

Yufeng Lyu

BAAI

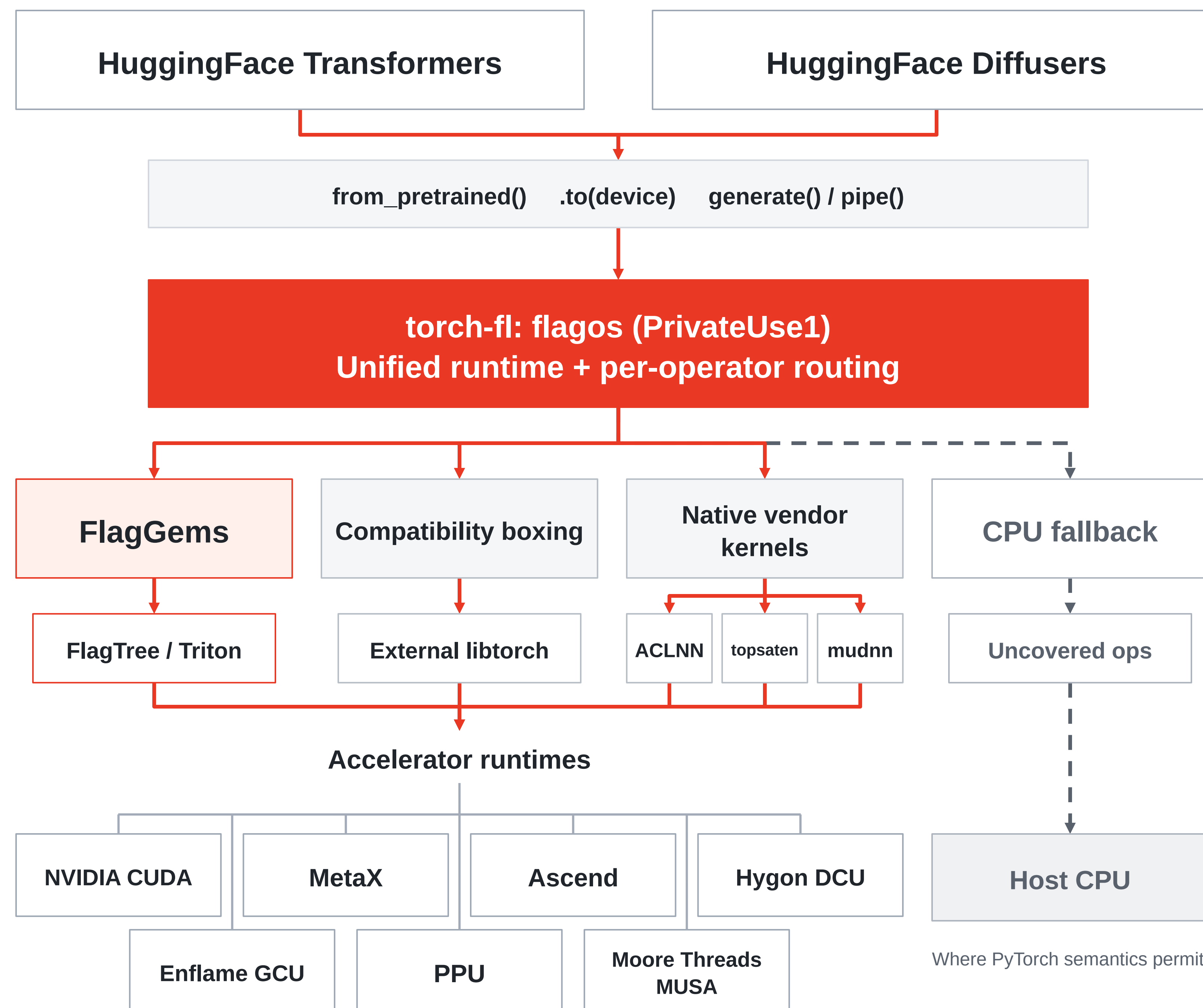
Torch-FL is a PyTorch plugin based on the FlagOS multi-chip technology stack. Accelerator adaptation happens in the PyTorch runtime and operator dispatch.

0 model-code changes

1 unified device: flagos

Model portability at the operator boundary

Unmodified upstream model code



Stable: NVIDIA CUDA, MetaX. Beta: Ascend, DCU, GCU. Experimental: PPU, MUSA.

