

WebEvolve 2024

Updates on WebGPU and WebAI

Yang Gu

Manager of Intel WebGraphics and WebAI Team

May 28, 2024





Latest News

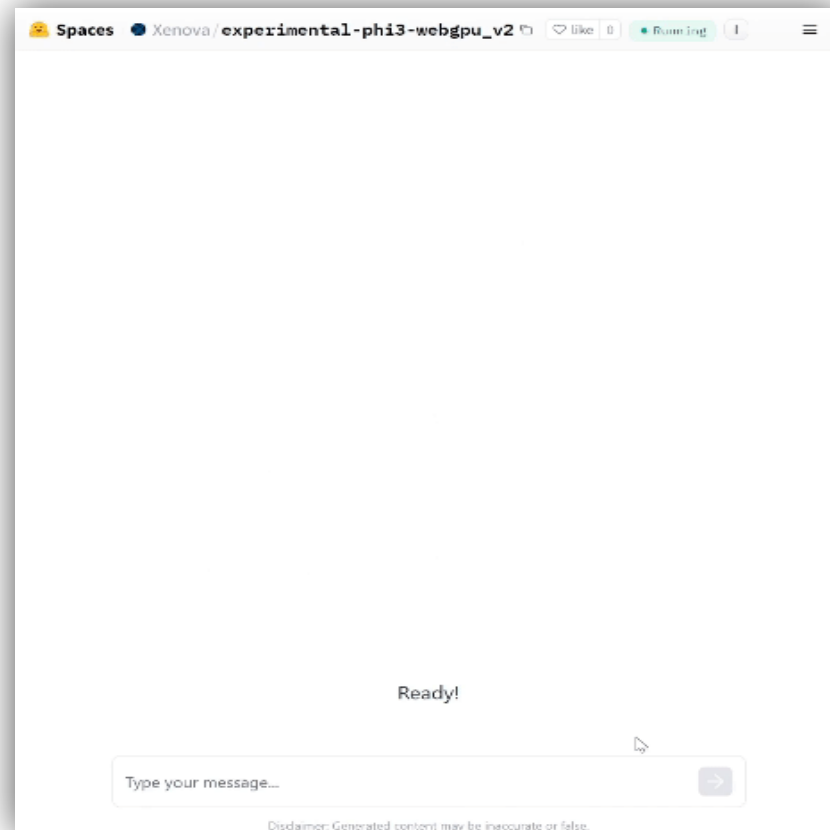
Google I/O 2024

WebGPU and Web Assembly are the backbone technologies that enable on-device AI on the web.



Microsoft Build 2024

[Enjoy the power of Phi-3 with ONNX Runtime WebGPU](#)





Core Updates

WebGPU

W3C Working Draft, 18 May 2024



▼ More details about this document

- This version:**
<https://www.w3.org/TR/2024/WD-webgpu-20240518/>
- Latest published version:**
<https://www.w3.org/TR/webgpu/>
- Editor's Draft:**
<https://gpuweb.github.io/gpuweb/>
- Previous Versions:**
<https://www.w3.org/TR/2024/WD-webgpu-20240517/>
- History:**
<https://www.w3.org/standards/history/webgpu/>

TABLE OF CONTENTS

- 1 Introduction**
 - 1.1 Overview
 - 1.2 Syntax Notation
 - 1.3 Mathematical Terms and Notation
- 2 WGSL Module**
 - 2.1 Shader Lifecycle
 - 2.2 Errors
 - 2.3 Diagnostics
 - 2.3.1 Diagnostic Processing
 - 2.3.2 Filterable Triggering Rules
 - 2.3.3 Diagnostic Filtering
 - 2.4 Limits
- 3 Textual Structure**
 - 3.1 Parsing
 - 3.2 Blankspace and Line Breaks
 - 3.3 Comments
 - 3.4 Tokens
 - 3.5 Literals
 - 3.5.1 Boolean Literals
 - 3.5.2 Numeric Literals
 - 3.6 Keywords
 - 3.7 Identifiers
 - 3.7.1 Identifier Comparison
 - 3.8 Context-Dependent Names
 - 3.8.1 Attribute Names
 - 3.8.2 Built-in Value Names
 - 3.8.3 Diagnostic Rule Names
 - 3.8.4 Diagnostic Severity Control Names

