptual development of prototyping and simulation environments.
- Design & E-commerce: Design product catalogs and user experiences to be customized.
- Intellectual property and ethical application of generated media, and the threat of misinformation through realistic synthetic imagery have also become an issue of industry discussion with the advent of SDXL.

4 Potential Applications

- Creative Arts: SDXL is used by artists and hobbyists to provide illustrations, book covers or album art.
- Education: Showing abstract concepts to students.
- Product Development: Pre-manufacturing design prototypes.
- Personalization: Customized wallpapers, avatars and visual narrations.

```
=== Fastest by prompt ===
                                                prompt  height  width  steps  \
2   A golden retriever puppy sitting in a field of...     512    512     16
19  A cozy wooden cabin in snowy mountains at suns...     512    512     16

    guidance scheduler  latency_s  \
2        5.5     euler   3.894388
19       5.5       dpm   4.126153

                                           image_path
2   sdxl_bench/A_golden_retriever_puppy_sitting_in...
19  sdxl_bench/A_cozy_wooden_cabin_in_snowy_mounta...

=== Highest steps & size (quality-biased) by prompt ===
                                                prompt  height  width  steps  \
12  A golden retriever puppy sitting in a field of...     768    768     28
31  A cozy wooden cabin in snowy mountains at suns...     768    768     28

    guidance scheduler  latency_s  \
12       3.0     euler  16.810893
31       5.5       dpm  17.040551
```

5 Summary (for SDXL section)

SDXL is an advanced image generation model that is both fast, high quality, and scalable. The fact that it can be used to generate photorealistic or artistic images based on textual prompts shows how diffusion models can be used to change the processes of creativity. Although its usage opens up new possibilities in various sectors, it requires a responsible approach toward the ethical, legal and social implications.

Model 2: Qwen2-VL (Text and Vision Large Model).

1 Introduction

The Qwen2-VL-7B-Instruct is a large language model of the Qwen team at Alibaba built in multimodal fashion. In contrast to text-only models, Qwen2-VL is text and visual understanding oriented, allowing reasoning, summarization and question answering of various types of inputs.
This project experimented with the model in text-only mode to demonstrate its reasoning, summarization and safety-checking capabilities, as a contrast to SDXL image synthesis capabilities.

2.Technical Overview

Architecture:

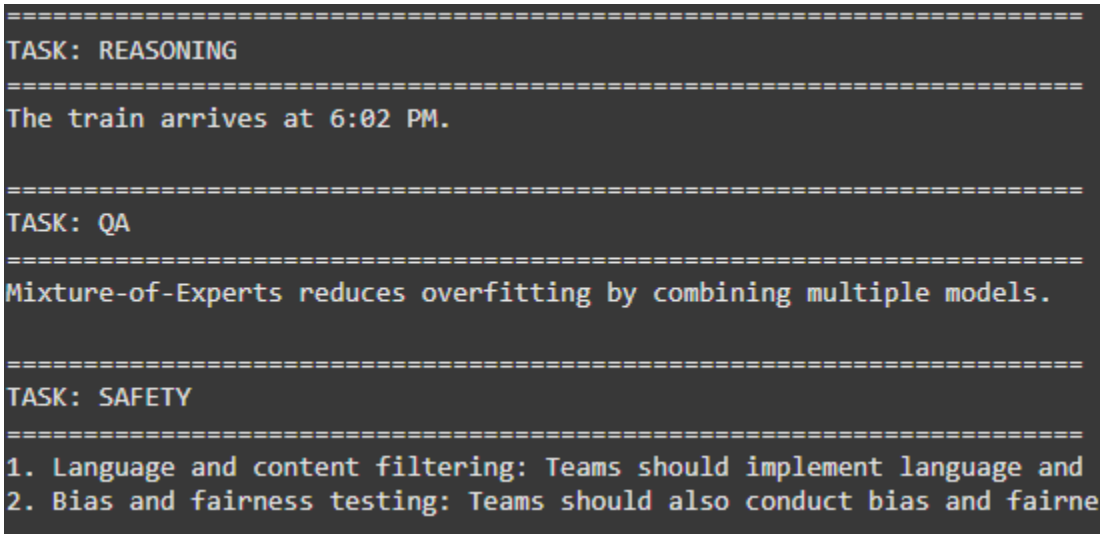
- According to transformer decoder stack that is optimized to multimodal alignment.
- Uses encoders of vision together with LLM reasoning layers, although also works in text-only mode.

Innovations:

- Fine-tuned to be instructed after, and in multimodal tasks.
- Effective inference as well as mixed-precision (FP16) and dynamic token throughput monitoring.
- Delivers general purpose intelligence, as opposed to SDXL which is quite specialized in image synthesis.

Comparative Advantages:

- Practices sophisticated reasoning (e.g. time calculations).
- Favors abstractive summarizing and writing in a stylistically controlled manner.
- Combined safety guardrails through task specific outputs.



3 Industry Impact

- The design of Qwen2-VL is designed to be multimodal:
- Research & Education: Automated summarization, concept explanation and cross modal tutoring.
- Enterprise: Document processing, meeting summarization and compliance auditing.
- Productivity Tools: Customer support agents, personal assistants and safety filters to implement AI.
- Its capability to work in both text and vision space puts it on the same level as OpenAI GPT-4V and Google Gemini, and it is taking the industry towards actual multimodal AI solutions.

4 Potential Applications

- Knowledge Work Rapid deposition of research papers or technical notes.
- Time dependent Reasoning: Calendar and time schedule assistants.
- Content Moderation: Auto safety and compliance.
- Accessibility: Multimodal support of visually or language impaired users.

	A	B	C	D	E	F	G	H
1	task	model	text	latency_s	input_tok	output_to	throughpu	quality_proxy
2	reasoning	Qwen2-VL-7B-Instruct	The train arrives at 6:02 PM.	13.87674	46	11	0.792693	1
3	reasoning	Qwen2-VL-7B-Instruct	The meeting will end at 11:55 AM.	13.97133	44	13	0.930477	1
4	reasoning	Qwen2-VL-7B-Instruct	1:10 PM	6.137382	41	5	0.81468	1
5	summariz	Qwen2-VL-7B-Instruct	FP8 mixed-precision training reduces memory bandwidth and accele	16.86779	41	16	0.948554	1
6	qa	Qwen2-VL-7B-Instruct	Mixture-of-Experts reduces overfitting by combining multiple model	14.94405	40	14	0.936827	0
7	rewrite	Qwen2-VL-7B-Instruct	"We significantly accelerated our computations by utilizing FP8 comp	16.08241	39	15	0.932696	1
8	safety	Qwen2-VL-7B-Instruct	1. Language and content filtering: Teams should implement	127.8042	35	128	1.001532	1

5 Summary (for Qwen2-VL section)

Qwen2-VL shows how modern LLMs can be used in a general-purpose way, with a variety of tasks ranging through logical reasoning to safety checks. It is more flexible than special architectures such as SDXL, which exemplify the various ways in which different architectures apply to different fields: creative visual synthesis and cognitive and multimodal reasoning.