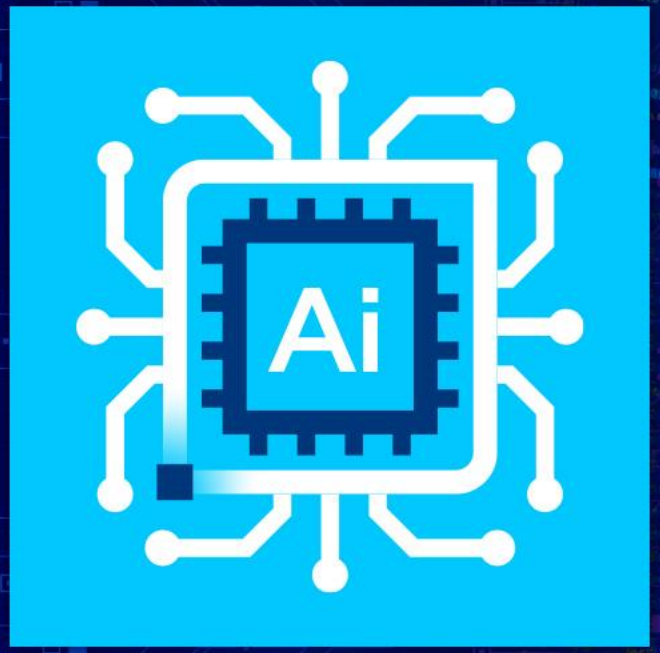
OpenVINO™ DEVCON Workshop Series 2023

Generative AI with OpenVINO

Computer Vision Specialist

Technical Sales Specialist, Sales & Mktg Grp.

What is Generative AI?



Generating and manipulating data at a large scale with increased flexibility

Compelling Creative Use Cases

Gaming Experiences

Global Generative AI Market Size*

$1.3T by 2032
(195兆円 @ 1USD=150JPY)

42% CAGR over 10 years

*Source: Bloomberg Intelligence

Room Design

Novel Illustration

Fashion & E-Commerce

Web Design

Generative AI: Pain Points

Large model size

ファイルサイズ

Large memory footprint

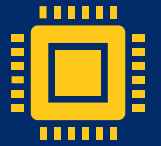
メモリ使用量

Slow inference speed

長い推論時間

Difficulty training + optimizing

学習と最適化の困難さ



No flexibility to run workloads on different HW

ハードウエア
スケーラビリティー

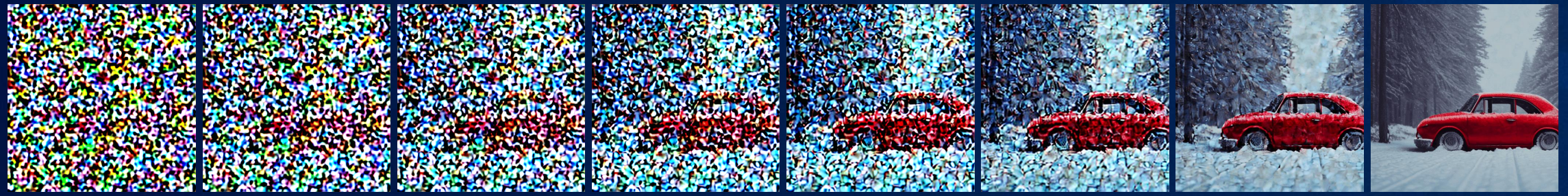
What is Stable Diffusion

Prompt: “red car in snowy forest”

Diffusion Training Process

Fixed Forward Diffusion Process

Slowly adds random noise to training data



Reverses noise to reconstruct the data samples

Generative Reverse Denoising Process

Generative AI increases the overall inference time

High inference time

Poor customer experience

Is the solution to use powerful and expensive machines?

Or use an edge device or my laptop?

