ChatGPT

Benchmarking Top Image Generation Models (November 2025)

Six Popular & Diverse Models on Hugging Face

We selected six of the most downloaded or liked text-to-image models on Hugging Face as of Nov 2025, ensuring diversity in style focus and developer origins. Each model's key characteristics, popularity stats, and specialties are outlined below.

Sample outputs from FLUX.1 [dev] (Black Forest Labs), a 12 billion-parameter flow-based transformer model. **FLUX.1 [dev]** – Developed by Black Forest Labs, FLUX.1-dev is a cutting-edge open-weight model (12B params) known for exceptional output quality. It achieves state-of-the-art prompt adherence and even renders readable in-image text accurately [1]. In benchmarks it **outperforms proprietary models like DALL-E 3 and Midjourney 6** [1]. FLUX.1-dev has amassed **11.9k likes** on Hugging Face [2] and saw **~1.51 million downloads in the last month** [3], reflecting its popularity. (License: non-commercial license for dev version.) Key features include a **rectified flow transformer** architecture (not diffusion-based) and training via guidance distillation for efficiency [4]. The result is best-in-class image fidelity across diverse subjects – from faces and landscapes to crisp typography [1].

Demo images from HiDream I1-Full (HiDream.ai), a 17 billion-parameter diffusion model excelling at photorealism and text rendering. **HiDream I1 (Full)** – Released by HiDream.ai in 2025, HiDream-I1-Full is a **17 billion-parameter** open-source diffusion model focusing on photorealistic quality and fast generation [5]. It achieved **state-of-the-art results** on internal benchmarks, with superior image quality across styles (photographic, artistic, etc.) and **industry-leading prompt following** fidelity [6]. Notably, HiDream-I1 excels at generating **legible text and logos within images**, outperforming other open models on text-specific evaluations. It's fully open under an MIT license, enabling commercial use [7]. HiDream-I1-Full is newer and a bit less viral (≈40k downloads last month [8], ~1k likes [9]), but it represents the bleeding edge of open generative models. (HiDream also provides distilled "fast" variants for efficiency.)

Stable Diffusion XL 1.0 – The flagship model from Stability AI, **Stable Diffusion XL (SDXL)** is a **latent diffusion** text-to-image model known for its versatility and photorealistic outputs. SDXL uses an ensemble of diffusion networks (base + refiner) to produce high-resolution 1024×1024 images [10] [11]. With an **OpenRAIL++ license** and millions of users, SDXL has become a workhorse in the community. On Hugging Face it has **~2.76 million downloads per month** [12] and ~7.1k likes [13], making it one of the most widely used models. It's adept at a broad range of tasks (portraits, landscapes, fantasy scenes, etc.) with a neutral default style and supports further fine-tuning. However, earlier SD versions struggled with in-image text; this was later addressed by Stable Diffusion 3 which introduced improved typography [14]. Overall, SDXL offers a strong **general-purpose foundation** with an unmatched ecosystem of extensions and optimizations built around it.

Example outputs from SDXL Lightning (ByteDance), a distilled 2–4 step diffusion model for ultra-fast generation. **SDXL Lightning** – Created by ByteDance's AI lab, SDXL-Lightning is a distilled version of Stable Diffusion XL designed for **lightning-fast image generation**. Through progressive distillation, ByteDance reduced SDXL's sampling to as few as **2 diffusion steps**, enabling **1024×1024 images in under a second** [15] on high-end GPUs. Despite the drastic speedup, SDXL Lightning maintains surprisingly high image quality – it outperforms earlier quick-sampling models like SDXL "Turbo" in detail and prompt fidelity [15]. This model (released with an OpenRAIL++ license) has gained ~2.1k likes and saw **~112k downloads last month** [16] as practitioners experiment with its speed. It's especially useful for scenarios needing rapid image previews or real-time generation, at some cost to absolute image perfection. The repository provides 1, 2, 4, and 8-step checkpoints (the 2-step and 4-step being the best quality-speed tradeoffs). SDXL Lightning showcases ByteDance's contribution to open-source diffusion by pushing **inference efficiency** to new extremes.

