DeepSeek V4 Flash Vision: Charts, Tables & Document Understanding

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · deepseek v4 flash vision · deepseek vision · document understanding ai · chart understanding benchmark · table extraction ai · multimodal llm · deepseek v4 flash vision exp · ocr ai model · gpt-4 vision comparison

Executive Summary

DeepSeek released **DeepSeek-V4-Flash-Vision-Exp** on the DeepSeek API Platform on **August 21, 2026**, describing it as its first experimental multimodal model in the DeepSeek-V4 family (api-docs.deepseek.com) (huggingface.co). Built by adding visual modules and continued training to the existing **DeepSeek-V4-Flash** text architecture, the model accepts JPEG, PNG, GIF, and WebP images alongside text and is explicitly marketed for reading screenshots and analyzing charts (api-docs.deepseek.com). This report evaluates that claim specifically for document understanding: charts, tables, forms, and scanned pages, using DeepSeek's own disclosed benchmarks, independent hands-on testing, and cross-checked competitor documentation, since the earlier IntuitionLabs analysis of DeepSeek's inference economics did not cover the vision variant at all (intuitionlabs.ai).

The architecture imposes a hard, published ceiling: every image is resized toward roughly 800x800 pixels (about 0.64 megapixels) and capped at **384 tokens per image** regardless of source resolution (api-docs.deepseek.com) (www.eesel.ai). DeepSeek's own model card discloses only one chart-adjacent benchmark, an internal

"Chartography" evaluation on which the model scored **64.3** against Claude Opus 4.8's 65.0, and an agentic "ApexBench" score of 36.5 versus Opus 4.8's 39.4 (huggingface.co) (huggingface.co); no ChartQA, DocVQA, TableBench, OCRBench, MMMU, or AI2D score appears anywhere in DeepSeek's own materials. Independent hands-on testing included one reviewer's estimate of roughly **97 to 98 percent** field-level accuracy extracting a dense four-column financial table from a low-quality scan (www.mindstudio.ai).

Pricing is identical to the text-only DeepSeek-V4-Flash model: **$0.22 per million input tokens**, **$0.007 per million cached-hit input tokens**, and **$0.66 per million output tokens** off-peak, with images billed as converted input tokens rather than a separate per-image rate, and peak-hour pricing (01:00 to 04:00 and 06:00 to 10:00 UTC, Monday through Friday; all other hours are off-peak) roughly doubling those rates (api-docs.deepseek.com) (api-docs.deepseek.com). Third-party hosts Fireworks AI and OpenRouter both list matching rates, cross-confirming the official figures (fireworks.ai) (openrouter.ai). Against competitors, no vendor, including OpenAI, Google, and Anthropic, has published a directly comparable ChartQA or DocVQA figure for its current flagship vision model; OpenAI's own July 2026 comparison table shows every tested frontier model, including its own GPT-5.6 variants, Claude, and Gemini, scoring below 31 percent on a real-world PDF-parsing evaluation, indicating that structured document parsing remains difficult industry-wide rather than a gap unique to DeepSeek (openai.com). DeepSeek's own documentation confirms no dedicated OCR, PDF-ingestion, or bounding-box mode exists for this model; developers must rasterize documents themselves and validate output empirically (omniakey.com). As of September 2026, DeepSeek-V4-Flash-Vision-Exp is a low-cost, low-resolution, experimental option best suited to lightweight chart and screenshot reading rather than production-grade full-page document OCR.

Introduction and Background

Document understanding, the ability of a model to read charts, extract structured tables, and interpret scanned forms accurately, has become one of the more consequential and least standardized capabilities in applied artificial intelligence (AI). Where text-only large language model (LLM) benchmarks are relatively mature, vision-language benchmarks for documents remain fragmented across academic suites (ChartQA, DocVQA, OCRBench) and vendor-specific agentic evaluations that are not directly comparable to one another. This creates a real risk for buyers: a model's strong text-benchmark reputation does not automatically transfer to its ability to read a chart's axis labels or a table's footnotes.

