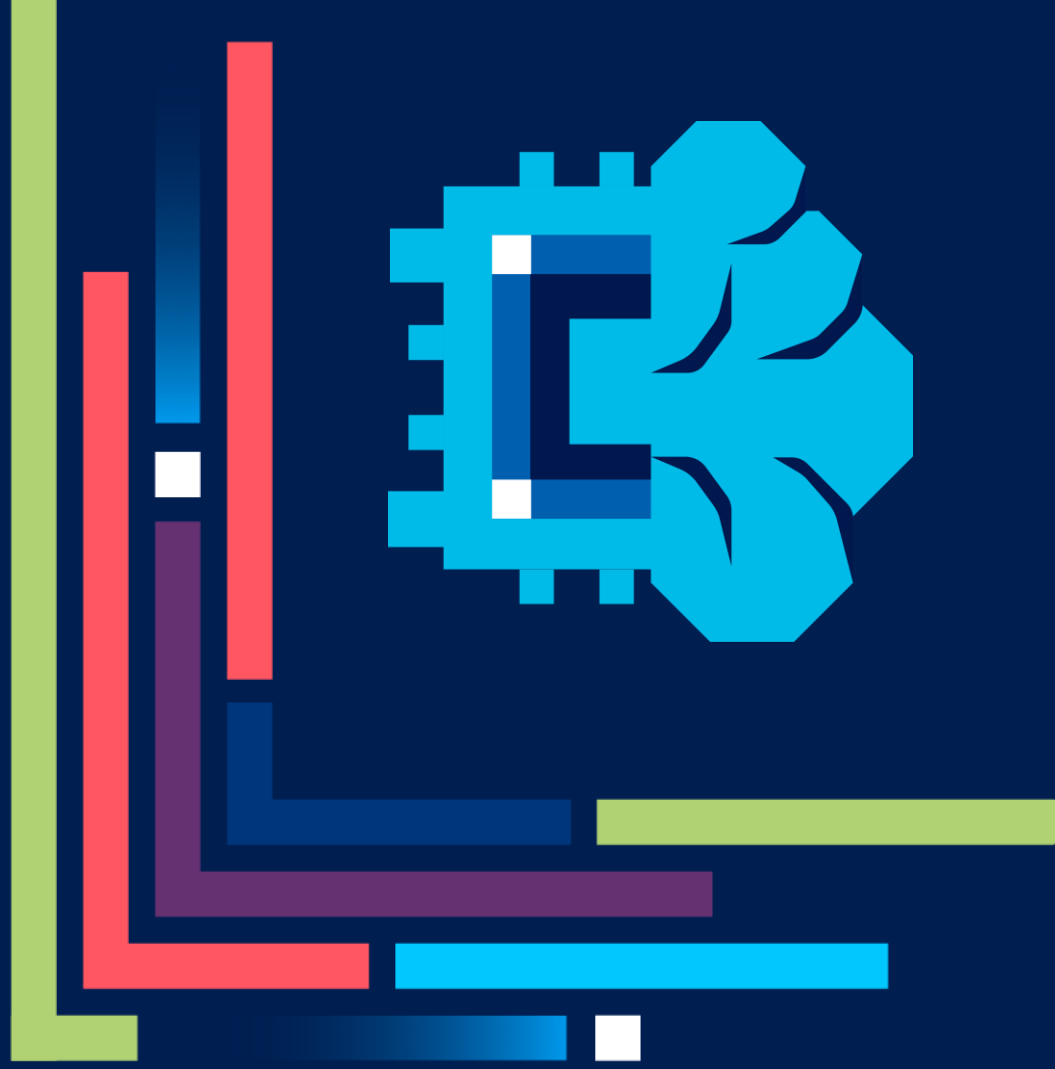
intel ai summit
英特爾AI科技論壇

Bringing AI Everywhere



Intel ARC ™ A770
AI application development & deployment vehicle

Ely Zeng曾智暉
Discrete Graphics Software Engineer
March 27th, 2024

Benefits of Running AI Locally

Increased Privacy

No need to transmit sensitive data over the internet / third parties / tech corporations

Lower Latency

No need for data to be transmitted across the planet

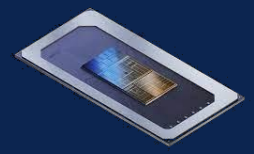
Increased Control

Flexibility on what you can and cannot run

intel Ai summit

How AI Runs on Intel Hardware

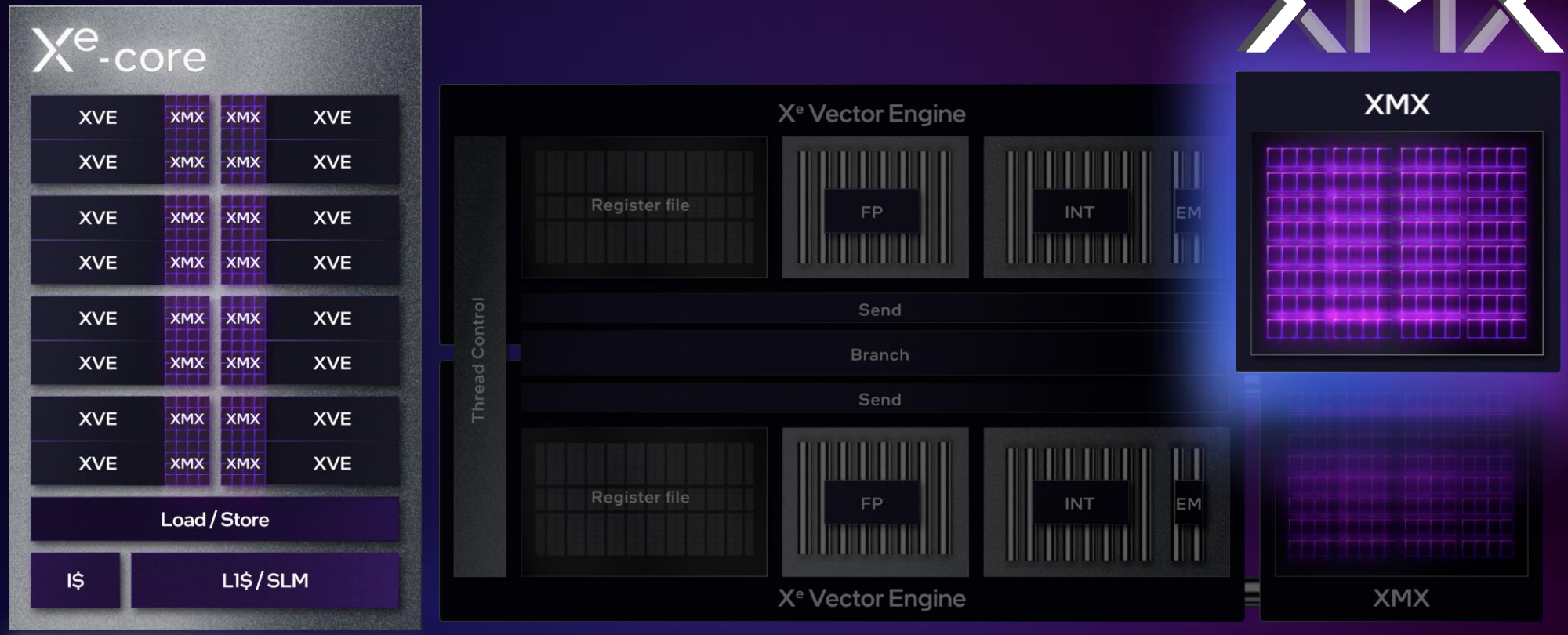
3 main pillars of the Intel AI PC

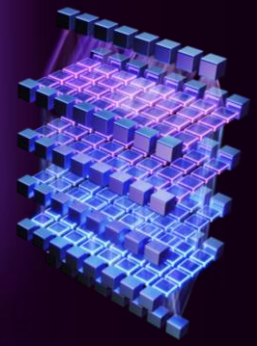
	Ideal for	Product Range
CPU 	Ideal for quick, lightweight AI tasks such as chat bots	Intel® Core™ processors
NPU 	Efficiently runs sustained AI workloads as webcam filters	Intel® Core™ Ultra processors
GPU & Built-in GPU 	Runs heavy AI workloads such as generative content	Intel® Core™ Ultra processors Intel® Arc™ A-Series Graphics

What this presentation will cover

intel Ai summit

Intel® Arc™ Graphics have
Dedicated Hardware for AI Acceleration





AI development & deployment on Intel® Arc™ A770

Gaming

Intel® XeSS AI upscaling

Get the latest visual technologies like AI-accelerated XeSS image upscaling for faster gaming with high image quality

Content Creation

Faster Creation Enhancements

Execute post-processing, upscale images, 3D rendering and edit videos quickly easily with powerful XMX AI capabilities.