WebGPU Shading Language

W3C Working Draft, 14 May 2024



▼ More details about this document

- This version:**
<https://www.w3.org/TR/2024/WD-WGSL-20240514/>
- Latest published version:**
<https://www.w3.org/TR/WGSL/>
- Editor's Draft:**
<https://gpuweb.github.io/gpuweb/wgsl/>
- Previous Versions:**
<https://www.w3.org/TR/2024/WD-WGSL-20240507/>
- History:**
<https://www.w3.org/standards/history/WGSL/>
- Feedback:**
public-gpu@w3.org with subject line "[wgsl] - message topic -" ([archives](#))
[GitHub](#)
- Editors:**
[Alan Baker \(Google\)](#)
[Mehmet Oguz Derin](#)
[David Neto \(Google\)](#)
- Former Editors:**
[Myles C. Maxfield \(Apple Inc.\)](#)
[dan sinclair \(Google\)](#)
- Participate:**
[File an issue \(open issues\)](#)
- Tests:**
[WebGPU CTS shader/](#)

Copyright © 2024 World Wide Web Consortium. W3C® liability, trademark and mission statement licenses apply.

Labels **Milestones**

5 Open ✓ 1 Closed

Sort ▾

Milestone 0

No due date ⌚ Last updated 3 days ago



92% complete
115 open
1,420 closed

Formerly "V1.0". Fixes or spec work for existing V1.0 features.

Milestone 1

No due date ⌚ Last updated 3 days ago

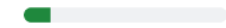


61% complete
40 open 63 closed

This is the current milestone for new features that we may land in the spec when they're resolved. Later milestones mean we're not landing them yet.

Milestone 2

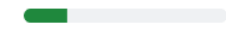
No due date ⌚ Last updated 5 days ago



13% complete
33 open 5 closed

Milestone 3+

No due date ⌚ Last updated 12 days ago



21% complete
176 open 48 closed

Items that need to be triaged into "Milestone 3" vs "Milestone 4+" when we define Milestone 3. (The items initially here were formerly "Post-V1".)

Browser Status

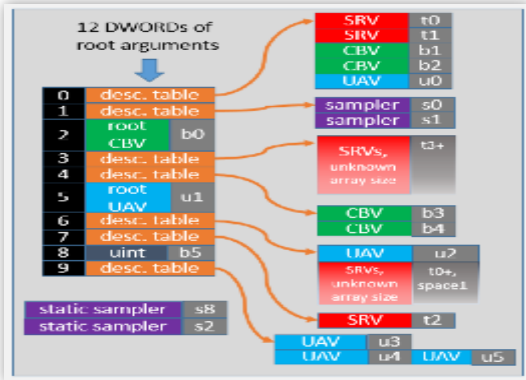


- Overall
 - WebGPU CG was created in Feb 2017. Contributions come from all major browser vendors, Intel and individuals
 - Chrome (MS Edge is the same) supports WebGPU on Windows, ChromeOS and MacOS in M113 (May 2023)
 - Chrome supports WebGPU on Android in M121 (Feb 2024)
 - Safari supports WebGPU in [Technical Preview 185](#)
 - Firefox supports WebGPU in Nightly
- Specific features in Chrome
 - DXC (shader compiler) support
 - Timestamp Query, F16, DP4A (INT8), etc.
 - Graphite D3D11 (Replacement of Skia Ganesh)
 - [What's New in WebGPU](#) for more details

