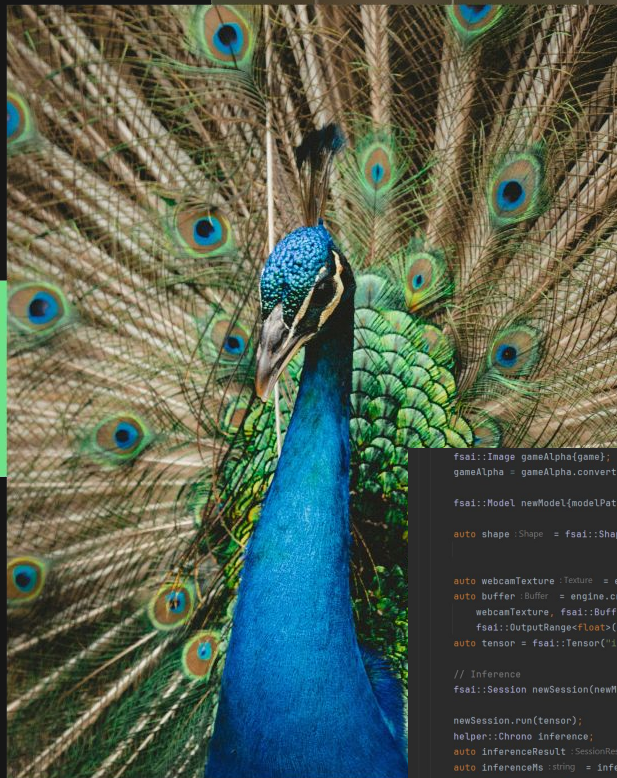


Full GPU driven AI workloads with GStreamer and Raven

Andoni Morales
Nacho García
Aleix Figueres
Izan Leal
Sergio Sánchez

GStreamer Conference 2025



```
fsai::Image gameAlpha(game);
gameAlpha = gameAlpha.convertTo(fsai::ColorSpace::RGBA);

fsai::Model newModel(modelPath);

auto shape :Shape = fsai::Shape::createMatrix(1 /*batch, one webcamAlpha
tensorInputChannels);

auto webcamTexture :Texture = engine.createTexture(webcamAlpha);
auto buffer :Buffer = engine.createBufferFromTexture(
    webcamTexture, fsai::BufferType::float32, fsai::BufferLayout::RGBA,
    fsai::OutputRange<float>(0.0F, 1.0F), fsai::Transform::resize(tensor));
auto tensor = fsai::Tensor("input_1", shape, buffer);

// Inference
fsai::Session newSession(newModel, engine);

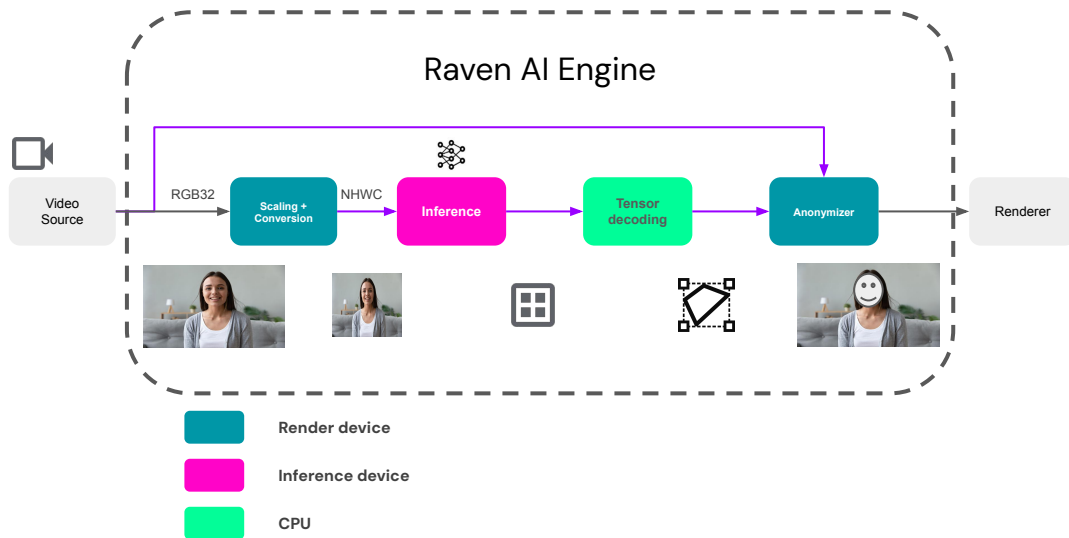
newSession.run(tensor);
helper::Chrono inference;
auto inferenceResult :SessionResult = newSession.run(tensor);
auto inferenceMs :string = inference.stop();
inferenceMs += " : ";
inferenceMs += engine.getAdapter().getDescription();
```

Raven AI Engine

Raven is our in-house AI engine designed for real-time multimedia workflows, combining high-throughput AI inference with graphics-native processing.

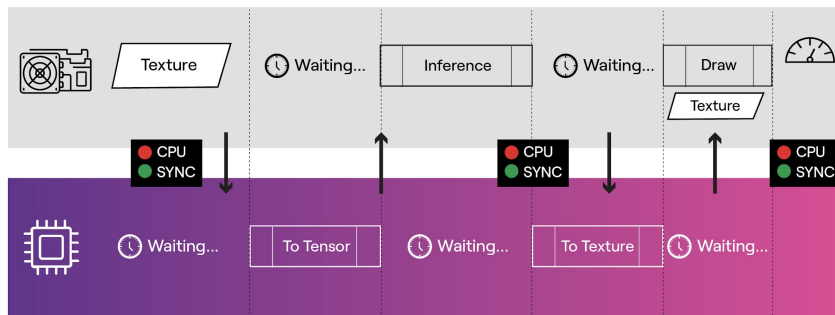
It gives full control over the entire GPU pipeline, from memory allocation to execution, allowing for deep customization across hardware and environments.

- 100% GPU driven processing
- Multivendor compatibility
- Multiplatform support
- Multipurpose design
- Low code integration

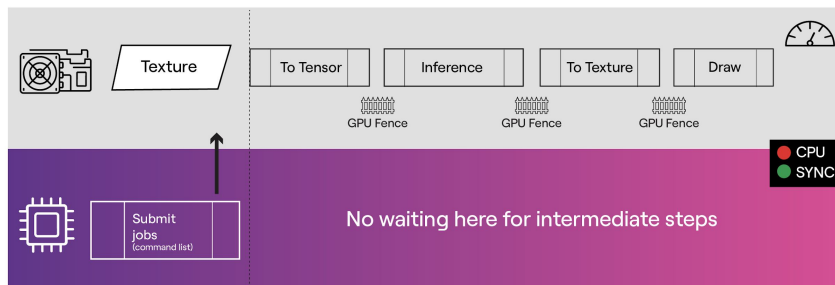


Full GPU-Driven AI pipeline

Conventional AI pipeline



Full GPU-Driven AI pipeline



- Task parallelization
- Post-processing algorithms on the GPU (eg: Non-Max Supression)
- GPU rendering engine

Anonymization task runs at 500 fps for 4K (3840x2160) input a in an Nvidia RTX 4060 Desktop.