DeepSeek-V4-Flash-Vision-Exp entered this landscape on **August 21, 2026**, as DeepSeek's first experimental multimodal release, built on the existing **DeepSeek-V4-Flash** architecture with added visual modules and continued training (api-docs.deepseek.com) (fireworks.ai). The company's own release note states the vision variant "matches DeepSeek-V4-Flash on text capabilities," positioning vision as an addition rather than a replacement for the existing text model (api-docs.deepseek.com). A companion agent framework, **DeepSeek Harness 0.1.1**, shipped the same day with built-in support for the model (api-docs.deepseek.com).

This report was commissioned as a distinct, vision-specific companion to IntuitionLabs' earlier analysis of DeepSeek's inference cost structure, which examined the economics of DeepSeek's text models but did not evaluate the vision variant's document-understanding accuracy (intuitionlabs.ai). It does not repeat that cost analysis; instead it asks a narrower, harder question: when this specific model is shown a chart, a scanned table, or a multi-column report page, what does it actually get right, what does it get wrong, and how does that compare to the disclosed and independently tested behavior of competing vision-capable models from OpenAI, Google, Anthropic, and Alibaba. The method throughout is to separate **vendor-published claims** from **independently**

observed evidence, to report benchmark names and numeric scores exactly as published rather than translated into other suites, and to flag every volatile fact (price, benchmark score, availability) with the date it was observed, since all three change quickly for an experimental model still under active revision.

Product Overview: Architecture and Specifications

DeepSeek-V4-Flash-Vision-Exp is a **305-billion-parameter Mixture-of-Experts (MoE)** model, per third-party hosting documentation, that adds a vision encoder and continued multimodal training on top of the DeepSeek-V4-Flash text backbone (fireworks.ai). It is released under an **MIT license**, and only this specific model variant, not the standard deepseek-v4-flash or deepseek-v4-pro models, accepts image input; sending an image to a non-vision endpoint returns an HTTP 400 error (huggingface.co) (api-docs.deepseek.com).

The model's most consequential design decision for document work is its **image-resolution policy** (see above): every submitted image is capped at an upper bound of 384 tokens after resizing, and independent analysis of that policy found the effective target resolution is approximately 800 by 800 pixels, or 0.64 megapixels, regardless of the original scan's dimensions (www.eesel.ai). A separate independent review made the practical implication explicit: fine chart labels and precise screen coordinates can disappear during that resizing step (modelflare.dev). The API otherwise supports a generous batch ceiling of up to **600 images per request**, and image files can be as large as 64 mebibytes (MiB) when referenced through DeepSeek's Files API, versus 32 MiB for inline base64 images (api-docs.deepseek.com) (api-docs.deepseek.com).

Supported formats are limited to **JPEG, PNG, GIF, and WebP**, detected from file content rather than extension or declared MIME type (api-docs.deepseek.com). Notably, and consistent with the resolution cap, DeepSeek's documentation contains **no dedicated PDF-ingestion, OCR, grounding, or bounding-box mode**; an independent guide published after launch confirmed this omission directly against the API reference (www.digitalapplied.com). In practice this means a developer must rasterize a PDF into page images before sending it, and the model's answer format is free-text rather than structured coordinates.

Table 1 below summarizes the model's core specifications as documented across DeepSeek's own API reference and independent hosting platforms.

Attribute	Value	Source
Model ID	deepseek-v4-flash-vision-exp	DeepSeek official docs
Release date	August 21, 2026	DeepSeek official docs (see above)
Parameters	305B (Mixture-of-Experts)	(fireworks.ai)
License	MIT	Hugging Face model card (see above)
Context window	1M tokens (DeepSeek); listed as 1,040k tokens by Fireworks	Fireworks AI listing (see above)
Max output	384K tokens	DeepSeek pricing page
Image formats	JPEG, PNG, GIF, WebP	DeepSeek official docs (see above)
Tokens per image	Capped at 384 (~800x800px effective resolution)	DeepSeek official docs (see above)

Attribute	Value	Source
Max images/request	600	DeepSeek official docs (see above)
Concurrency limit	2,500 (vs. 500 for deepseek-v4-pro)	(api-docs.deepseek.com)
PDF/OCR mode	None documented	(www.digitalapplied.com)

The table summarizes documented specifications for the model.