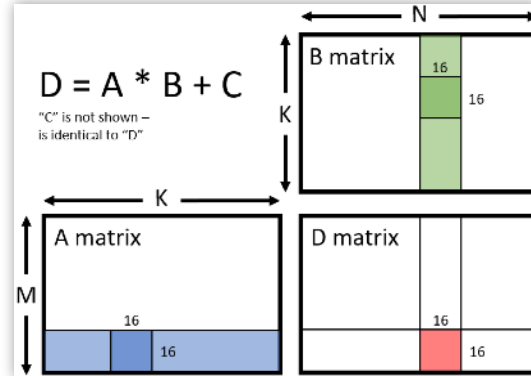


What's New in
WebGPU
blog series

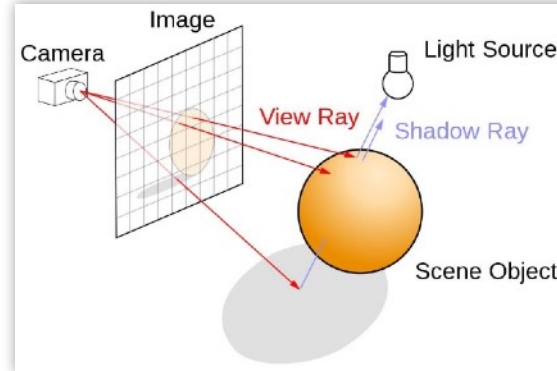
Upcoming New Features



Push Constants



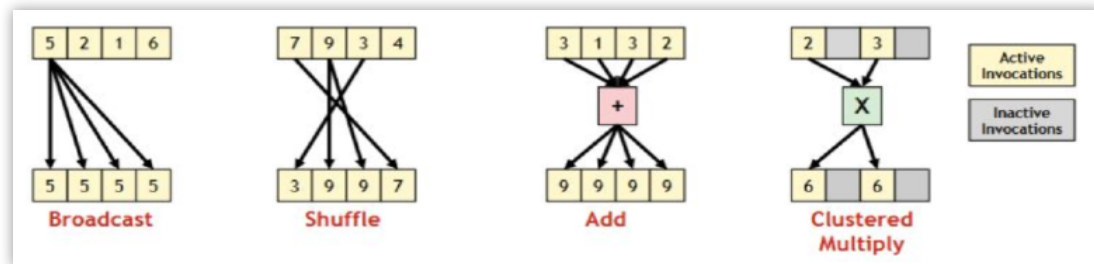
Wave Matrix



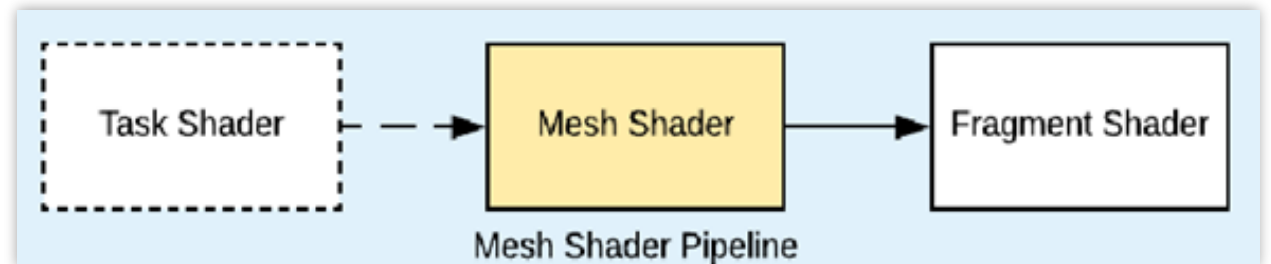
Ray Tracing



Variable Rate Shading



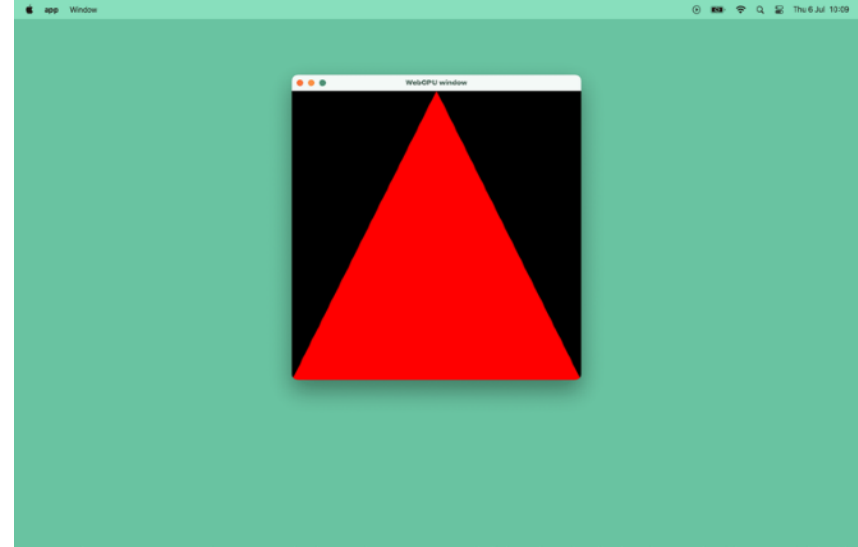
Subgroup



Mesh Shading

WebGPU Native

- WebGPU is a graphics and compute API for both the web and native
- Cross-platform
- [WebGPU headers](#)