Outputs from Waifu Diffusion (hakurei/harubaru), an anime-styled Stable Diffusion fine-tune popular for its specialized aesthetic. **Waifu Diffusion v1.4** – An **anime-focused** image generation model, Waifu Diffusion is a fine-tuned Stable Diffusion 1.x model tailored to produce **Japanese anime/manga style** outputs. Created by community developers (hakurei, a.k.a. harubaru), it has become hugely popular in the anime art community. In fact, *Waifu Diffusion was the most downloaded custom model on Hugging Face*, with over **176k downloads in one month** at its peak [17], far surpassing other style-specific models. It has ~2.5k likes on HF and is widely used in anime fan-art generation and NSFW content (users note it tends to generate adult content if not guided carefully [18]). Architecturally, Waifu Diffusion is the same latent diffusion backbone as Stable Diffusion, but fine-tuned on anime image datasets (e.g. Danbooru). This specialization allows it to **excel at anime faces, characters, and illustrations**, producing on-model styles that would be difficult for a generic model to replicate. Organizations and hobbyists looking for anime-style renders often choose this model for its quality in that domain.

DeepFloyd IF – Developed by Stability AI's DeepFloyd team, IF is a novel **two-stage diffusion model** that rivals the quality of Google's Imagen. It uses a frozen T5-XXL text encoder and a **cascade of three diffusion models** (base 64px, upsamplers to 256px and 1024px) [19] [20]. The result is **exceptional photorealism and language understanding** – DeepFloyd IF can **render written text within images** (e.g. signs, logos) far more reliably than prior open models [20] [21]. For example, IF achieved a **zero-shot FID of 6.66 on COCO** (a very low score indicating highly realistic outputs) [22]. The model is available on Hugging Face with a research license (initially quite restrictive), and while its usage is somewhat gated, it has ~11k users and significant buzz in the community. DeepFloyd IF essentially provides an **open-source alternative to closed models like DALL-E 3**, especially shining in cases that involve **complex prompts or text rendering in images** [19]. Its multi-step pipeline is heavier to run, but rewards with fidelity that was previously unseen in open models. (In practice, IF is often used for generating things like logos with specific text, which it handles with high accuracy.)

Key Parameters to Vary in Benchmarks

To conduct an insightful benchmarking study on an NVIDIA RTX 5090 (which supports TensorFloat-32 precision), we recommend varying the following technical parameters for each model:

- **TF32 Precision (on vs. off):** TF32 is a tensor core precision mode that can accelerate throughput on newer NVIDIA GPUs. Enabling TF32 for matrix ops often speeds up diffusion inference significantly, with minimal impact on image quality. Comparing *TF32 enabled* vs. *standard FP32* will show how much faster generation runs with tensor cores and whether there's any subtle quality difference. (In code, this corresponds to settings like `torch.backends.cuda.matmul.allow_tf32` [23] [24].) This test highlights the 5090's capability to trade a bit of precision for speed.

- **Number of Diffusion Steps:** The number of denoising steps (iterations) in the diffusion process greatly affects both output quality and latency. Benchmark a **lower step count vs. a higher step count** (e.g. 20 vs 50 steps) to quantify the quality trade-off. Fewer steps yield faster images but can appear grainier or less faithful, while more steps improve fidelity up to a point of diminishing returns. Tracking metrics like FID can reveal how image quality improves with more steps [25]. This parameter is critical for analyzing speed-quality efficiency of each model.

- **Guidance Scale (CFG Scale):** The classifier-free guidance scale influences how strongly the model follows the text prompt. Testing different guidance values (say, 5 vs 8 or a range) is insightful: *Lower guidance* tends to produce more creative or varied outputs, while *higher guidance* forces closer adherence to the prompt at risk of visual artifacts or loss of diversity. By varying guidance, you can measure effects on image-text alignment (e.g. via CLIP score) and aesthetic appeal. This helps identify the optimal guidance for each model (some models may prefer lower CFG for best results [26]). Monitoring CLIP similarity vs. guidance is one way to automate this analysis.