Generative Content

Generative Enhancements

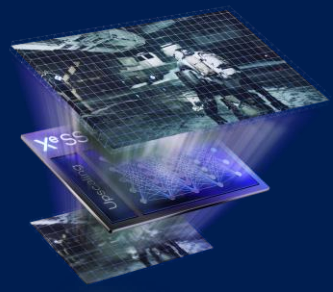
Generative content with Intel XMX AI engines . Example run text-to-image content generation locally on your own system.

intel Ai summit

Intel® XeSS AI upscaling

intel Ai summit

XeSS Overview



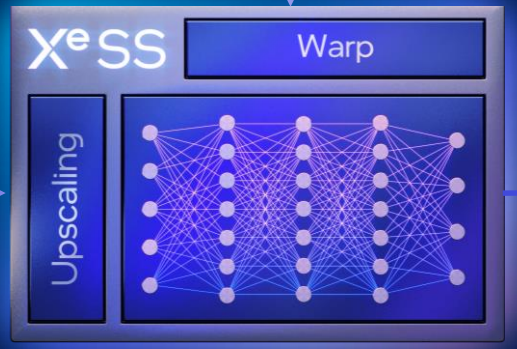
Jitter

Raster, Lighting & Pre-Effects

Motion Vectors

Low-Res Frame Render

Frame History



Super Sampled Render

Post-Effects

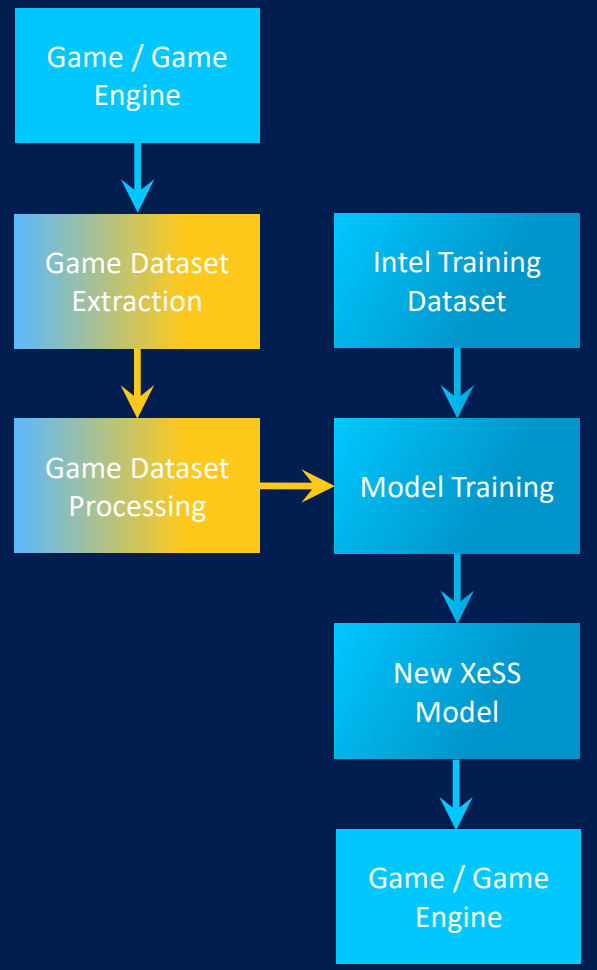
Use of AI In Gaming: Higher Quality Upscaling

intel Ai summit

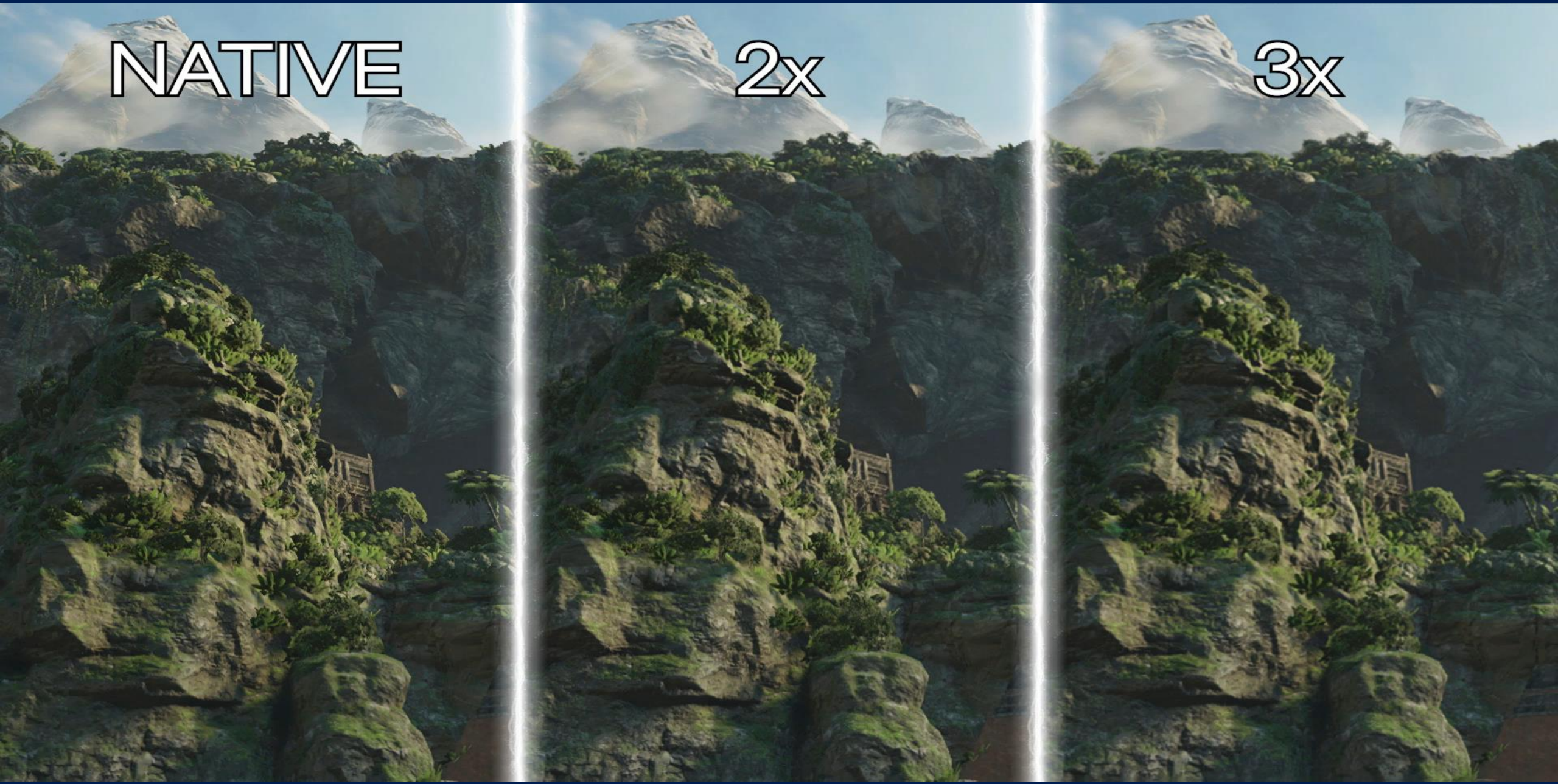
Improved AI models with data set extraction

Dataset extraction API

- Integrate with the game engine to capture training data
- Captures:
  - Jitter offsets
  - Low-res jittered HDR color
  - Low-res depth buffer
  - Low-res motion vectors / hi-res dilated motion vectors
  - Optional parameters
- The game/engine:
  - Freezes timestamps and suspends animations, dynamic effects, etc
  - Captures jittered data over frozen scenes