Fast Growing Ecosystem

Game Engine - Unity

- Most popular game engine
- Unity [announced](#) the official support of WebGLGPU in Unity 6
- Demo [link](#) and [source code](#)



Game Engine – PLAYCANVAS

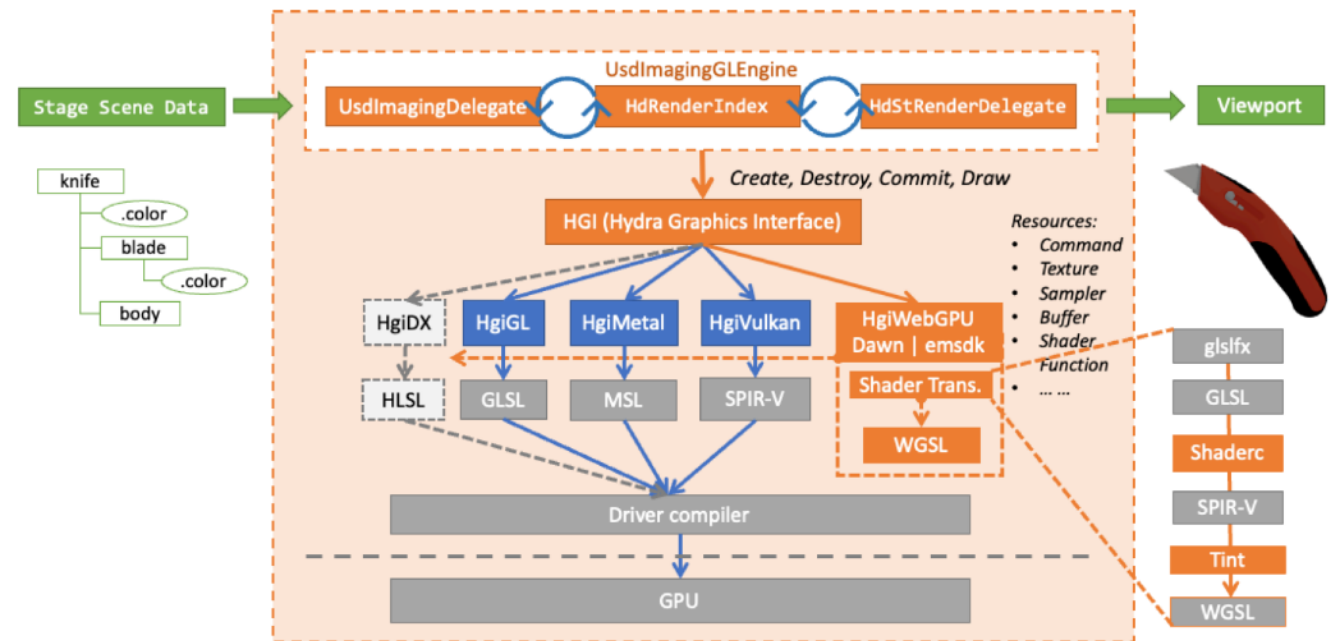
- Open source HTML5 game engine
- WebGPU support has officially arrived in the PlayCanvas Editor on Apr 18, 2024