(Optionally, you might also explore image resolution (e.g. 512×512 vs 1024×1024) if resources permit. Higher resolution outputs stress the models' performance and memory, revealing how each scales. However, in this study the generation resolution is capped at 768×768 for consistency [27].)

Automated Image Quality Evaluation Metrics

When benchmarking image generation, it's important to use **automated metrics** to compare quality without human bias. We suggest computing the following for each model's outputs:

- **CLIP Score (Image-Text Similarity):** CLIP score measures how well an output image matches its prompt by using a pretrained CLIP model to encode both and calculating cosine similarity [28]. A higher CLIP score means the image is more semantically aligned with the text prompt. This is a good proxy for prompt adherence and relevance. CLIP-based metrics correlate reasonably with human judgment of alignment [29]. For each generated image, we can compute the CLIP similarity to its prompt; then average these scores per model or prompt type.

- **Fréchet Inception Distance (FID):** FID is a popular metric to quantify visual quality by comparing the distribution of generated images to that of real images. It uses an Inception-v3 network to get feature vectors for images, then computes the Fréchet distance between the real vs generated feature distributions [30]. Lower FID indicates the synthetic images are closer to real imagery. In our context, we might compute FID by using a large set of prompts with outputs from each model and comparing against a real image set (if available) or at least compare models' output distributions. FID is sensitive to diversity and visual fidelity, and is standard for generative model evaluation. (We should generate a sufficiently large sample per model – e.g. 500 images – for meaningful FID calculation [25].)

- **Aesthetic Score:** This is an attempt to quantify how "visually pleasing" an image is. One approach is to use a model like the LAION-Aesthetic Predictor, which is a lightweight regression on CLIP embeddings trained to predict human aesthetic ratings [31]. It outputs a score (typically 1–10) for image aesthetics. By averaging the **aesthetic score** of outputs, we can see which model tends to produce more beautiful or artistically appealing images. This metric doesn't consider prompt alignment, but rather composition, color, and other subjective qualities as learned from human preference data [31]. It's fully automated – for example, LAION's model was trained on a dataset where people rated images on a 1–10 scale [31]. In our benchmark, we can compute an aesthetic score for each image and summarize the results (mean, distribution) per model.

Other possible metrics include **Inception Score (IS)** for realism/diversity and specialized scores for face fidelity or structural consistency, but CLIP, FID, and aesthetic scores should suffice for an automated evaluation of prompt fidelity, realism, and attractiveness respectively. Together, these metrics provide a well-rounded picture: CLIP score for text relevance, FID for realism (compared to real data), and aesthetic score for overall visual appeal.

Diverse Prompt Set (JSON)

To benchmark fairly, we curated **50 diverse text prompts** covering a range of content categories. Each prompt is labeled with a category tag indicating its type: `"text-to-image"`, `"animals"`, `"people"`, or `"landscapes"`. The **"text-to-image"** prompts specifically test the models' ability to generate legible text (e.g. signs, logos), **"animals"** prompts involve various creatures in different scenes, **"people"** prompts cover portraits and human activities, and **"landscapes"** prompts are about natural or scenic environments. The prompts vary in complexity and style – some are simple descriptions, others are more detailed or imaginative. This helps ensure the benchmark explores different strengths (e.g. realism, anatomy, typography, composition).