Generative AI: Pain Points

Large model size

ファイルサイズ

Large memory footprint

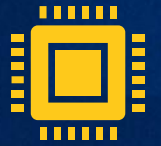
メモリ使用量

Slow inference speed

長い推論時間

Difficulty training + optimizing

学習と最適化の困難さ



No flexibility to run workloads on different HW

ハードウエア
スケーラビリティー

Generative AI: Make it easier with OpenVINO™

Reduce model size

ファイルサイズ削減

Reduce memory footprint

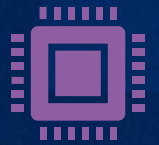
メモリ使用量削減

Faster inference speed

高速な推論

Strategy optimizing

モデル最適化



Flexibility to run workloads on CPUs and Intel GPUs

ハードウエア
スケーラビリティー
向上

HuggingFace
Optimum with OpenVINO

optimum-intel

https://huggingface.co/docs/optimum/intel/index

Let's Run the Demo!

Stable Diffusion

+ intel

HuggingFace
Optimum with OpenVINO

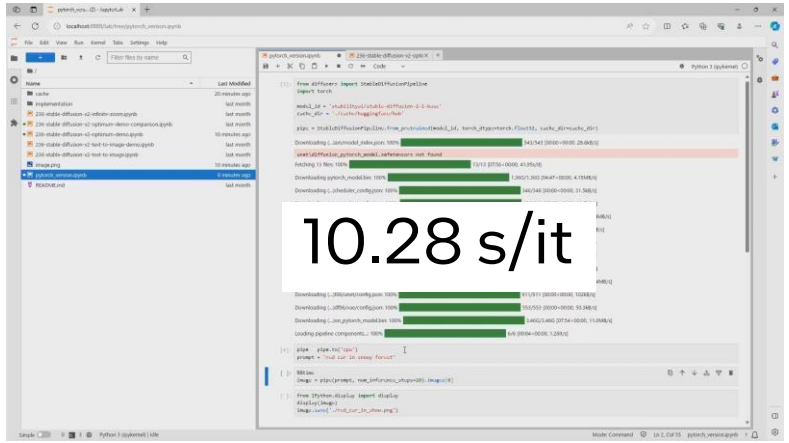
Text to Image Pipeline:
generation to create images from a text description as input

"red car in snowy forest"