Creator – AUTODESK®

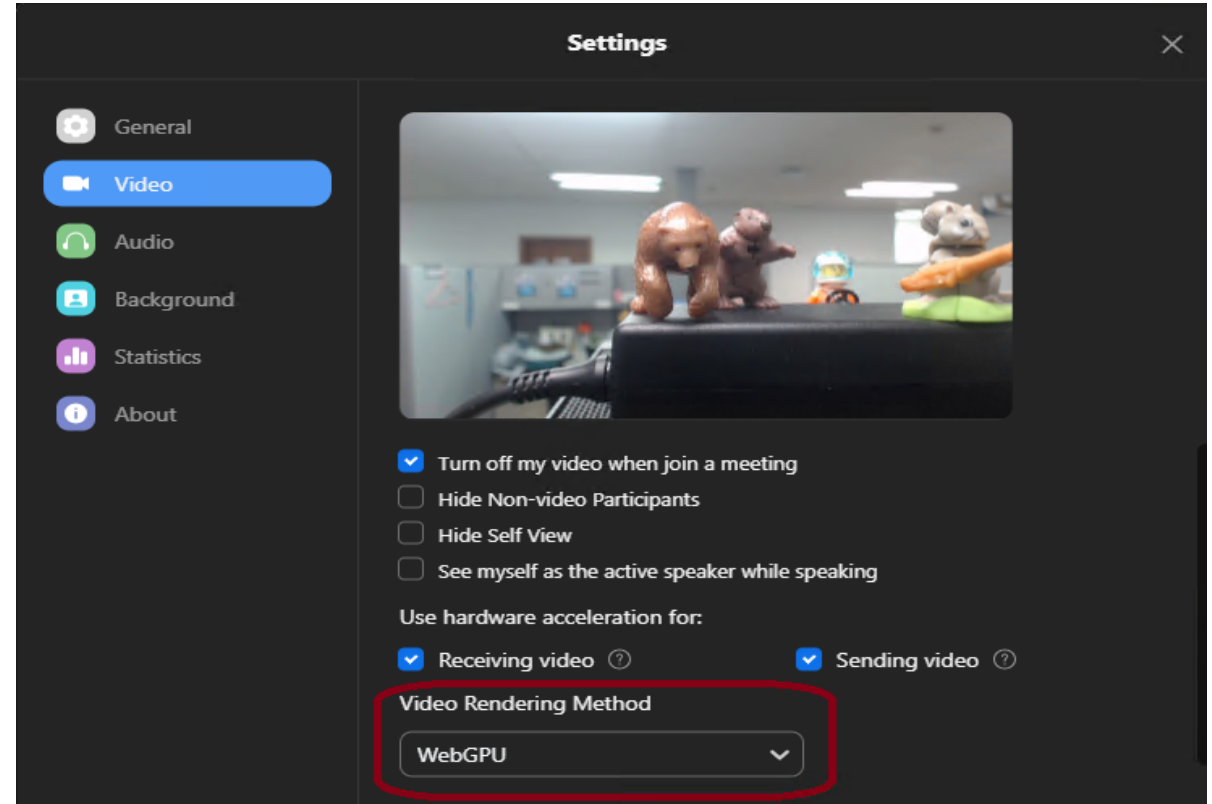
- Provides software products and services for the architecture, engineering, construction, manufacturing, media, education, and entertainment industries.
- USD and MaterialX on the WebGPU (Hydra Storm Renderer)