BPU / TsingMicro: runtime or graph-path integration only. Model and feature coverage varies by platform. [3]

[1] QwenLM/Qwen-Image-2.1 [2] Operator support report (MetaX, 2026-09-19) [3] Architecture & compatibility github.com/flagos-ai/PyTorch-Plugin-FL

FlagOS: Open-sourced System SW

20+

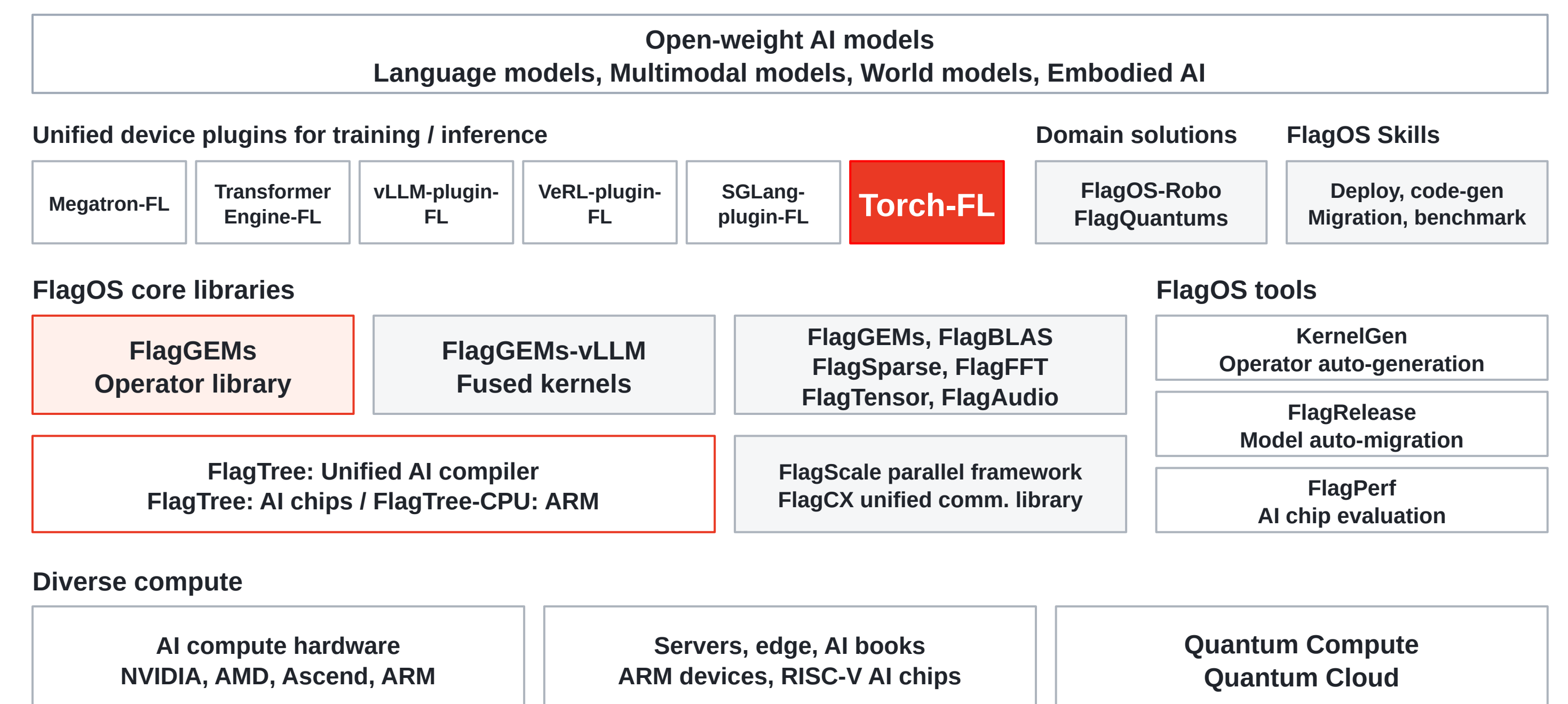
Vendors

30+

Chips

for Multi-Vendor AI chipsets

URL: flagos.io GitHub: github.com/flagos-ai



Day-0: Qwen-Image-2.1

September 20, 2026 8 chip-platform release bundles

Qwen names Torch-FL and FlagGems as its Diffusers portability path and reports inference accuracy aligned with the official implementation. [1]

Full pipeline: Qwen3-VL text encoder, 32-layer DiT with prefix KV cache, and causal 3D VAE. [1, 3]

Measured example: MetaX C550



Seconds per image (lower is better)

1024 × 1024, BF16, 40 steps, batch 1, seed 42. Warmup 1, 40 s timing window.

September 19 measurement. [2]

Paired latents: 46.47 dB PSNR against vendor output. Backend agreement, not image quality. [2]