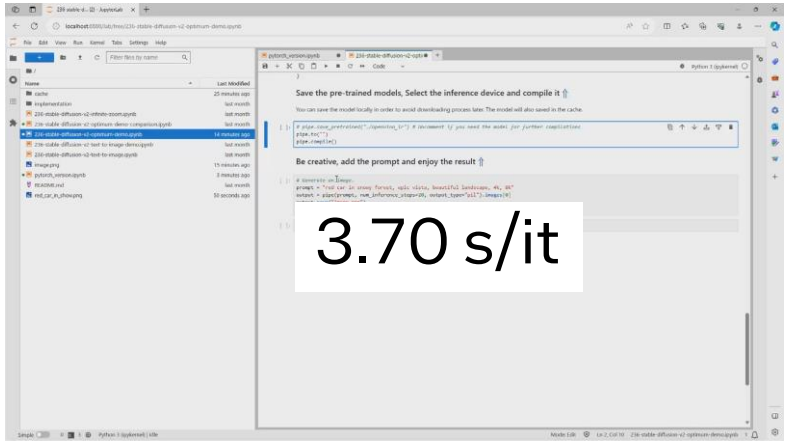
1. FP32 PyTorch Model CPU

2. FP32 OpenVINO Model CPU

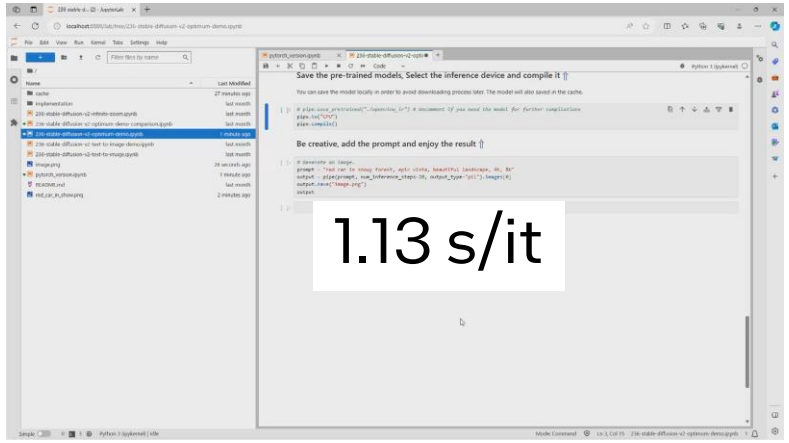
3. FP16 OpenVINO Model Intel discrete GPU



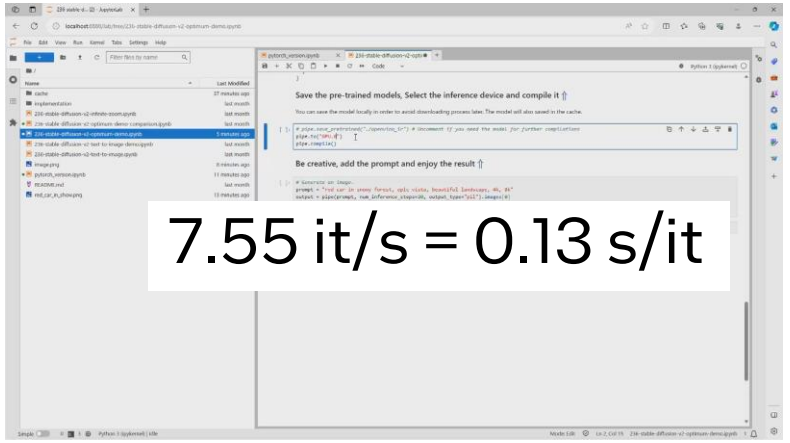
PyTorch CPU FP32



OpenVINO CPU FP32



OpenVINO 内蔵GPU FP16



OpenVINO 外付けGPU Intel Arc A770m FP16

HW: NUC Serpent Canyon, Core i7-12700H, RAM 32GB,
SW: Windows 11 22H2, OpenVINO 2023.1.0

OpenVINO™ Notebooks

https://github.com/openvinotoolkit/openvino_notebooks

（参考）PyTorch版テストコード（jupyter notebook）

```
from diffusers import StableDiffusionPipeline
import torch

model_id = 'stabilityai/stable-diffusion-2-1-base'
cache_dir = './cache/huggingface/hub'

pipe = StableDiffusionPipeline.from_pretrained(model_id, torch_dtype=torch.float32, cache_dir=cache_dir)
```

```
pipe = pipe.to('cpu')
prompt = 'red car in snowy forest'
```

```
%%time
image = pipe(prompt, num_inference_steps=20).images[0]
```

```
from IPython.display import display
display(image)
image.save('./red_car_in_show.png')
```

Hugging Face Optimum with OpenVINO
Accelerate diffusers inference

```
- from diffusers import StableDiffusionPipeline
+ from optimum.intel.openvino import OVStableDiffusionPipeline

model_id = "stabilityai/stable-diffusion-2-1-base"
- pipe = StableDiffusionPipeline.from_pretrained(model_id)
+ pipe = OVStableDiffusionPipeline.from_pretrained(model_id, export=True, compile=False)
- pipe.save_pretrained("./stabilityai_cpu")
+ pipe.save_pretrained("./openvino_ir")
pipe.to("DEVICE_NAME")
pipe.compile()
prompt = "red car in snowy forest"
output_cpu = pipe(prompt, num_inference_steps=17).images[0]
```

+ intel

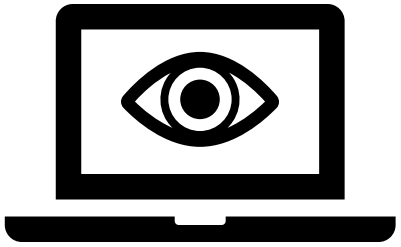
https://github.com/huggingface/optimum-intel