Extended HdStorm Render Pipeline



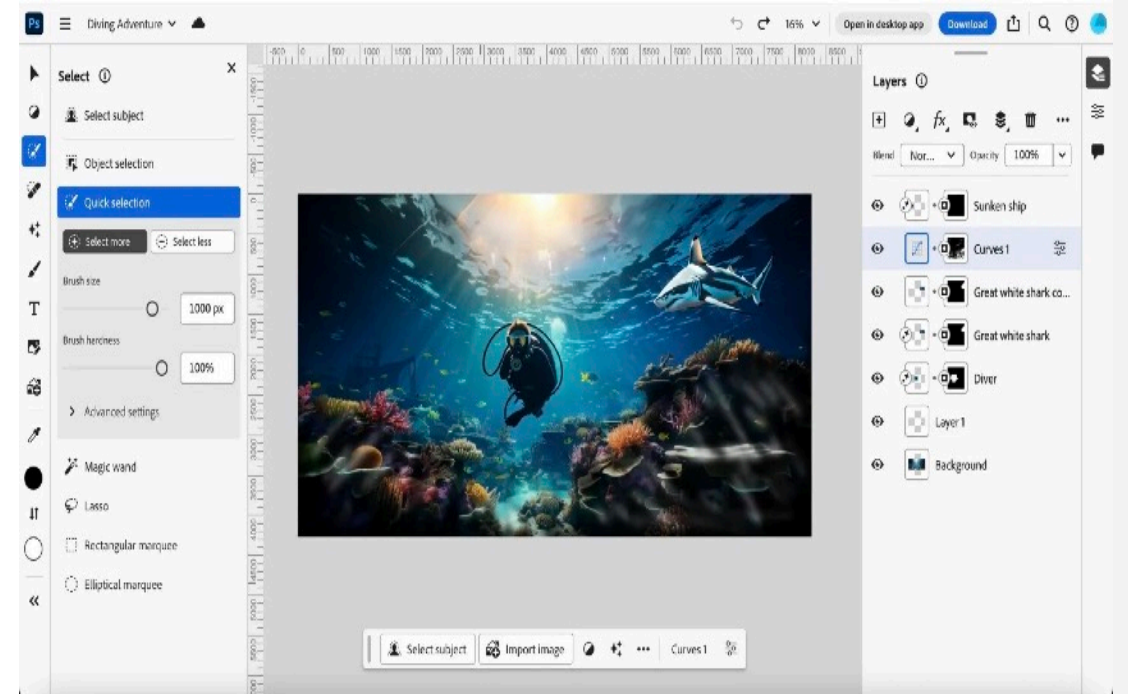
Video Conference - zoom

- Popular video conference solution
- WebGPU for rendering was officially released



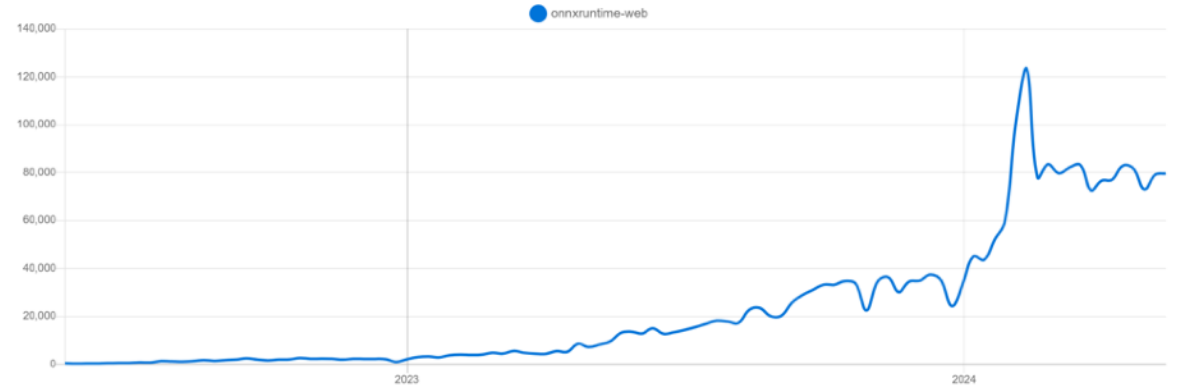
AI Framework – TensorFlow.js

- WebAI framework by Google and Intel has been working on its WebGPU backend since Sep 2019
- WebGPU backend was officially released in May 2023




AI Framework – ONNX RUNTIME

- Microsoft's major machine learning framework for both native and the web
- Intel joined the WebGPU EP (Execution Provider) effort in July 2023
- 1.17 release on Feb 3 is the first official release of WebGPU EP




Spaces | Xenova/experimental-phi3-webgpu like 6 Running



Phi-3 WebGPU

A private and powerful AI chatbot that runs locally in your browser.

You are about to load **Phi-3-mini-4k-instruct**, a 3.82 billion parameter LLM that is optimized for inference on the web. Once downloaded, the model (2.3 GB) will be cached and reused when you revisit the page.

Everything runs directly in your browser using  **Transformers.js**, meaning your conversations are not sent to a server. You can even disconnect from the WiFi after the model has loaded.