Below is the JSON array of 50 prompt objects, each with its text and category. These will be used for generating images from each model:

```json
[
  {"prompt": "a logo that says 'Landscaping Pros' in green sans-serif font", "category": "text-to-image"},
  {"prompt": "a tiger leaping across a river at sunset", "category": "animals"},
  {"prompt": "a portrait of an elderly woman laughing, wearing traditional Indian attire", "category": "people"},
  {"prompt": "a serene beach with palm trees at sunrise, waves gently crashing", "category": "landscapes"},
  {"prompt": "a neon sign that reads 'Open 24/7' on a brick wall, at night", "category": "text-to-image"},
  {"prompt": "two golden retriever puppies playing in a grassy field", "category": "animals"},
  {"prompt": "a futuristic cyborg girl with neon circuit patterns on her face", "category": "people"},
  {"prompt": "a misty forest in autumn with sunlight beaming through the trees", "category": "landscapes"},
  {"prompt": "a book cover with the title 'The Lost City' in ornate gold lettering", "category": "text-to-image"},
  {"prompt": "a colorful parrot flying over a tropical jungle", "category": "animals"},
  {"prompt": "a group of friends taking a selfie at a beach during sunset", "category": "people"},
  {"prompt": "a desert landscape with towering sand dunes under a starry night sky", "category": "landscapes"},
  {"prompt": "a road sign shaped like a cat silhouette, with the word 'Caution' underneath", "category": "text-to-image"},
  {"prompt": "a majestic elephant standing in front of Mount Kilimanjaro", "category": "animals"},
  {"prompt": "an astronaut floating in space, with Earth visible in the visor reflection", "category": "people"},
  {"prompt": "snow-covered mountains reflecting in a crystal clear lake at dawn", "category": "landscapes"},
  {"prompt": "a white coffee mug with the text 'World's Best Dad' printed in blue", "category": "text-to-image"},
  {"prompt": "a cat dressed as an astronaut floating in space", "category": "animals"},
  {"prompt": "a medieval knight in armor standing before a castle gate", "category": "people"},
  {"prompt": "a bustling city skyline at night, illuminated by skyscraper lights", "category": "landscapes"},
  {"prompt": "an album cover featuring the text 'Love & Time' in handwritten style, with abstract background", "category": "text-to-image"},
  {"prompt": "a school of fish swimming around a coral reef", "category": "animals"},
  {"prompt": "a child holding a balloon and looking up at the sky, cartoon style", "category": "people"},
  {"prompt": "an alien planet landscape with two moons and purple rocky terrain", "category": "landscapes"},
  {"prompt": "a movie poster with the title 'Galactic Odyssey' at the top and stars in the background", "category": "text-to-image"},
  {"prompt": "a dragon breathing fire atop a mountain", "category": "animals"},
  {"prompt": "a business meeting with five people around a conference table, whiteboard in background", "category": "people"},
  {"prompt": "a waterfall cascading into a jungle pool with rainbows in the mist", "category": "landscapes"},
  {"prompt": "a colorful parrot perched on a pirate's shoulder", "category": "animals"},
  {"prompt": "Mona Lisa painting but with a robotic face, digital art", "category": "people"},
  {"prompt": "rolling hills of lavender fields under a golden sunset", "category": "landscapes"},
  {"prompt": "a panda bear wearing sunglasses and eating bamboo", "category": "animals"},
  {"prompt": "a firefighter rescuing a kitten from a burning house", "category": "people"},
  {"prompt": "an underwater coral reef scene with sunlight filtering from above", "category": "landscapes"},
  {"prompt": "a hummingbird hovering next to a red flower", "category": "animals"},
  {"prompt": "a fashion model walking on a runway wearing a vibrant avant-garde
```

```
dress", "category": "people"},
  {"prompt": "a volcanic eruption at dusk, with lava flowing down a mountain",
"category": "landscapes"},
  {"prompt": "a white rabbit in a magician's hat on stage", "category":
"animals"},
  {"prompt": "a close-up portrait of a man with tribal face paint, high
detail", "category": "people"},
  {"prompt": "a tranquil Japanese garden with a red bridge over a koi pond",
"category": "landscapes"},
  {"prompt": "a wolf howling at the full moon in a dark forest", "category":
"animals"},
  {"prompt": "two medieval warriors dueling with swords in a forest",
"category": "people"},
  {"prompt": "a frozen tundra with polar lights dancing in the sky",
"category": "landscapes"},
  {"prompt": "a peacock displaying its colorful feathers in a garden",
"category": "animals"},
  {"prompt": "a couple dancing tango under a street lamp at night", "category":
"people"},
  {"prompt": "a canyon with stratified red rock formations and a river at the
bottom, midday", "category": "landscapes"},
  {"prompt": "a mouse in a tuxedo standing on a stage", "category": "animals"},
  {"prompt": "a pirate captain at the helm of an old ship, oil painting style",
"category": "people"},
  {"prompt": "a ballerina performing on stage with dramatic lighting,
photography", "category": "people"},
  {"prompt": "a family of four having a picnic in a park on a sunny day",
"category": "people"}
]
```