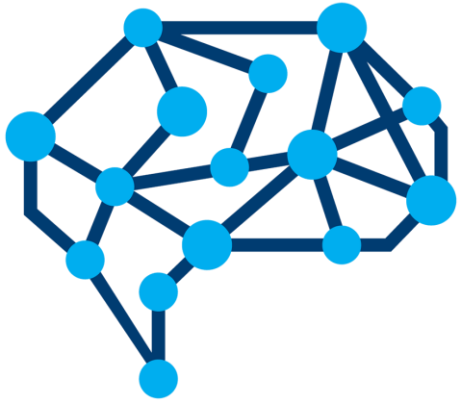
OpenVINO™

OpenVINO™

Open Source Visual Inference Neural Network Optimizations

High performance inference

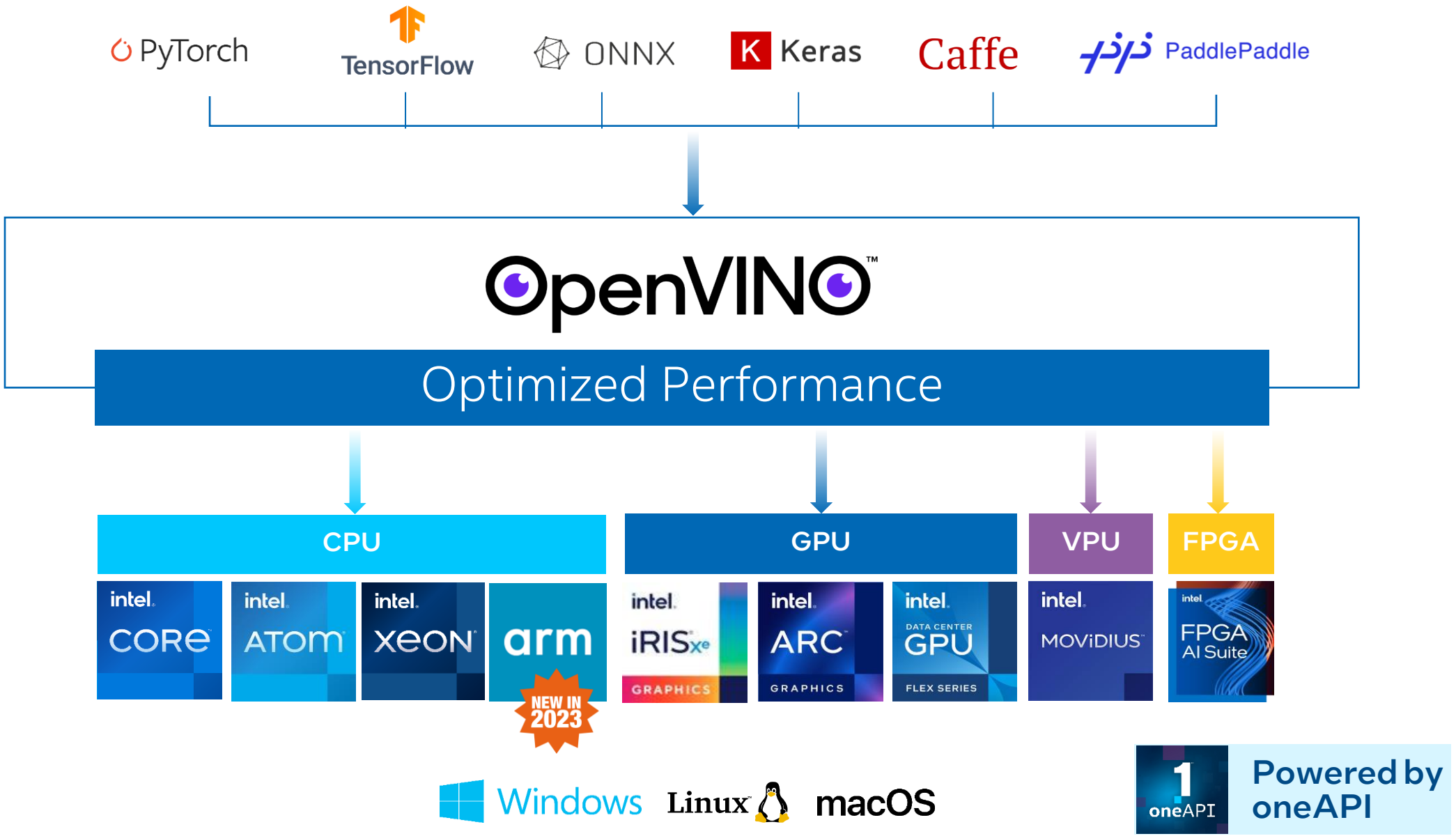




Generative AI

2017 — 5 years — 2023

OpenVINO over the years


PyTorch
TensorFlow
ONNX
Keras
Caffe
PaddlePaddle
OpenVINO™
Optimized Performance
CPU
GPU
VPU
FPGA
intel CORE
intel ATOM
intel XEON
arm
NEW IN 2023
intel iRIS Xe GRAPHICS
intel ARC GRAPHICS
intel DATA CENTER GPU FLEX SERIES
intel MOVIDIUS
intel FPGA AI Suite
Windows
Linux
macOS
oneAPI
Powered by oneAPI

