

DeepSeek open-sources DeepInfer, the inference framework behind its production services

Source copy kept by wheresthe.ai for the story "DeepSeek open-sources DeepInfer, the inference framework behind its production services" (<https://wheresthe.ai/d/cn-deepseek-open-sources-deepinfer-inference-framework-2026-10/>).

Source: <https://github.com/deepseek-ai/DeepInfer>

Text saved 21 Oct 2026 (the page text through a reader, because the site refused a direct fetch).

SHA-256 of the saved text: 2e3a1454770849e55daa362b865603de312917929d0c9ed04d93b63288361184

DeepInfer is the high-performance inference framework that powers DeepSeek's large-scale production services. This open-source release provides offline inference with the core features needed for serving at scale.

Supported models:

* DeepSeek-V4.1-Flash

Supported features:

* Optimized kernels, model and sampler runtime * Parallelism: Attention DP / balanced CP, MoE EP + EPLB * DSpark and confidence-scheduled verification * SWA bounded replay * Asynchronous scheduling (zero-overhead batch scheduler) * Batch-invariant inference * Deterministic sampling * Structured outputs

Feature details and configuration * **Optimized kernels:**FlashMLA, DeepGEMM, DeepEP, TileKernels, DeepSelect, and custom TileLang kernels. * **DSpark and confidence-scheduled verification:** accelerate decoding with draft tokens and use confidence estimates to schedule verification. Set ``DEEPINFER_ENABLE_MTP_VARLEN=1``; ``DEEPINFER_USE_NEXTN_DECODE`` controls the maximum draft length (default: 5). * **Decoder SWA bounded replay:** reduce decoder prefill work by replaying only a bounded suffix. Set ``DEEPINFER_YOCO_PREFILL_TRUNCATION_MODE=compact`` and ``DEEPINFER_YOCO_PREFILL_KEEP_TOKENS=140`` for the default V4.1 settings; the retained suffix must cover the SWA window plus the maximum backtracking length. * **Asynchronous scheduling:** overlap CPU batch scheduling with device execution. Enabled by default. * **Batch-invariant inference:** keep model results identical across different batch compositions. Enable with ``DEEPINFER_BATCH_INVARIANT=1`` (NVIDIA only). * **Deterministic sampling:** produce identical outputs for identical input tokens and sampling parameters. Set ``DEEPINFER_BATCH_INVARIANT=1`` `DEEPINFER_ENABLE_MTP_VARLEN=0`` and pass a fixed ``seq_seed`` to ``SamplingParams`` (NVIDIA only). \* \*\*Structured outputs:\*\* constrain generation to a regex, JSON schema, or EBNF grammar. Pass ``so_args={"schema_type": ..., "schema": ...}`` to ``SamplingParams``; see examples.

Supported hardware:

* NVIDIA SM100 GPUs * Huawei Ascend 950 NPUs

Required for a complete inference service (not included in this release):

* A KV cache system * VE/Prefill/Decode disaggregation (most required engine functionality is already implemented) * Load balancing across data-parallel replicas for both prefill and decode

Getting started

Prerequisites

* Linux * GCC >= 13.0 * Python 3.12 * CUDA 13 (NVIDIA only) * CANN 9.2 (Ascend only)

Build from source

Before building DeepInfer, please create and activate an empty Python 3.12 environment.

Use ``ascend`` instead for the Ascend platform # Keep it exported in the environment where you run DeepInfer
`export DEEPINFER_ACCELERATOR=nvidia ./build.sh`

Offline inference

The default checkpoint, ``deepseek-ai/DeepSeek-V4.1-Flash-1017``, is downloaded into the Hugging Face cache

on the first run; set `DEEPINFER\_HF\_MODEL\_DIR=/path/to/DeepSeek-V4.1-Flash-1017` to run from a local copy instead.

python examples/offline_inference.py

See examples for details.

Performance benchmarks

| Prefill | Decode | | --- | --- | | | |

Measured on 4 SM100 GPUs for prefill and 64 for decode — each within one scale-up domain. To reproduce the benchmarks, see run_sweep.sh.

Evaluation benchmarks

See eval/README.md.

License

DeepInfer is released under the Apache License 2.0. See LICENSE for details.

Acknowledgments

DeepInfer was originally built primarily on vLLM, with additional inspiration from SGLang and TensorRT-LLM. We thank their developers.

Citation

```
@misc{deepinfer2026, title = {DeepInfer: High-performance, large-scale inference framework for DeepSeek models}, author = {DeepSeek-AI}, year = {2026}, publisher = {GitHub}, howpublished = {\url{https://github.com/deepseek-ai/DeepInfer}}, }
```