(The JSON above can be saved as `prompts.json`*. Each prompt's index in the array can serve as an ID for logging results later.)*

We have intentionally included some prompts that challenge specific capabilities – for example, **logos or signs with text** (to test models like DeepFloyd IF or HiDream on text rendering), prompts with **animals** and fine details (to see how models handle fur, feathers, etc.), prompts with **people** in various styles (to evaluate photorealism, face accuracy, or stylization of humans), and diverse **landscapes** (testing composition, lighting, and environment details). This ensures the benchmark isn't overly skewed to one model's strength.

Logging and Output Format for Benchmark Results

For each model and each prompt (and each setting variation), the system should record key results so we can analyze performance and quality quantitatively. We recommend logging the following for **every generation run**:

- **Model Name** (or ID)

- **Prompt ID** (index from the JSON above, or some identifier for the prompt)
- **Prompt Category** (optional, could be derived from ID via the JSON)
- **Parameters:** e.g. number of steps, guidance scale used, resolution, and whether TF32 was enabled. These context fields ensure we know exactly what settings produced the result.
- **Generation Time**: how long it took to generate the image (in seconds or milliseconds). This is crucial for throughput benchmarking.
- **CLIP Score**: the image-text similarity for the output (as discussed, using a CLIP model). This helps measure prompt adherence.
- **Aesthetic Score**: the predicted aesthetic quality rating of the image (using an automated model).
- (Optional) **Other metrics**: e.g. a flag if safety filter triggered or if NSFW content was detected (since the notes indicate black images or NSFW might be rewritten [32] – logging such events could be useful to evaluate how often each model produces unsafe content).

Format: It's best to output these records in a **CSV file** (comma-separated values) or another structured table format. CSV is human-readable and easily importable into spreadsheets or databases. Each row would represent one image generation result. For example:

```
model,prompt_id,category,steps,guidance,TF32,resolution,time_ms,clip_score,aesthetic_score
flux_dev,0,text-to-image,28,5.0,True,768,1250,0.312,6.5
flux_dev,0,text-to-image,28,5.0,False,768,1470,0.315,6.5
... (and so on)
```

In the above schema, we list the model, the prompt id (0 corresponds to `"a logo that says 'Landscaping Pros'..."` in the JSON), the category, the steps and guidance used, whether TF32 was enabled, resolution (pixels), the generation time in milliseconds, and the computed CLIP and aesthetic scores. We would have one line for each run (so if we generate each prompt with each model under a couple of different settings, that could be many lines – easily handled in a CSV).

This **CSV** can be directly opened in analysis tools or imported into a SQL database. If using SQL, one could create a table with columns for each of those fields and insert the records accordingly. The data is "SQL-ready" since it's in a structured, delimited text format (just ensure to escape commas or newlines in the prompt text if you log the full prompt; using the prompt ID as we suggest avoids that issue).

For clarity, you might not include the full prompt text in each row (to avoid very large CSV cells or formatting issues). Using an ID that links back to the prompt JSON is a cleaner approach. Analysts can join on the prompt ID to get the text and category from the JSON if needed. Alternatively, you could output JSON Lines (each line a JSON object) with the same info, but CSV is simpler for many.