OpenVINO™ Stable Diffusion

High performance inference

Reduce model size

ファイルサイズ削減

Reduce memory footprint

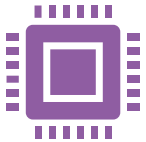
メモリ使用量削減

Faster inference speed

高速な推論

Strategy optimizing

モデル最適化



Flexibility to run workloads on CPUs and Intel GPUs

ハードウエア
スケーラビリティー
向上

OpenVINO™ Stable Diffusion

High performance inference

Reduce model size

ファイルサイズ削減

Reduce memory footprint

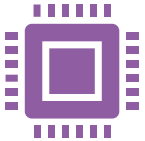
メモリ使用量削減

Faster inference speed

高速な推論

Strategy optimizing

モデル最適化



Flexibility to run workloads on CPUs and Intel GPUs

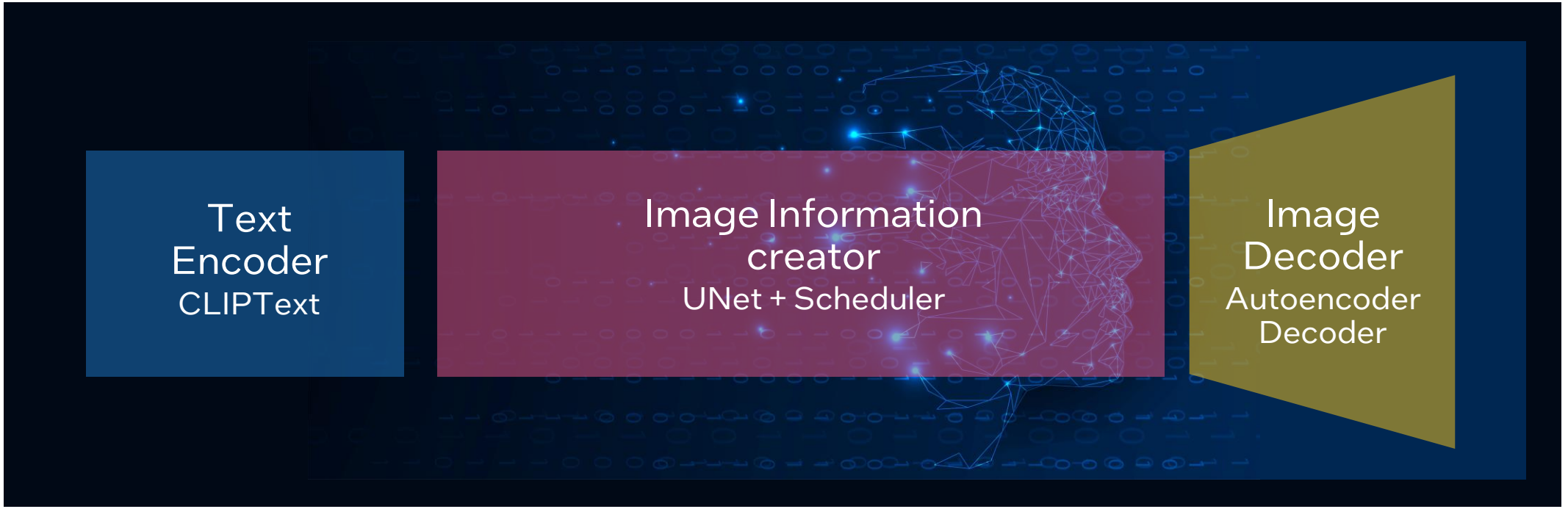
ハードウエア
スケーラビリティー
向上

OpenVINO™ Stable Diffusion

High performance inference

SD Pipeline

Downloading process



OpenVINO™ Stable Diffusion

High performance inference

SD Pipeline

Compiling models into the Intel HW

```
from openvino.runtime import Core
core = Core()
text_enc = core.compile_model(TEXT_ENCODER_OV_PATH, "GPU")
unet_model = core.compile_model(UNET_OV_PATH, 'GPU')
vae_decoder = core.compile_model(VAE_DECODER_OV_PATH, 'GPU')
```

FP16 conversion on the fly for GPU devices

OpenVINO™ Stable Diffusion

High performance inference

SD Pipeline

OVStableDiffusionPipeline

```
from transformers import CLIPTokenizer

scheduler = LMSDiscreteScheduler.from_config(conf)
tokenizer = CLIPTokenizer.from_pretrained('openai/clip-vit-large-patch14')

ov_pipe = OVStableDiffusionPipeline(
    tokenizer=tokenizer,
    text_encoder=text_enc,
    unet=unet_model,
    vae_decoder=vae_decoder,