3x Scaling Sneak Peek



Developer Resources

Add XeSS support to your game

https://github.com/intel/xess
https://github.com/GameTechDev/XeSSUnrealPlugin

Check our developer website for all the latest info

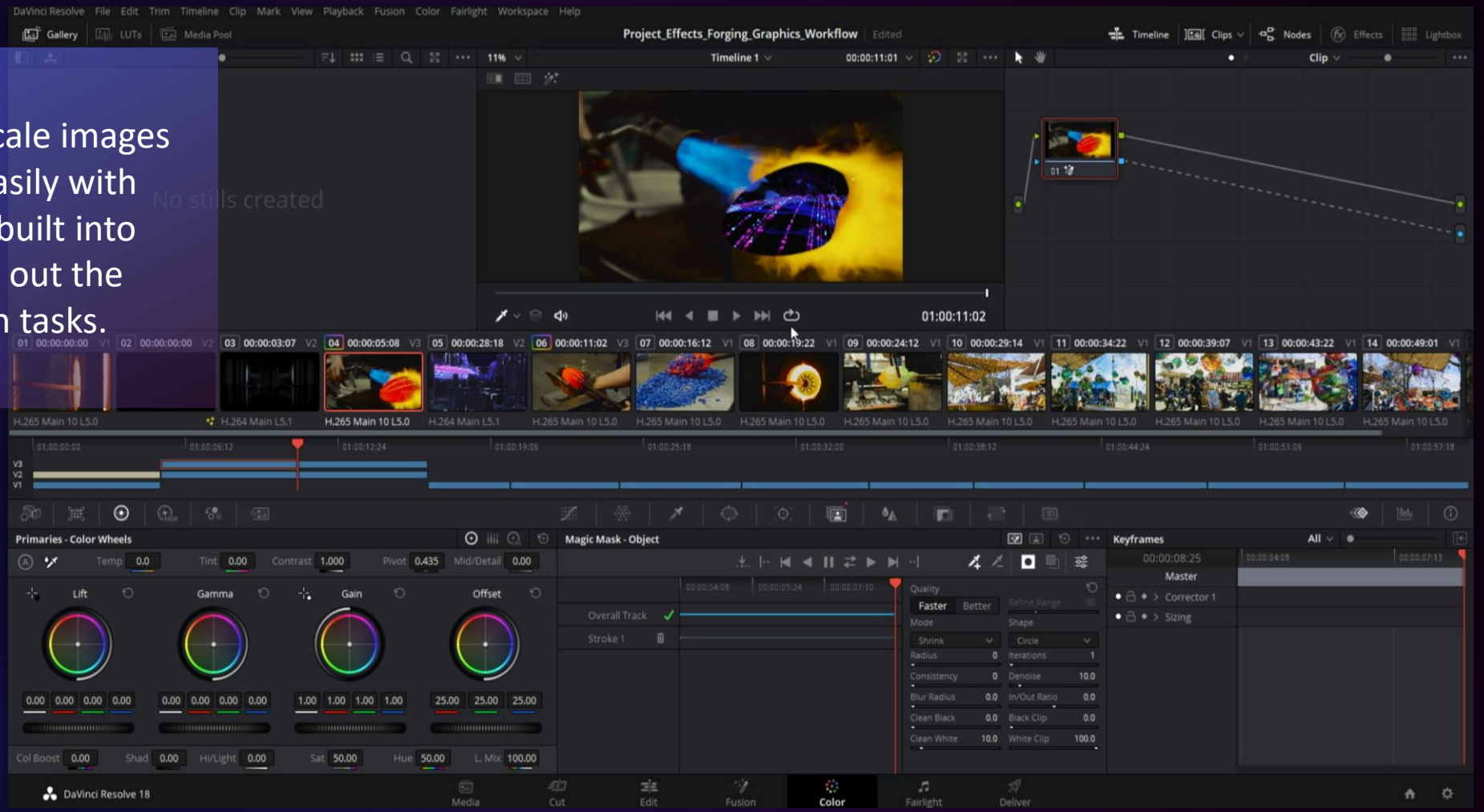
http://gamedev.intel.com
https://github.com/GameTechDev
gamedevtech@intel.com

Use of AI in Creation

intel Ai summit

Use of AI In Creation: Faster Enhancements

Execute post-processing, upscale images and edit videos quickly and easily with powerful XMX AI capabilities built into Intel® Arc™ graphics, bringing out the best in AI accelerated creation tasks.



Use of AI In Creation: 3D Rendering

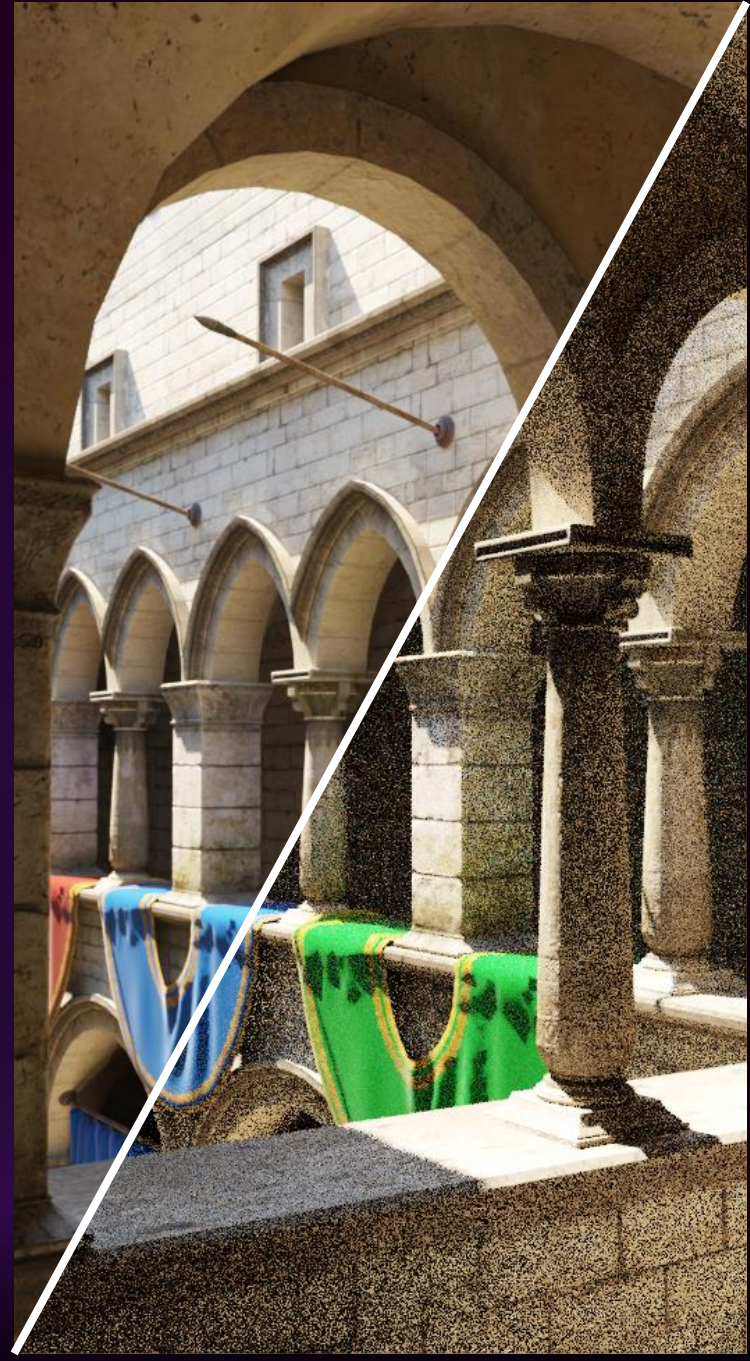
Intel® Open Image Denoise (OIDN)

Denoising library for ray traced images that easily integrates into existing popular rendering applications, featuring:

- High-quality deep learning based denoising filters
- Final frames and baked lightmaps
- Suitable for interactive and offline rendering

www.openimagedenoise.org/

Open Source under Apache 2.0 license



Scene courtesy of Frank Meinl, downloaded from Morgan McGuire's Computer Graphics Archive.