Performance and Benchmark Analysis

DeepSeek's own model card discloses two multimodal-relevant benchmark rows: **"ApexBench" (Pass@1)**, an agentic tool-use evaluation, and **"Chartography,"** described by third-party aggregators as a chart-reading evaluation (huggingface.co) (www.datalearner.com). DeepSeek-V4-Flash-Vision-Exp scored **36.5 on ApexBench** and **64.3 on Chartography**, against Claude Opus 4.8's disclosed 39.4 and 65.0 on the same two benchmarks in DeepSeek's comparison table.

Crucially, **no standard academic document-understanding benchmark score is disclosed anywhere in DeepSeek's own materials for this model**: not ChartQA, not DocVQA, not TableBench, not OCRBench, not MMMU, not AI2D. A benchmark-aggregator page that tracks the model's full disclosed score set lists only agentic and tool-use evaluations (Terminal-Bench, DeepSWE, NL2Repo, DSBench-Hard, Chartography, ApexBench, and similar), confirming the absence rather than an omission by any single secondary source (www.buildfastwithai.com). Press coverage of the launch was explicit that these figures come from DeepSeek's own test harness with no third-party replication at time of writing (datanorth.ai) (www.mindstudio.ai).

Independent hands-on document testing, as distinct from headline benchmark scores, produced a more mixed and more specific picture. One reviewer's structured test set found the model **"strong at chart interpretation and structured document extraction, but inconsistent on handwriting,"** and reported that on a three-series business chart it correctly extracted specific figures, including a 26-week peak value and an 8x drop between two series, and identified a leading-indicator relationship between the series without being asked (www.mindstudio.ai). On a dense four-column financial table taken from a low-quality scan, the same tester estimated overall extraction accuracy at **roughly 97 to 98 percent**, with subtotal markers and footnote superscripts preserved correctly (www.mindstudio.ai).

Two developer guides published shortly after launch converged on the same practical recommendation: treat the model's document and table output as a **draft requiring verification**, not a deterministic OCR result, and validate accuracy empirically against a representative sample before production use. One guide specifically advises testing at least 20 to 50 representative images and recording field-level OCR accuracy and chart/table interpretation accuracy separately before deployment (omniakey.com) (omniakey.com).

Comparative Landscape: DeepSeek Vision Against GPT, Gemini, Claude, and Qwen-VL

Answering "DeepSeek Vision vs. GPT-4 Vision" requires acknowledging that OpenAI's vision offering has moved well past the original GPT-4V and GPT-4o generations; as of September 2026, OpenAI has introduced **GPT-6 Astra**, which is available through the OpenAI API (openai.com). OpenAI's own developer documentation states its models can perform tasks "such as optical character recognition (OCR), small-object detection, or computer use," but discloses two specific weaknesses directly relevant to charts and reports: models "struggle to understand graphs or text where colors or styles... vary," and accuracy is weaker on non-Latin scripts such as Japanese or Korean (developers.openai.com) (developers.openai.com). OpenAI's own GPT-5.6 launch comparison, dated July 9, 2026, reports an **MMMU Pro** (multimodal reasoning) score of 83 percent for its top GPT-5.6 variant against 80.5 percent for Google's Gemini 3.1 Pro Preview (openai.com). More directly relevant to this report, the same comparison table includes a real-world PDF-parsing evaluation labeled "gdp.pdf," on which **every model tested scored below 31 percent**, including OpenAI's own newest models, Anthropic's Claude models, and Gemini 3.1 Pro Preview (openai.com). Because this comparison table is published by OpenAI, it is vendor-published evidence rather than independent measurement and should not alone establish an industry-wide conclusion.