    scheduler=scheduler
)
```

OpenVINO™ Stable Diffusion

High performance inference

Reduce model size

ファイルサイズ削減

Reduce memory footprint

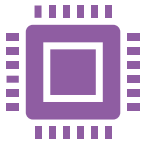
メモリ使用量削減

Faster inference speed

高速な推論

Strategy optimizing

モデル最適化



Flexibility to run workloads on CPUs and Intel GPUs

ハードウエア
スケーラビリティー
向上

Reduce model size

ファイルサイズ削減

```
8.0K    stabilityai_cpu/feature_extractor
1.3G    stabilityai_cpu/text_encoder
8.0K    stabilityai_cpu/scheduler
1.6M    stabilityai_cpu/tokenizer
3.3G    stabilityai_cpu/unet
[illegible]  stabilityai_cpu/vae
4.9G    stabilityai_cpu/
```

FP32 Native Model

```
8.0K    openvino_ir/feature_extractor
1.3G    openvino_ir/text_encoder
131M    openvino_ir/vae_encoder
8.0K    openvino_ir/scheduler
1.6M    openvino_ir/tokenizer
3.3G    openvino_ir/unet
[illegible]  openvino_ir/vae_decoder
4.9G    openvino_ir/
```

FP32 OpenVINO Model

```
8.0K    modelSD21_dGPU_OV/feature_extractor
652M    modelSD21_dGPU_OV/text_encoder
8.0K    modelSD21_dGPU_OV/scheduler
1.6M    modelSD21_dGPU_OV/tokenizer
1.7G    modelSD21_dGPU_OV/unet
[illegible]  modelSD21_dGPU_OV/vae_decoder
2.4G    modelSD21_dGPU_OV/
```

FP16 OpenVINO model

OpenVINO™ Stable Diffusion

High performance inference

Reduce model size

ファイルサイズ削減

Reduce memory footprint

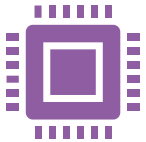
メモリ使用量削減

Faster inference speed

高速な推論

Strategy optimizing

モデル最適化



Flexibility to run workloads on CPUs and Intel GPUs

ハードウエア
スケーラビリティー
向上

What can you do next?