Development Opportunities Examples

Smart & Easy Photo Editor

- Upscale photos
- Change photo ratio without cropping
- Remove elements from photos

Fooocus: example of easy and free to use photo editing

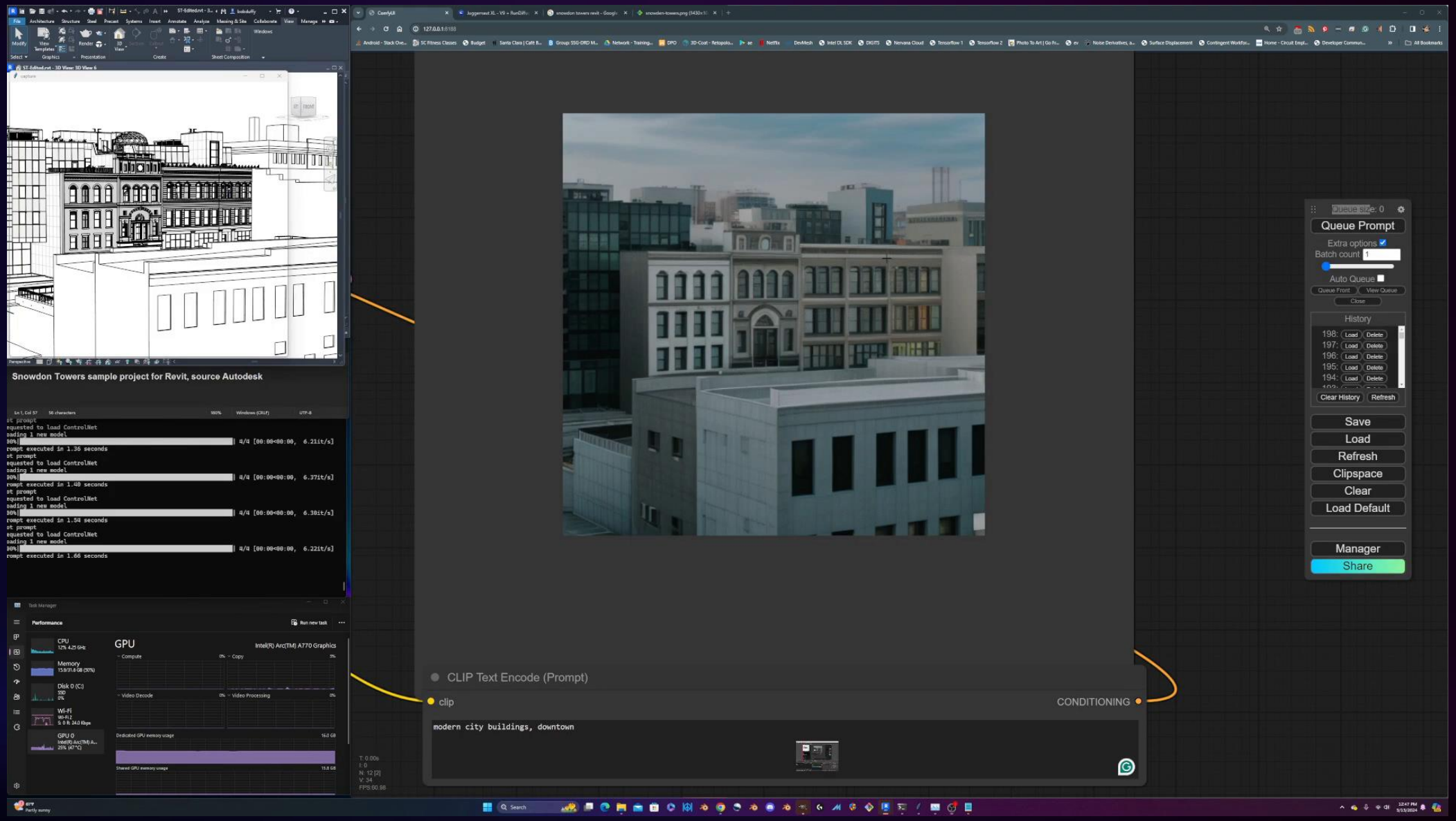
AI Design Visualizer

- Beautify, clean up faces, and objects
- Quickly render architectural buildings
- Visualize and ideate interior design ideas

ComfyUI: example of quick renders of complex CAD files

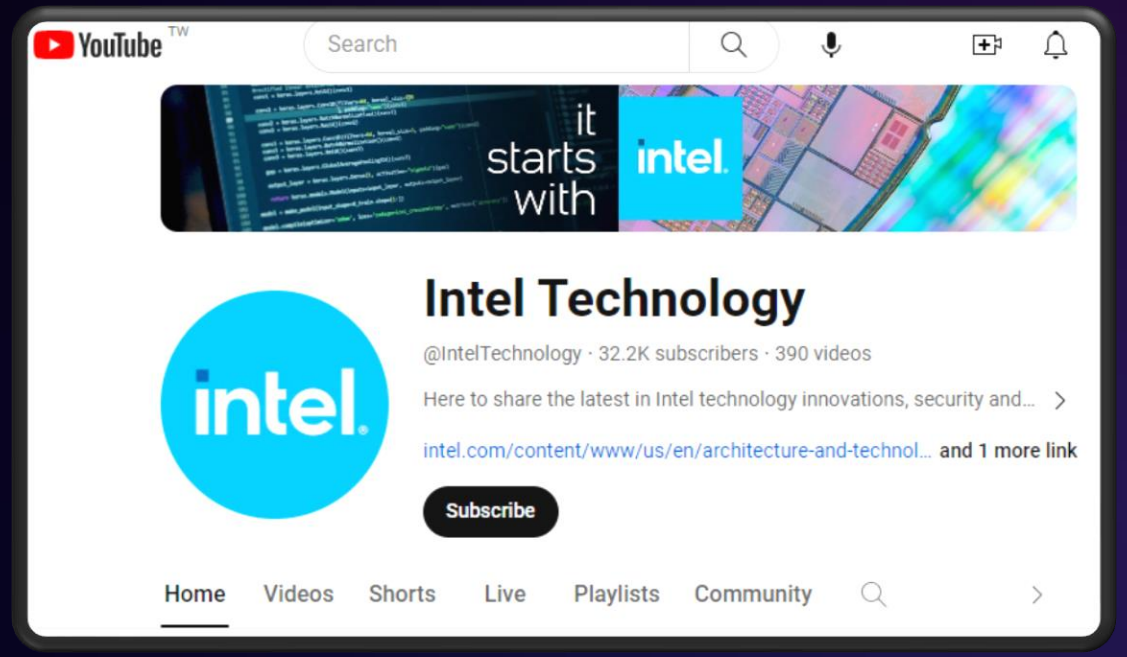
intel Ai summit

AI Rendering of CAD files using ComfyUI



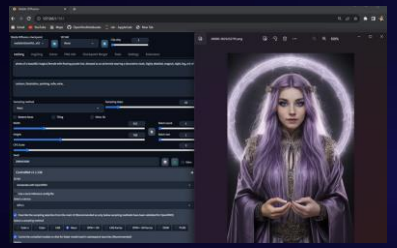
intel Ai summit

More tutorials @ youtube.com/@IntelTechnology



AI Image Generation using Intel® Arc™ Pro A40 GPUs
For this workstation proof-of-concept, we're using a simple wireframe preview in Revit from Autodesk as real-time input for stable diffusion image generation. All on an #IntelArc Pro A40 GPU, with great results.
https://www.youtube.com/watch?v=i6FvA2YaEGI

A1111 WebUI with OpenVINO™ Toolkit for Intel® Arc™ GPUs (video)
Watch Intel's Bob Duffy walk through various AI image generation techniques using the A1111 WebUI, running on an #IntelArc GPU and powered by the #OpenVINO toolkit.
https://www.youtube.com/watch?v=5X0RmlH6JI4



Three Ways to Generate AI Art Using Intel® Arc™ GPUs (Article)
AI image generation tools now allow for higher control, iteration, and custom data sets, all possible on your PC and running on Intel Arc Graphics.
https://game.intel.com/story/intel-arc-graphics-generative-ai-art/

Using Stable Diffusion on GIMP with Intel Arc Graphics (video)
Learn to use your Intel Arc GPU to generate AI text to image art. The technique is accomplished with the OpenVINO toolkit and Stable Diffusion. Intel's Bob Duffy walks you through the process in this demo and tutorial.
https://www.youtube.com/watch?v=q8xJjPBjqso

