GENSPECTRUM: A MULTI-MODEL, MULTI-TASK TEXT-TO-IMAGE CORPUS

Juncheng Yan[*1,2], **Yifei Li**[*1,2], **Yanran Zhang**[1], **Yu Zheng**[1], **Borui Zhang**[1], **Bin Fu**[2], **Yu Qiao**[2], **Jiwen Lu**[1]

[1]Tsinghua University [2]Shanghai AI Laboratory

[*]Equal contribution.

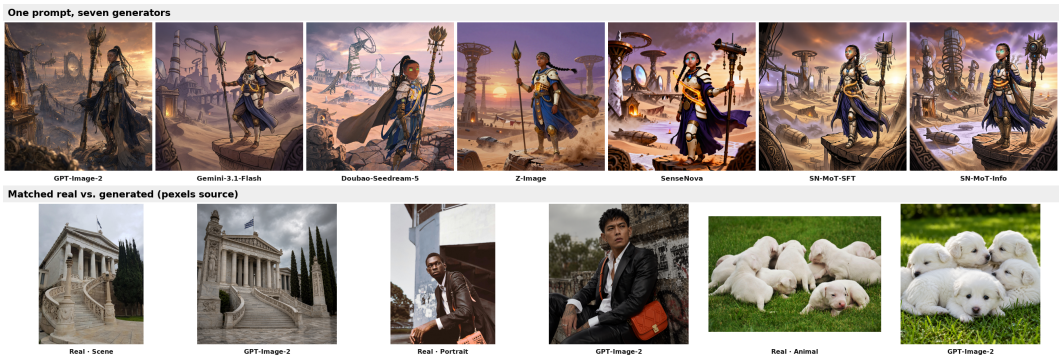

Figure 1: **What GenSpectrum captures.** *Top:* one prompt rendered by seven of GenSpectrum's eight generators, which diverge markedly in composition, palette, and fidelity. *Bottom:* real photographs paired with the image a generator produced from their captions (the `pexels` source). These are GenSpectrum's two signatures, cross-model alignment and matched real/generated pairs. More galleries appear in Appendix B.

ABSTRACT

The rapid progress of text-to-image (T2I) generation is increasingly driven by frontier open and proprietary systems, yet public resources for systematically studying their behavior remain limited. Existing corpora often focus on a single generator, cover outdated models, or lack prompt-aligned real images, which makes it hard to compare modern generators or to analyze why they succeed and fail. We present GenSpectrum, a large-scale prompt-aligned dataset and benchmark for evaluating frontier T2I models. The first release contains 9,817 prompts: 7,818 generation-oriented prompts and 1,999 derived from real photographs. We render them through eight state-of-the-art generators (GPT-Image-2, Gemini-3.1-Flash, Doubao-Seedream-5.0, Z-Image, Qwen-Image-2.0-Pro, and three SenseNova variants), yielding 72,523 images in total: 70,525 generated and 1,998 real. All samples are organized under a unified schema with prompt metadata and content-domain annotations.

GenSpectrum serves as both a data release and a benchmark for interpretable comparison of modern image generators. Its prompt-aligned structure supports controlled evaluation along complementary dimensions, including visual quality and aesthetics, prompt adherence, in-image text rendering, realism, AI-generated image detection, generator attribution, and domain-specific failure analysis. As an initial study, we benchmark 16 detectors (8 discriminative and 8 generative) on the prompt-matched real and generated images. No single detector handles every generator: the strongest, DDA, reaches 0.87 mean balanced accuracy, while detectors that look solved on older generators, such as SAFE, fall to near chance on the two newest, Gemini-3.1-Flash and Doubao-Seedream-5.0. We further provide evaluation protocols for comparing generator strengths and weaknesses across prompt categories, enabling more mechanistic analysis of where specific models succeed

or fail on particular visual or compositional dimensions. We release GenSpectrum as the first version of a living dataset and benchmark, and will expand it with more generators, larger scale, image-editing pairs, video generation data, and broader evaluation tasks.

1 INTRODUCTION

Text-to-image (T2I) generation has entered a phase of rapid, frontier-model iteration. Recent systems such as OpenAI's GPT-Image-2 (OpenAI, 2025), Google's Gemini-3.1-Flash (Google DeepMind, 2026), ByteDance's Seedream 5.0 (ByteDance Seed Team, 2025a;b), Alibaba's Qwen-Image 2.0 Pro (Wu et al., 2025; Qwen Team, 2026), Z-Image (Tongyi-MAI Z-Image Team, 2025), and SenseTime's SenseNova have substantially improved photorealism, instruction following, composition, and in-image text rendering. Many of these models are accessible primarily through APIs and are frequently updated, which makes it difficult for the research community to track their capabilities, limitations, and forensic signatures with public data.

Existing datasets only partially cover this need. Benchmarks for AI-generated image detection, such as GenImage (Zhu et al., 2023) and DRCT-2M (Chen et al., 2024a), are valuable but were largely built around earlier open-source diffusion or GAN models. More broadly, current resources often have three limitations for studying frontier T2I systems: limited generator coverage, limited prompt and content diversity, and few prompt-matched real images. As a result, they are not well suited for controlled cross-model comparison, analysis of prompt following and text rendering, generator attribution, or robust real-versus-generated evaluation.