Convert, optimize, and run models using OpenVINO and Optimum-Intel

Try stable diffusion using Optimum-Intel OpenVINO

Text-to-Image

"valley in the Alps at sunset, epic vista, beautiful landscape, 4k, 8k"

Try stable diffusion using OpenVINO Notebooks

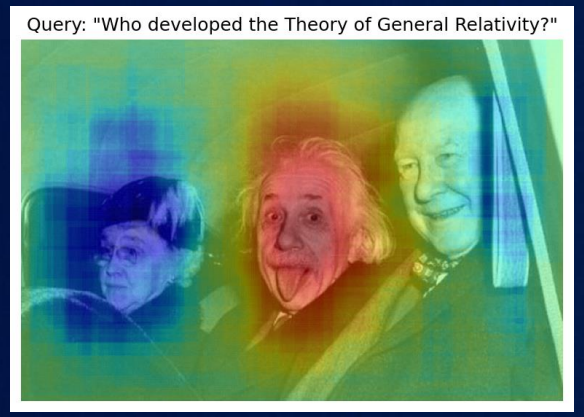
Text-to-Image

"valley in the Alps at sunset, epic vista, beautiful landscape, 4k, 8k"

OpenVINO™ Notebooks

launch binder | Open in Colab

Language-Visual Saliency with CLIP



Visual QA with BLIP

AI Trends with OpenVINO

Infinite Zoom Stable Diffusion

'a photo of a blue T-shirt with yellow text "OPENVINO"'

Generative AI: Pain Points

Large model size

ファイルサイズ

Large memory footprint

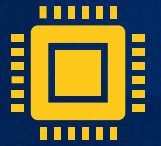
メモリ使用量

Slow inference speed

長い推論時間

Difficulty training + optimizing

学習と最適化の困難さ



No flexibility to run workloads on different HW

ハードウエア
スケーラビリティー

Generative AI: Make it easier with OpenVINO™

Reduce model size

ファイルサイズ削減

Reduce memory footprint

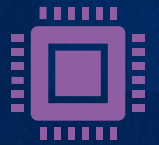
メモリ使用量削減

Faster inference speed

高速な推論

Strategy optimizing

モデル最適化



Flexibility to run workloads on CPUs and Intel GPUs

ハードウエア
スケーラビリティー
向上

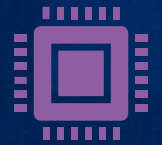
Generative AI: Make it easier with OpenVINO™

Reduce model size

Reduce memory footprint

Faster inference speed

Strategy optimizing



Flexibility to run workloads on CPUs and Intel GPUs

Install your way, today!

```
pip install openvino
pip install openvino-dev
```

ubuntu

Red Hat

docker

ANACONDA

CONDA-FORGE

www.openvino.ai

Read more about OpenVINO in the Industry

OpenVINO Blogs

https://medium.com/openvino-toolkit

Edge AI Industry Solutions

https://intel.ly/3NHkA7t

Scan/Click the QR codes for more information

New to OpenVINO™? Learn and test

Stable Diffusion

Intel® Developer Cloud for the Edge

Edge AI Certification

intel.ly/edgeaicert

Scan/Click the QR codes for more information

New to Computer Vision? Want to rapidly develop CV models

Intel Geti

geti.intel.com

WebUIベースのモデル学習ツール

Scan/Click the QR code for more information

Notices and disclaimers

Intel is committed to respecting human rights and avoiding complicity in human rights abuses. See Intel's Global Human Rights Principles. Intel® products and
software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

Intel® technologies may require enabled hardware, software, or service activation. No product or component can be absolutely secure. Your costs and results may vary.

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.

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.