Enable repository to support Intel Arc products

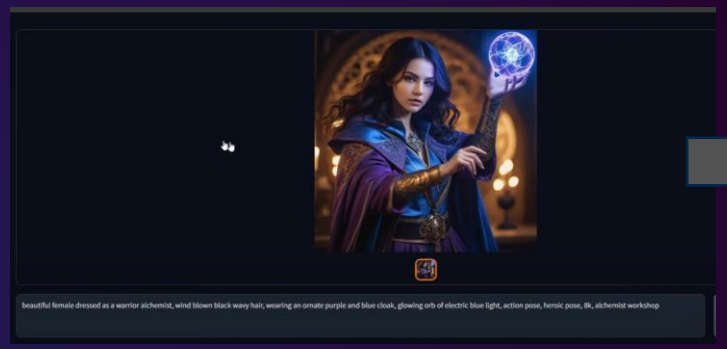
How: Forked repos with IPEX (few lines of code + public IPEX for Windows workaround library)

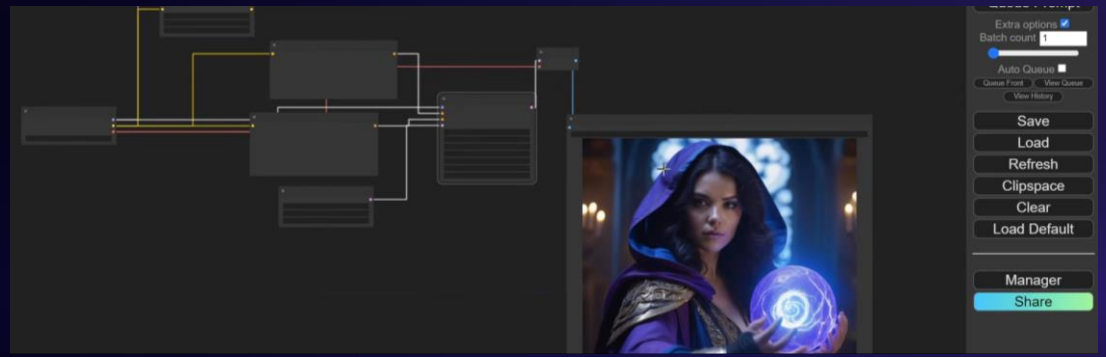
Output:

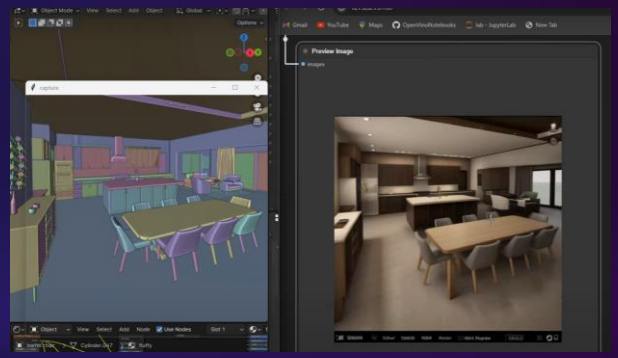
- Powerful unique features in each.
- Designed for different types of users
- Real world use cases with each tool
- Run on Intel Arc discrete GPUs

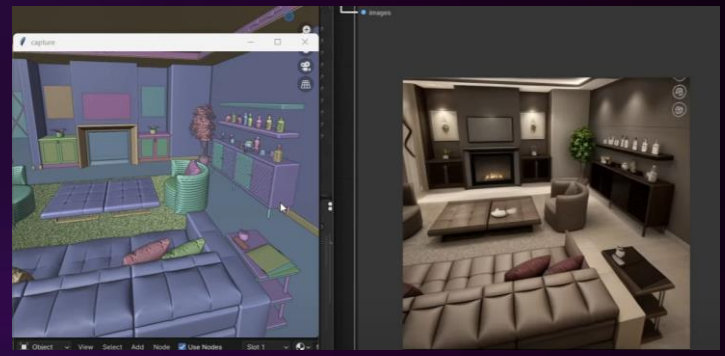
https://youtube.com/watch?v=L0RKVvdIpFA

Fooocus: Simple and smart interface, Powerful image editing, quality output









ComfyUI: Visual programming UI & custom node architecture. 3rd party apps can influence output (Blender used as a control for image generation)

Example implementation (IPEX & OpenVINO)

CUDA

Diffusers

```
from diffusers import StableDiffusionPipeline
import torch

model_id = "runwayml/stable-diffusion-v1-5"
pipe = StableDiffusionPipeline.from_pretrained(model_id, torch_dtype=torch.float1
pipe = pipe.to("cuda")

prompt = "a photo of an astronaut riding a horse on mars"
image = pipe(prompt).images[0]

image.save("astronaut_rides_horse.png")
```

IPEX

```
from diffusers import StableDiffusionPipeline
import torch
import intel_extension_for_pytorch

model_id = "runwayml/stable-diffusion-v1-5"
pipe = StableDiffusionPipeline.from_pretrained(model_id, torch_dtype=torch.float16)
pipe = pipe.to("xpu")

prompt = "a photo of an astronaut riding a horse on mars"
image = pipe(prompt).images[0]

image.save("astronaut_rides_horse.png")
```

Intel Extension for Pytorch (IPEX)

- Run the sample code from huggingface with couple line changes
  1. Setup python env with anaconda
  2. Pip install dependencies
  3. Copy paste sameple code from huggingface
  4. Import intel_extension_for_pytorch
  5. Change pipe.to(“cuda”) to pipe.to(“xpu”)
  6. Run sample code.

intel AI summit

Example implementation (IPEX & Intel ® OpenVINO ™)

```python
from optimum.intel import OVStableDiffusionPipeline
import time
hugging_face_model = "runwayml/stable-diffusion-v1-5"
pipe = OVStableDiffusionPipeline.from_pretrained(hugging_face_model, export=True)
height=512
width=512
batch_size=1
steps=20
pipe.reshape(batch_size=batch_size, height=height, width=width, num_images_per_prompt=batch_size)
# default to GPU.0, specify device id if needed. i.e GPU.1
pipe.to("GPU")
pipe.compile()
prompt = "a majestic lion jumping over a stone at night"
output = pipe(prompt,
              height=height,
              width=width,
              num_inference_steps=steps,
              num_images_per_prompt=batch_size,
              output_type="pil"
              ).images[0]
```

Intel ® OpenVINO ™

Run stable diffusion with pipelines from Optimum

OpenVINO can accelerate stable diffusion pipelines, with Hugging Face Optimum. This provides acceleration and cross-platform support, with an easy conversion from PyTorch to OpenVINO IR format.