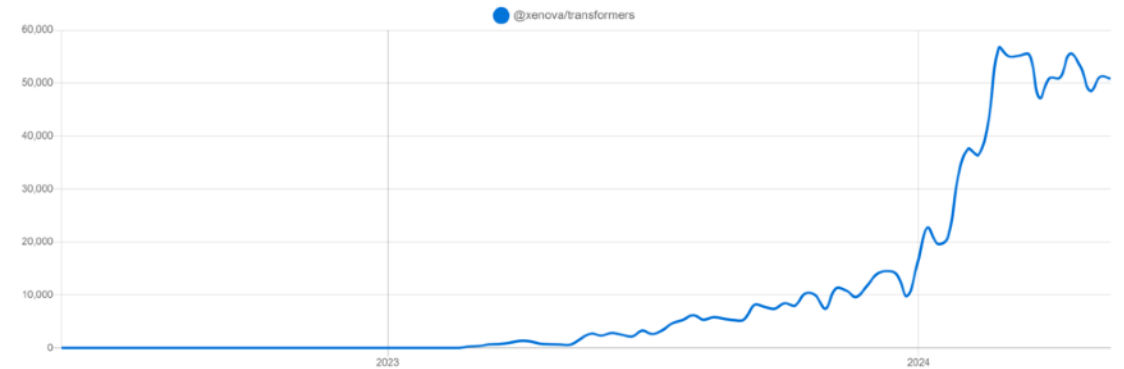
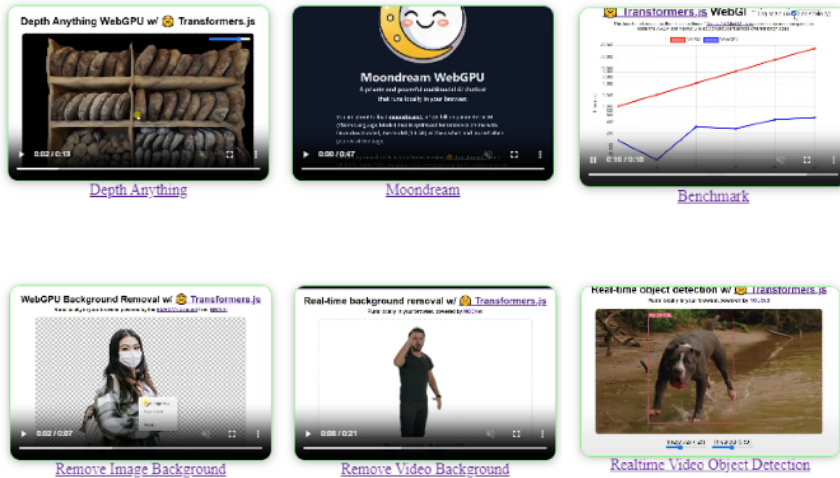
[Load model](#)

Type your message... →

Disclaimer: Generated content may be inaccurate or false.

AI Framework — 🧠 Transformers.js

- Functionally equivalent to HuggingFace's transformers python library
- Base on ONNX Runtime Web, WASM and WebGPU (V3)
- Maintained by Joshua Lochner (HuggingFace), and [more demos](#) can be found at HF



Spaces | Xenova/webgpu-clip | like 0 | Running

Real-time zero-shot image classification (WebGPU)