Independent benchmarking firm Roboflow measured OpenAI's models directly rather than relying on vendor tables: it recorded **88.9 percent accuracy (8 of 9 prompts passed)** on a document-understanding test for GPT-5.5, and found the newer GPT-5.6 "Sol" variant scored a 90.7 percent mean OCR similarity score, essentially matching GPT-5.5's 91.2 percent (blog.roboflow.com) (blog.roboflow.com). No comparably specific independent document-accuracy figure was located for DeepSeek-V4-Flash-Vision-Exp; the closest analogues are the qualitative hands-on tests described in the previous section.

Google's Gemini 2.5 technical report defines its own chart-focused internal evaluation, "BetterChartQA," as covering nine disjoint capability buckets, and positions the Gemini 2.5 family as a model that "excels at multimodal understanding," though this is a vendor characterization rather than an externally reproduced score (storage.googleapis.com). An independent, open academic OCR benchmarking project, whose leaderboard tracks document-parsing robustness across many vision-language models, found that "closed-source models (notably Gemini3-Pro) prove relatively robust" on multilingual document parsing while several open-source alternatives it tested showed sharp accuracy declines on the same tasks, though DeepSeek-V4-Flash-Vision-Exp itself was not confirmed present on that specific leaderboard at the time of this report (raw.githubusercontent.com).

Anthropic's official documentation confirms that all of its current models, Claude Fable 5.1, Opus 5, Sonnet 5, and Haiku 4.5, support image input, but the company's own public materials emphasize coding and agentic benchmarks rather than a published DocVQA- or ChartQA-style document score, so no primary Anthropic figure for chart or table extraction accuracy could be cited here; every Claude vision comparison found in this research originated from a competitor's table, not Anthropic's own materials (platform.claude.com).

Alibaba's open-weight **Qwen2.5-VL-72B-Instruct** is the most directly comparable model in terms of published academic benchmark transparency: its model card states the model is "highly capable of analyzing texts, charts, icons, graphics, and layouts within images" and publishes a full comparison table against GPT-4o, Claude 3.5 Sonnet, Gemini-2-flash, and InternVL2.5-78B on the exact benchmarks DeepSeek omits (huggingface.co). On that table, Qwen2.5-VL-72B scored **96.4 on DocVQA_VAL** (versus GPT-4o's 91.1), **89.5 on ChartQA_TEST** (versus

GPT-4o's 86.7 and Claude 3.5 Sonnet's 90.8), and **885 on OCRBench** (versus GPT-4o's 736) (huggingface.co) (huggingface.co) (huggingface.co). These are Qwen's own selected comparison figures, so they should be read as a vendor table rather than independent proof.

Reception and Community Perspectives

The following section reports developer and community sentiment, which is qualitative and anecdotal by nature and is presented as such rather than as measured evidence. The model's release generated substantial developer interest: coverage citing Hacker News activity reported the launch post **reached 458 points on release day**, framed as reflecting long-standing community demand for a DeepSeek vision option (codepick.dev). One reviewer noted that before launch, a frequently repeated question in DeepSeek's own agent-framework community discussion was how to add image support to the toolchain, illustrating pent-up demand rather than a reaction to the shipped product itself (medium.com). The Hugging Face repository for the model has recorded over 133,000 downloads, an early adoption signal independent of any qualitative review (huggingface.co).

Community feedback on document-specific use cases was more mixed than the general launch reception. Reviewers reported the model could grasp a screenshot's overall gist while still missing small but consequential details, such as a specific error message buried in a user-interface screenshot (www.buildfastwithai.com), and separately flagged weaker results on dense enterprise-software (ERP) screenshots than on simpler charts or forms (www.buildfastwithai.com). A hands-on comparison of webpage-screenshot replication found the model reproduced source layout, spacing, and proportions more faithfully than Claude Opus 5 in that specific test, even as reviewers judged Opus 5's visual polish superior overall, an interesting split result suggesting DeepSeek's strength may lie more in structural layout fidelity than in aesthetic rendering (medium.com).

