

DeepSeek releases V4.1-Flash and routes its V4-Pro traffic to it

Source copy kept by wheresthe.ai for the story "DeepSeek releases V4.1-Flash and routes its V4-Pro traffic to it" (<https://wheresthe.ai/d/cn-deepseek-v4-1-flash-2026-09/>).

Source: <https://api-docs.deepseek.com/news/news260910>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: e5d8e62e7c3311327d69b06b1a6ccbe5bd83cb504560d942d08f94ffc9d90063

On this page DeepSeek-V4.1-Flash: Smarter, Faster, More Efficient

 Introducing DeepSeek-V4.1-Flash: smarter, faster, more efficient.

- ◆ Introducing the smallest model in our new architecture family, with native visual understanding.
- ◆ Designed for greater capability, faster inference, higher throughput, and scaling to larger models.

Asymmetric architecture. More intelligence, less cost.

- ◆ 552B-parameter MoE.
- ◆ New Causal Encoder–Decoder architecture: just 8B active parameters for input, 16B for output.
- ◆ New pre-training methods + larger-scale RL post-training deliver benchmark results ahead of flagship models, including DeepSeek-V4-Pro.

 Smaller KV cache. Bigger savings.

Compared with the previous generation, V4.1-Flash's KV cache needs just:

- ◆ 1/4 the HBM
- ◆ 1/8 the SSD storage

Cache-hit charges often account for a large share of agent costs. Compressing the cache cuts those costs significantly.

✂ V4.1-Flash is now live on the DeepSeek API with native multimodal support.

Set your model to `deepseek-flash`.

- ◆ V4-Flash & V4-Flash-Vision-Exp are retired. For compatibility, `deepseek-v4-flash` and `deepseek-v4-flash-vision-exp` temporarily route to V4.1-Flash.
- ◆ Tests by multiple parties put V4.1-Flash ahead of V4-Pro on performance, cost, speed & total runtime. We're phasing out V4-Pro.
- ◆ Starting at 04:00 UTC on Sept 14, 2026, all `deepseek-v4-pro` requests will route to V4.1-Flash at V4.1-Flash rates. This will continue until V4.1-Pro launches.

Official partners WorkBuddy (including CodeBuddy) & OpenCode now fully support V4.1-Flash. Try it today!

 More efficient architecture. Lower API prices.

V4.1-Flash lets us serve more users at a lower cost. We're passing the savings on to you.

- ◆ Peak/off-peak pricing continues to balance demand.
- ◆ Off-peak rates are 50% of peak rates. Schedule flexible workloads off-peak to save.
- ◆ New pricing takes effect at 04:00 UTC on Sept 10, 2026.

 Supporting open source. Expanding deployment options.

We'll work closely with the open-source community on V4.1-Flash inference support and explore more deployment options.

Planning a large-scale deployment with 2,000 GPUs + a storage cluster? Let's talk.

- ◆ Model: <https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash>
- ◆ Paper: https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash/blob/main/DeepSeek_V41_Tech_Report.pdf

Asymmetric architecture. More intelligence, less cost. 📄 Smaller KV cache. Bigger savings. ✂️ V4.1-Flash is now live on the DeepSeek API with native multimodal support. 💰 More efficient architecture. Lower API prices. 🌐 Supporting open source. Expanding deployment options.