Runs locally in your browser w/ 🧠 Transformers.js

chameleon: 0.27
eagle: 0.21
seal: 0.21

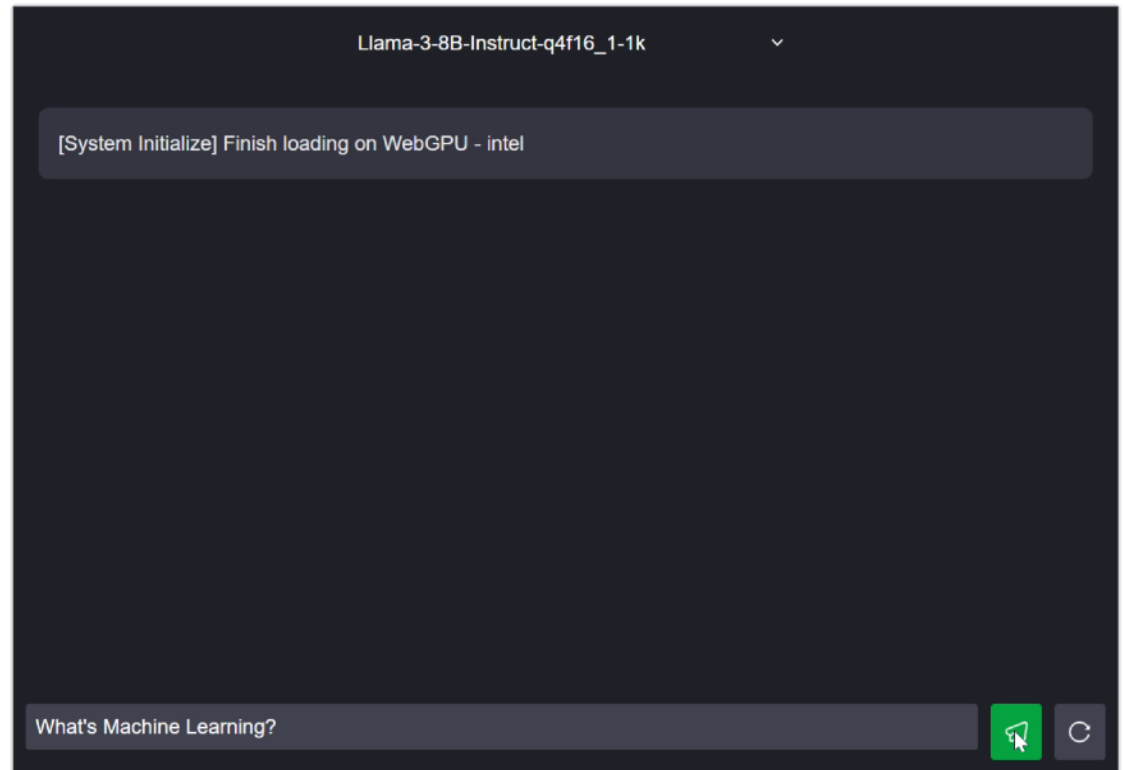
Labels (comma-separated)
chameleon, seal, eagle

Hypothesis template
A photo of a {}

FPS: 18.69

AI Framework –

- An End to End Machine Learning Compiler Framework for CPUs, GPUs and accelerators
- Web solution is based on WebGPU, and already ported many popular [LLM models to WebGPU](#)
- Intel contributed F16, while DP4A is WIP
- Enabled Llama3 WebGPU on the day 0 when it was released on Apr 19, 2024




AI Framework - MediaPipe

- LLM inference

- Falcon 1.3B
- Gemma 2.5B
- Phi-2 2.7B
- Stable LM 2.8B

- Gemini API

documentation for more details'. At the bottom, there is a dropdown menu for 'LLM model:' with 'gemma-2b-it-gpu-int4.bin' selected. On the right side of the interface, there is a text input field with the placeholder 'Enter some text.' and the text 'Write me an email to my friend Chris about how cake is so much better than pie. Use arguments like how good frosting is and how much better cakes look'. Below the input field is a label 'Inference time (ms):'." data-bbox="418 197 933 570"/>

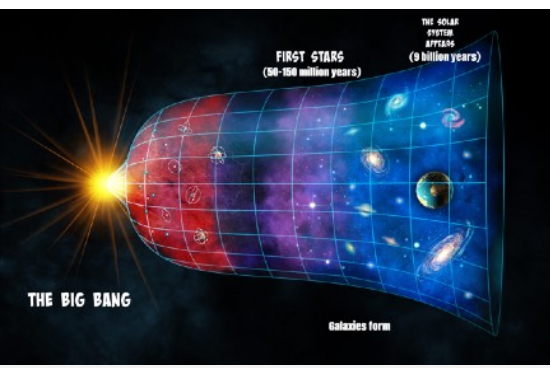


Browser - chrome

- Gemini Nano Integration
 - Integrated into Chrome, and will be released in M126 in June
 - The solution is based on WebGPU
 - “Help me write” feature and High-level APIs
- Graphite Project to replace Skia Ganesh



 Gemini Nano in Chrome



交流合作，共建生态！

- 邮箱: yang.gu@intel.com
- 任何Web图形的问题，欢迎发送到
<https://github.com/webatintel/webgraphicsforum/issues>