Example entry: For a single image generation: Model `HiDream-I1`, Prompt id 21 ("an album cover...'Love & Time'..."), 20 steps, guidance 4.0, TF32 enabled, 768×768 resolution, took 1100 ms, got CLIP score 0.28 and aesthetic 7.0 – the CSV line would look like:

```
HiDream-I1,21,text-to-image,20,4.0,True,768,1100,0.281,7.02
```

This systematic logging will allow you to create pivot tables or SQL queries – for example, **average time per model**, **average CLIP score per model**, distribution of aesthetic scores per category, etc. It also facilitates plotting (time vs steps, CLIP vs guidance, etc.).

Finally, ensure to **record any anomalies** (like if a model failed on a prompt or produced an inappropriate image that was filtered). These could be noted in additional columns or a separate log, so that the final data can be filtered to only successful generations for metric calculations.

By collecting these results and metrics, you will have a rich dataset to power your technical blog post. In the blog, you can highlight insights such as “Model A was 30% faster than Model B on average [33] with minimal quality loss when using TF32,” or “Model X achieved the highest CLIP score on text-to-image prompts, confirming its strength in handling text in images,” etc. You’ll also be able to include charts (like a bar chart of generation time per model, or a table of average CLIP/Aesthetic scores) to back up your thought leadership analysis.

With this comprehensive benchmarking input – the **diverse prompts**, the **varied parameters** tested, the **quantitative metrics**, and the **top models** spanning different niches – your blog post can confidently discuss the state of generative vision models in late 2025 and provide valuable guidance to the community on choosing the right model for the right task.

Sources:

- Hugging Face model cards and stats for popularity: FLUX.1-dev [3] [1], HiDream-I1 [6] [5], Stable Diffusion XL [12], SDXL Lightning [15] [16], Waifu Diffusion [17], DeepFloyd IF [19].
- Evaluation metrics definitions: CLIP score (measures image-caption compatibility) [28], FID (compares generated vs real distribution) [30], aesthetic scoring via LAION model [31].
- Technical notes on using TF32 and guidance/steps: internal study notes [23] [26] and community discussions [25].

---

[1] [14] [15] The best open-source image generation model
https://www.baseten.co/blog/the-best-open-source-image-generation-model/

[2] [3] [4] black-forest-labs/FLUX.1-dev · Hugging Face
https://huggingface.co/black-forest-labs/FLUX.1-dev

[5] [6] [7] [8] [9] HiDream-ai/HiDream-I1-Full · Hugging Face
https://huggingface.co/HiDream-ai/HiDream-I1-Full

[10] [11] [12] [13] stabilityai/stable-diffusion-xl-base-1.0 · Hugging Face
https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

[16] ByteDance/SDXL-Lightning · Hugging Face
https://huggingface.co/ByteDance/SDXL-Lightning

[17] [18] Best Custom (Fine-Tuned) Stable Diffusion Models | Blog
https://stablecog.com/blog/best-custom-fine-tuned-stable-diffusion-models

[19] DeepFloyd IF
https://huggingface.co/docs/diffusers/main/en/api/pipelines/deepfloyd_if

[20] [21] [22] IF — DeepFloyd: Pixel-Based Text-To-Image Generation Model | by Natasha | AI Vanguard | Medium
https://medium.com/ai-vanguard/if-deepfloyd-pixel-based-text-to-image-generation-model-6262f0a5469f

[23] [24] [26] [27] [32] [33] vision-models-study.md
file://file_000000005d14722f9d634f5b26f8f298

[25] [28] [29] [30] Performance Metrics in Evaluating Stable Diffusion Models | by BugendaiTech Deutschland GmbH | Medium
https://medium.com/@seo.germany/performance-metrics-in-evaluating-stable-diffusion-models-4ca8bfdcc2ba

[31] LAION-Aesthetics | LAION
https://laion.ai/blog/laion-aesthetics/