Data Analysis and Evidence

This section consolidates the quantitative facts governing cost and access, the two variables that most directly determine whether DeepSeek-V4-Flash-Vision-Exp is viable for a document-processing workload at scale, alongside the benchmark figures already introduced above.

Pricing, observed September 5, 2026, is identical across the vision and text variants of DeepSeek-V4-Flash: **$0.22 per million input tokens** (cache miss), **$0.007 per million input tokens** (cache hit), and **$0.66 per million output tokens**, all off-peak; DeepSeek's documentation states images are "converted into tokens based on their dimensions and billed as input tokens" rather than carrying a distinct per-image price (api-docs.deepseek.com). Peak-hour pricing is defined as 01:00 to 04:00 and 06:00 to 10:00 UTC, Monday through Friday; all other hours are off-peak, and roughly doubles these rates (api-docs.deepseek.com). The 384-token cap is an upper bound per image, and actual image-token usage depends on image dimensions.

Table 2 below cross-checks the official DeepSeek pricing against two independent hosting platforms.

Provider	Input $/M tokens	Cached input $/M tokens	Output $/M tokens	Context window	Source
DeepSeek (official)	$0.22	$0.007	$0.66	1M tokens	DeepSeek official docs (see above)

Provider	Input $/M tokens	Cached input $/M tokens	Output $/M tokens	Context window	Source
Fireworks AI	$0.22	$0.007	$0.66	1,040k tokens	(fireworks.ai)
OpenRouter	$0.2156 (discounted)	not separately listed	$0.6468 (discounted)	1M tokens	(openrouter.ai)

The near-exact agreement between DeepSeek's own listed price and two independent hosts is a useful cross-check: it confirms the headline rate is genuine platform pricing rather than a promotional or introductory figure limited to one channel. OpenRouter's modest discount reflects routing-level competition among the multiple infrastructure providers, including DeepSeek itself and Fireworks, that OpenRouter lists as backing the same model (openrouter.ai).

On the benchmark side, the clearest quantitative pattern across every source examined for this report, vendor and independent alike, is that **no organization, including DeepSeek, OpenAI, Google, or Anthropic, has published a fully reproducible, independently verified ChartQA/DocVQA/OCRBench-equivalent score specifically for a document use case that matches how these models are used in practice**. OpenAI's own "gdp.pdf" real-document evaluation is vendor-published evidence rather than an independent measurement; it should not alone establish an industry-wide conclusion (openai.com). Where a genuinely comparable academic benchmark table does exist, Qwen2.5-VL's own published comparison (see Comparative Landscape above), an open-weight competitor's scores (up to 96.4 on DocVQA_VAL, 89.5 on ChartQA_TEST, 885 on OCRBench) sit above the GPT-4o-generation baseline it compares against, though this remains a vendor-selected table rather than a neutral leaderboard.

Implications and Future Directions

For an organization evaluating DeepSeek-V4-Flash-Vision-Exp for document-heavy workflows, chart summarization, table digitization, form triage, the evidence in this report points to a specific and narrow fit rather than a general-purpose replacement for dedicated OCR or document-intelligence tooling:

- **Reasonable fit.** High-volume, low-stakes tasks such as bulk screenshot triage, rough chart summarization for internal dashboards, or a first-pass filter ahead of human review, where the model's low image-token cost that varies with image dimensions and generous 600-image batch limit are genuine advantages.
- **Additional evaluation may be appropriate.** Organizations should assess their own validation requirements before using a model for document extraction.