With Optimum pipelines, it converts the model automatically

https://huggingface.co/docs/optimum/en/intel/inference

https://github.com/huggingface/optimum-intel

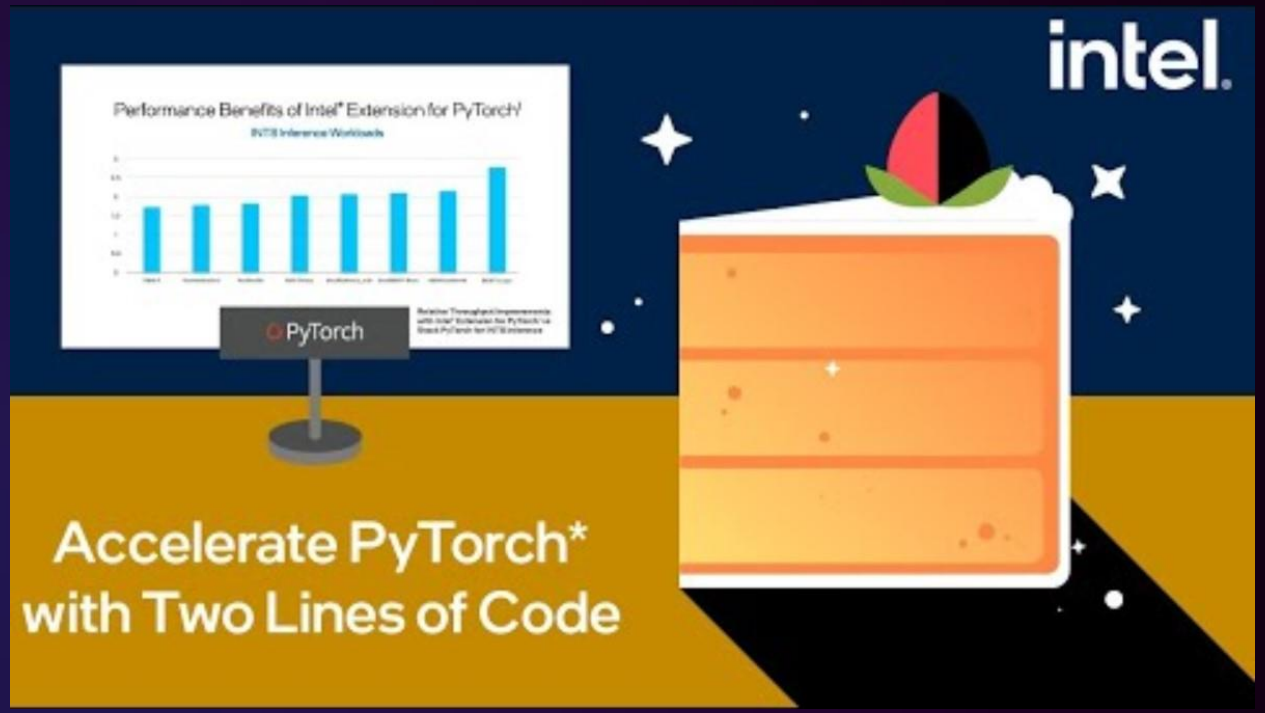
Developer Resources

- Intro to IPEX

  https://www.intel.com/content/www/us/en/developer/articles/technical/introducing-intel-extension-for-pytorch-for-gpus.html

- Deploying IPEX backend

  https://pytorch.org/tutorials/recipes/torch_compile_backend_ipex.html



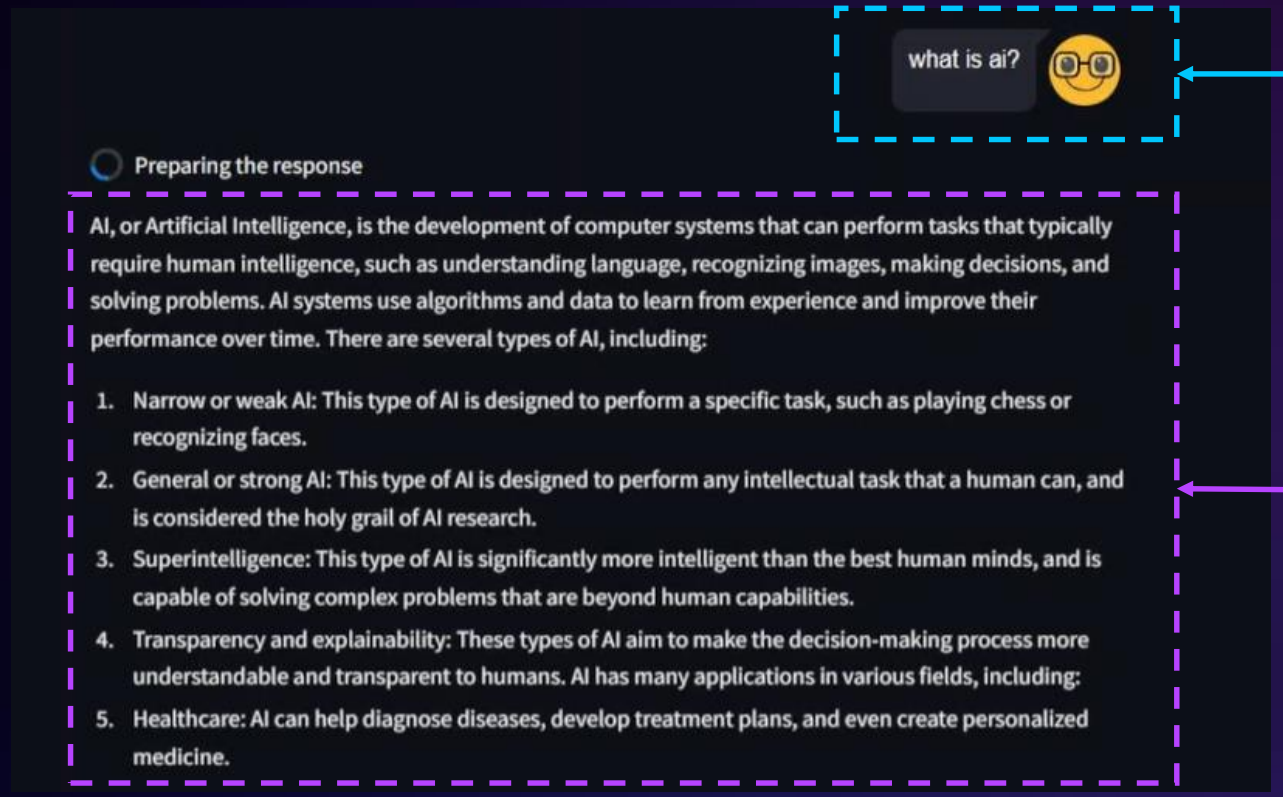
https://www.youtube.com/watch?v=W_rpNvTsE7E