This is precisely the evaluation discipline that governed AI adoption in regulated industries already requires for any model, general-purpose or specialized. As IntuitionLabs' own published framework for enterprise AI information architecture describes it, a defensible deployment "retrieve[s] relevant passages with metadata and inspectable citations so qualified users can verify what supports an output," treating model output as evidence to be checked rather than an answer to be trusted outright (intuitionlabs.ai). Applied to a vision model like this one, that means building an evaluation set from an organization's actual document types, scanned tables, specific chart styles, house-format reports, rather than assuming a headline chart-reading benchmark transfers to a specific use case; the objective, as that same framework puts it, "is not to make a probabilistic system appear certain. It is to make verification practical" (intuitionlabs.ai).

Looking forward, three developments would materially change this assessment:

- **Independent benchmark replication.** An independently reproduced ChartQA, DocVQA, or OCRBench score for this specific model would enable more direct comparisons against Qwen2.5-VL and the GPT and Gemini families.
- **A higher resolution ceiling.** An increase beyond the current 384-token, 800x800-pixel cap, plausible given DeepSeek's stated "experimental" label and rapid prior release cadence, would directly address the most consistently cited limitation across every independent review examined.
- **Industry-wide document-parsing progress.** OpenAI's own sub-31-percent "gdp.pdf" result suggests document understanding, unlike many text-only benchmarks, remains an open research problem for the entire sector, not a race DeepSeek is uniquely behind in.

Frequently Asked Questions (FAQs)

Is DeepSeek V4 Flash Vision good at reading charts? Independent hands-on testing (discussed above under Performance and Benchmark Analysis) found it could extract specific figures directly from a business chart and identify relationships between data series, but its own vendor-disclosed "Chartography" benchmark score of 64.3 (see above) sits close to, though slightly behind, Claude Opus 4.8's 65.0 on the same internal evaluation, and no standard ChartQA score has been published for it.

How accurate is DeepSeek Vision at extracting tables? One independent test estimated roughly 97 to 98 percent field-level accuracy on a dense financial table from a low-quality scan (discussed above), but this is a single documented test rather than a benchmark score, and developers are advised to validate accuracy on their own representative samples before production use (omniakey.com).

Does DeepSeek Vision have a dedicated OCR mode? No. DeepSeek's own API documentation contains no named OCR, PDF-ingestion, or bounding-box capability; an independent guide confirmed this omission directly against the reference documentation (www.digitalapplied.com).

How does it compare to GPT-4 Vision for document understanding? OpenAI has introduced GPT-6 Astra, which is available through the OpenAI API (openai.com). OpenAI's historical GPT-5.6 comparison table shows every model tested, including OpenAI models, scoring below 31 percent on a real-document PDF-parsing test; it should not be treated as a comparison of OpenAI's current model (openai.com).

What is the best multimodal LLM for document understanding right now? No source reviewed for this report identifies a single model as conclusively best; the most benchmark-transparent open-weight option is Qwen2.5-VL-72B (see Comparative Landscape above), which publishes DocVQA, ChartQA, and OCRBench scores directly, while DeepSeek, OpenAI, and Google either omit these specific benchmarks or report internal alternatives that are not directly comparable.

What does DeepSeek V4 Flash Vision cost? As of September 5, 2026: $0.22 per million input tokens, $0.007 per million cached-hit input tokens, and $0.66 per million output tokens off-peak, with images billed as converted input tokens, confirmed identically by DeepSeek's own pricing page and two independent hosts (see Table 2 above).

Conclusion

DeepSeek-V4-Flash-Vision-Exp is a genuine, if narrowly specified, entry into document-adjacent multimodal AI: officially released August 21, 2026, priced identically to DeepSeek's text-only Flash model, and explicitly marketed for reading screenshots and analyzing charts. The evidence gathered here supports a specific, bounded conclusion rather than a sweeping one. Its architecture, an 800x800-pixel effective resolution and a 384-token-per-image ceiling, makes it cost-efficient and fast for lightweight visual tasks, and independent testing found real competence on clean charts and moderately dense tables.

Readers evaluating this model, or any competing vision model, for chart, table, or document work should validate accuracy against their own representative documents before relying on any single figure in this report or elsewhere.