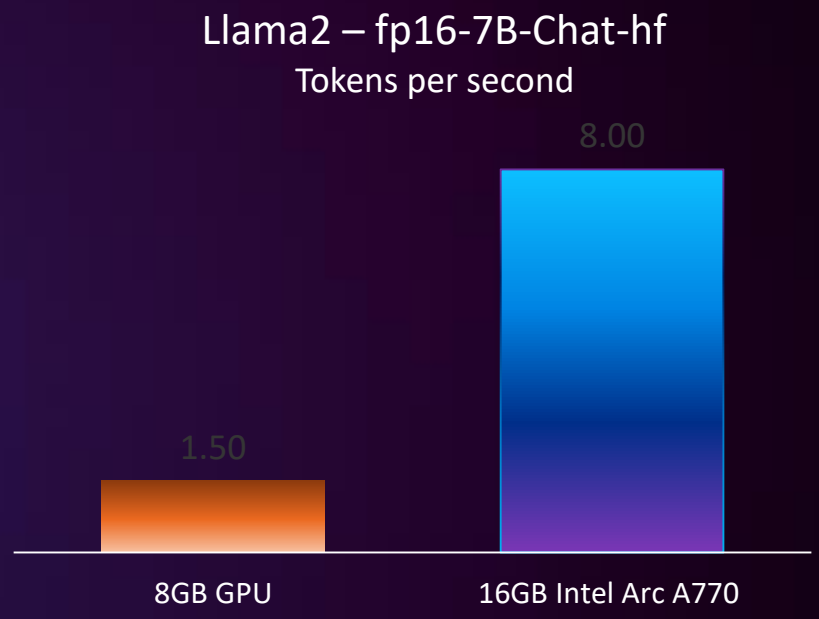
Generative Content

intel Ai summit

The advantage of more than 8 GB VRAM in Gen AI

As tested on Llama 2: offline large language model





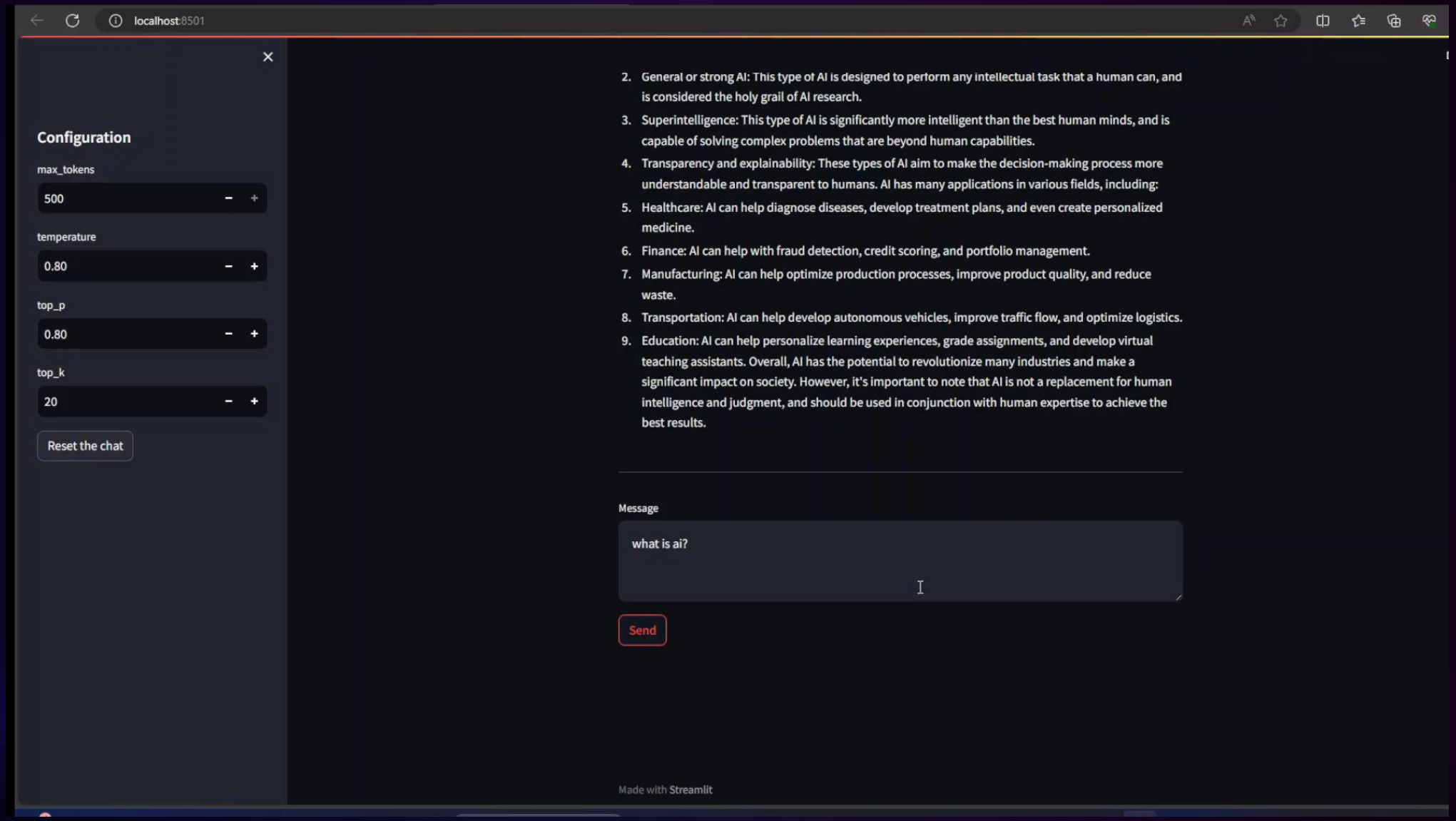
~14GB VRAM required for Llama2-7B fp16 LLM to perform optimally

Less than 14GB VRAM requires extra steps and accuracy may be reduced

intel AI summit

Performance varies by use, configuration and other factors. See backup for full configuration details.

Video Example of Chatbot



Example implementation (IPEX)

Intel Extension for Pytorch (IPEX)

- Run the sample code from huggingface with couple line changes
  1. Setup python env with anaconda
  2. pip install IPEX packages
  3. Copy and paste code example from Huggingface
  4. Import intel_extension_for_pytorch
  5. Change pipe.to(“cuda”) to pipe.to(“xpu”)
  6. Run sample code.

```
# Install transformers from source - only needed for versions <= v4.34
# pip install git+https://github.com/huggingface/transformers.git
# pip install accelerate

import torch
Import intel_extension_for_pytorch
from transformers import pipeline

pipe = pipeline("text-generation", model="TinyLlama/TinyLlama-1.1B-Chat-v1.0", torch_dtype=torch.bfloat16, device_map=“xpu")

# We use the tokenizer's chat template to format each message - see https://huggingface.co/docs/transformers/main/en/chat_templating
messages = [
   {
      "role": "system",
      "content": "You are a friendly chatbot who always responds in the style of a pirate",
   },
   {"role": "user", "content": "How many helicopters can a human eat in one sitting?"},
]
prompt = pipe.tokenizer.apply_chat_template(messages, tokenize=False, add_generation_prompt=True)
```

Different workloads require different engines

Performance

Heavier Workloads

BURST INFERENCE

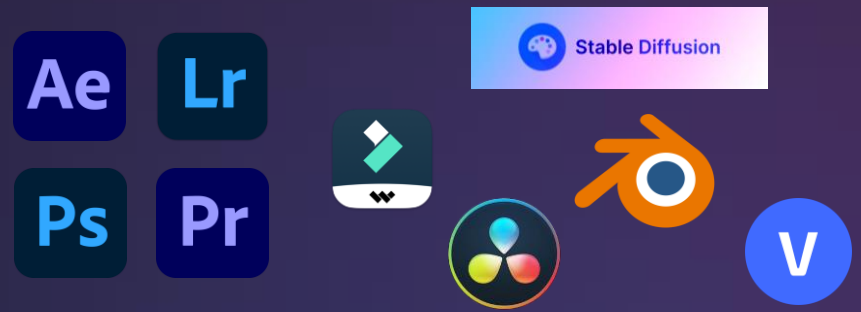


Image generation

Large Language Models

Neural Graphics

Lighter Workloads

CONTINUOUS
LOW-POWER INFERENCE

Image /Video Denoising

Image /Video Upscaling

Background Segmentation

Automated Summarization

Discrete GPU

Built-in GPU

NPU

CPU

Intel Arc Pro, A7, A5, A3, Product AI Related Capability

AI Usage / Arc Graphics	XMX AI Units	Ray Tracing & Xe Engine	Memory bandwidth	Generative Content	AI for Content Creation	AI Gaming of XeSS upscaling
Arc A770 16GB	**512 Units**	**32 Units**	256bit	Supported	Supported	Supported
Arc A750 8GB	**448 Units**	**28 Units**	256bit	Supported	Supported	Supported
Arc A580 8GB	**384 Units**	**24 Units**	256bit	Supported	Supported	Supported
Arc A380 6GB	**128 Units**	**8 Units**	96bit	Supported*	Supported	Supported
Arc A310 4GB	**96 Units**	**6 Units**	64bit	Supported*	Supported	Supported
Arc Pro A60 12GB	**256Units**	**16 Units**	192bit	Supported	Supported	Supported
Arc Pro A50 6GB	**128 Units**	**8 Units**	96bit	Supported*	Supported	Supported
Arc Pro A404GB	**128Units**	**8Units**	96bit	Supported*	Supported	Supported

*Quantize LLM model

Arc A770, A750, A580 with 256bit memory bandwidth for AI acceleration

XeSS | DirectX XII ULTIMATE | XMX AI Acceleration | Xe Media Engine | Vulkan

intel Ai summit

Experience the power of AI with Intel® Arc GPUs

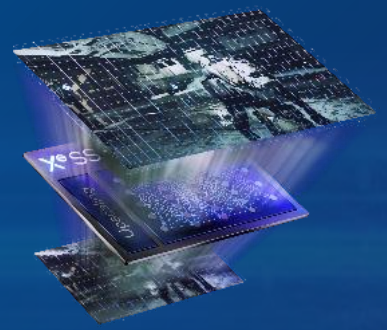
Ready for AI

Develop for Users

at Scale

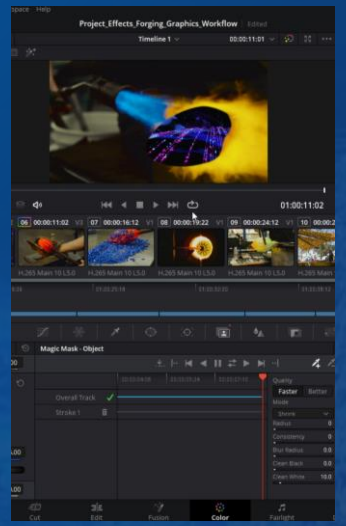
Gaming

XeSS



AI Upscaling

Creation

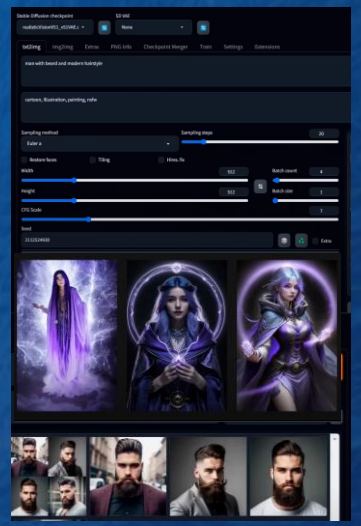


Already supports multiple

AI Effects

in Davinci Resolve, and growing

Generative AI



AI Enhancements

For Developers

OpenVINO

Mxnet

oneAPI

WindowsML

TensorFlow

PyTorch

ONNX Runtime

WebGPU

… and more

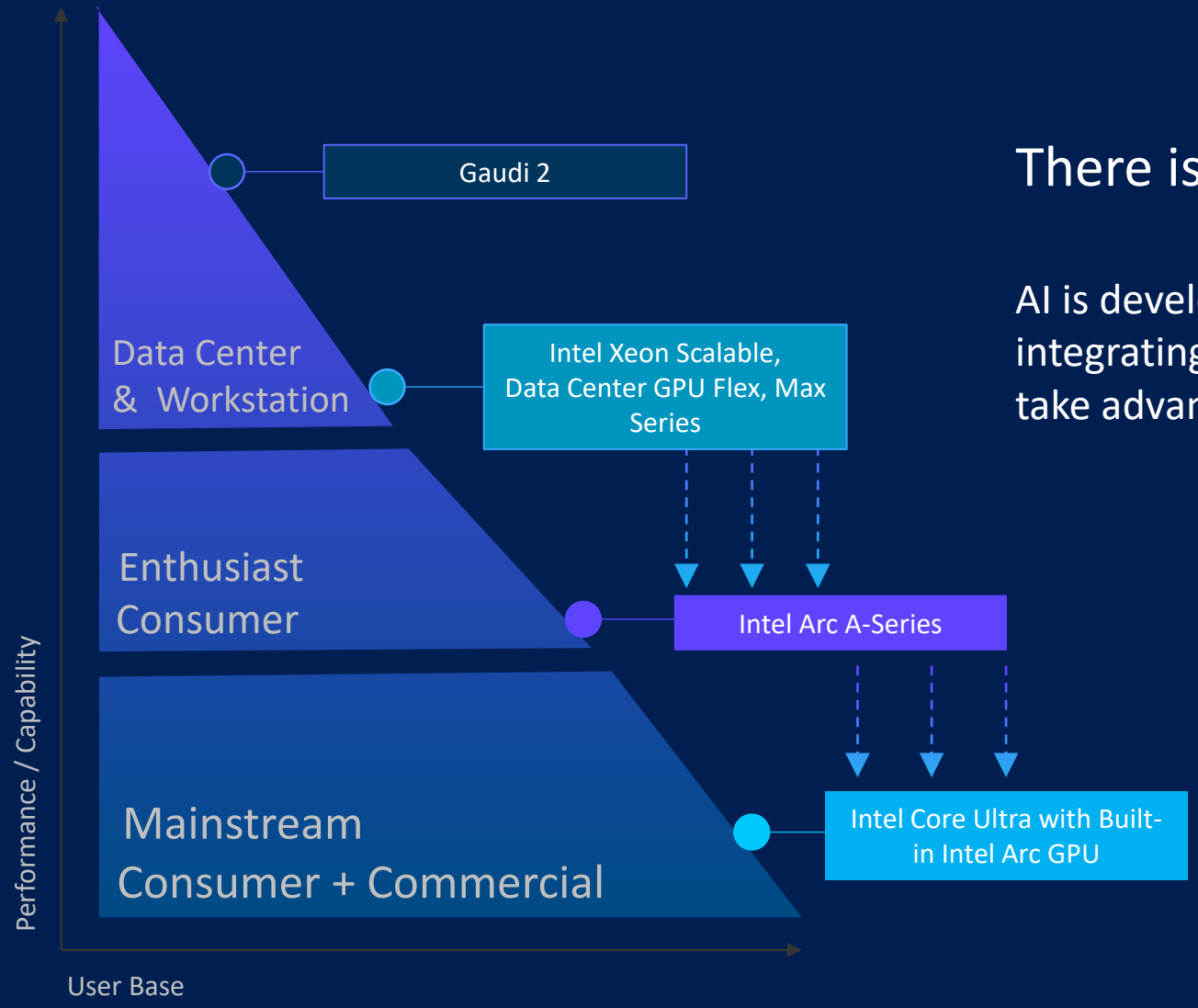
Widespread Support

for the latest frameworks

Ready for today and tomorrow, fast and optimized

intel AI summit

Snapshot of AI Hardware Capacity



There is a “trickle down” effect happening

AI is developed on the cutting edge, and Intel has been integrating AI capabilities into products for years, ready to take advantage of this new paradigm shift.

Intel is bringing AI hardware to the masses at scale

intel Ai summit

Thank You!

Appendix

Notices and Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

See backup for configuration details. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks

Intel® Arc™ graphics only available on select Intel® Core™ Ultra H-series processor-powered systems with at least 16GB of system memory in a dual-channel configuration. OEM enablement required; check with OEM or retailer for system configuration details.

AI features may require software purchase, subscription or enablement by a software or platform provider, or may have specific configuration or compatibility requirements. Details at www.intel.com/AIPC.

Intel technologies may require enabled hardware, software or service activation.

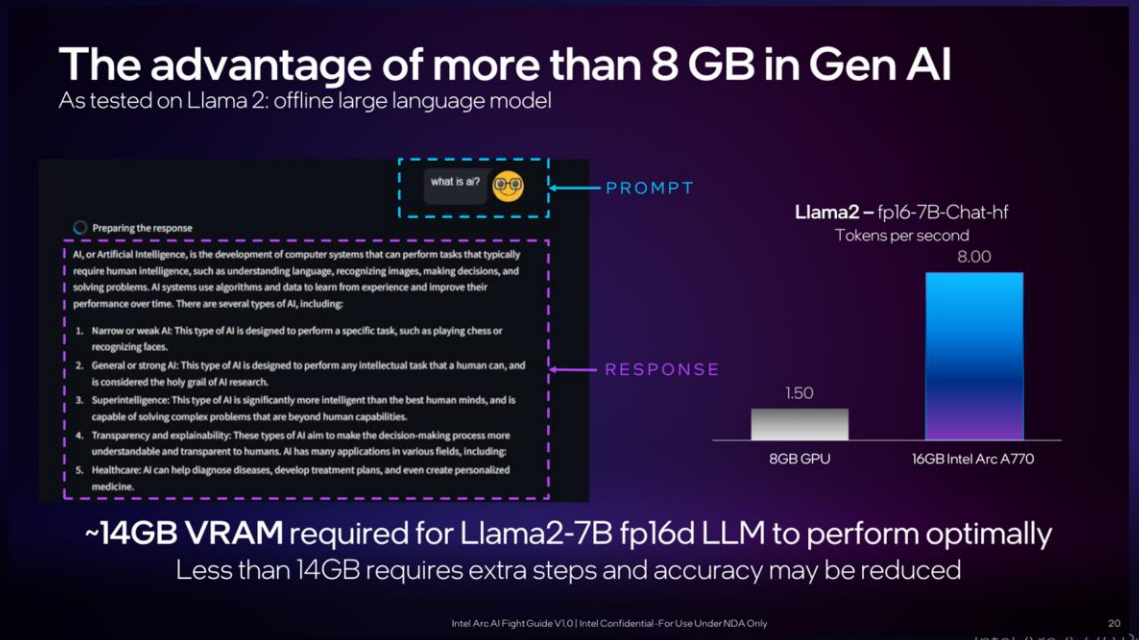
No product or component can be absolutely secure.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

intel Ai summit

Workloads and Configurations – LLM on 16GB of VRAM

CLAIM	SYSTEM CONFIGURATION	MEASUREMENT	MEASUREMENT PERIOD
Intel Arc A770 GPU can run large language models. Due to 16GB of memory, it is able to Llama 2 at 8 tokens per second – significantly faster than other GPUs that are limited by memory.	Motherboard - ROG MAXIMUS Z790 HERO BIOS - 1501 Memory - 32GB (2x16GB) DDR5 5600MHz Memory SKU - CMT32GX5M2B5600C36 XMP RAM - DDR5-5600 CPU - i9-13900K dGPU - Intel ARC A770 dGPU Driver - 31.0.101.5180 / 5122 / 5085 SSD - Corsair MP600 PRO XT OS - Win 11 Pro OS Version - 10.0.22631.2861 Power Policy - High Performance	As tested on Llama2-7B-fp16 Test was done by using streamlit chatbot scripts with LlamaTokenizer, OVModelforCausalLM and LlamaforCausalLM pipelines Results taken as a median from 3 runs measured Tokens per second: • A770 - OpenVINO Optimum chatbot app: 8 tokens/s • 4060 8GB – CUDA: 1.50 tokens/s	Dec 1 2023




intel