We introduce GenSpectrum, a prompt-aligned dataset and benchmark for studying modern T2I generators. GenSpectrum contains $9{,}817$ prompts collected from three complementary sources: community API galleries, design-platform prompts, and real-image captions. These prompts include $7{,}818$ generation-oriented prompts and $1{,}999$ derived from real photographs. We render them with up to eight frontier generators, including GPT-Image-2, Gemini-3.1-Flash, Doubao-Seedream-5.0, Z-Image, Qwen-Image-2.0-Pro, and three SenseNova variants, producing $72{,}523$ images in total: $70{,}525$ generated and $1{,}998$ real. All samples are organized under a unified schema with prompt metadata and content-domain annotations, enabling side-by-side comparison of different models under the same textual conditions.

Beyond releasing data, GenSpectrum is designed as a multi-dimensional benchmark. Its aligned structure supports evaluation of visual quality, aesthetics, prompt following, in-image text rendering, realism, AI-generated image detection, generator attribution, and domain-specific failure modes. These dimensions allow us to move from aggregate leaderboards toward more interpretable questions: which models are better at compositional reasoning, which models render text more faithfully, where realism breaks down, and why images from recent generators are harder for existing detectors to identify. In this preview report, we present the dataset construction pipeline and initial benchmark results, with a focus on real-versus-generated detection and cross-generator comparison. Our preliminary findings show that no single detector is robust across all generators: the strongest (DDA) leads by handling the newest models, while detectors that look solved on older generators, such as SAFE, collapse on Gemini-3.1-Flash and Doubao-Seedream-5.0.

Our contributions can be summarized as:

- **A prompt-aligned frontier T2I dataset.** We release $9{,}817$ prompts and $72{,}523$ images from eight modern T2I systems, together with $1{,}998$ matched real images and unified metadata across nine content domains.
- **A multi-dimensional benchmark.** We define evaluation protocols for cross-generator comparison, prompt following, in-image text rendering, realism, visual quality, AI-generated image detection, generator attribution, and domain-wise failure analysis.
- **Initial findings on modern generator forensics.** Using the matched real-image subset, we evaluate 16 detectors and find that none is robust to all generators; the newest models, Gemini-3.1-Flash and Doubao-Seedream-5.0, defeat detectors that look solved on older ones.

- **A preview release with a clear roadmap.** This report releases the first version of GenSpectrum and outlines extensions to more generators, larger scale, image editing pairs, video generation data, and more mechanistic analyses of model strengths and weaknesses.

2 Related Work

2.1 Text-to-Image Generation Models

The models we collect are the latest point in a long line of T2I architectures. Generative adversarial networks dominated early work, from attentional and stacked designs (Xu et al., 2018; Zhang et al., 2019) to stronger single-stage (Tao et al., 2022) and scaled-up (Kang et al., 2023) variants. Autoregressive models then recast generation as next-token prediction over discrete image codes, as in DALL-E (Ramesh et al., 2021), CogView (Ding et al., 2021), and Parti (Yu et al., 2022), with Muse (Chang et al., 2023a) switching to faster masked-token decoding. Diffusion models soon took over. Text-guided pixel-space diffusion (Nichol et al., 2022; Ramesh et al., 2022; Saharia et al., 2022) gave way to latent diffusion (Rombach et al., 2022), which moved the process into a compressed space and became the basis for later high-resolution systems (Podell et al., 2024; Pernias et al., 2024). Replacing the U-Net with a transformer (Peebles & Xie, 2023) and training under rectified-flow objectives produced a wave of efficient diffusion transformers (Chen et al., 2024c;b; Esser et al., 2024; Xie et al., 2025a; Zhuo et al., 2024; Li et al., 2024). The newest shift is toward unified models that both understand and generate images, whether through next-token prediction (Wang et al., 2024; Chen et al., 2025b; Han et al., 2025), joint autoregressive-diffusion training (Xie et al., 2025b; Zhou et al., 2025), or mixed-modal pretraining (Chameleon Team, 2024; Xiao et al., 2025; Deng et al., 2025). The strongest systems today, among them GPT-Image (OpenAI, 2025), Imagen 3 (Imagen-Team-Google, 2024), FLUX (Black Forest Labs, 2024), Qwen-Image (Wu et al., 2025; Qwen Team, 2026), Seedream (ByteDance Seed Team, 2025a;b), Z-Image (Tongyi-MAI Z-Image Team, 2025), and SenseNova, raise quality further still. Many of these are closed-source, reachable only through APIs, and revised often enough that public datasets struggle to keep pace.

2.2 Datasets and Benchmarks for Text-to-Image Generation

Work on these models rests on two kinds of resources: data and ways to score it. Web-scale image-text corpora (Schuhmann et al., 2022; Changpinyo et al., 2021) supply training data, and prompt galleries record how people actually use generators, such as DiffusionDB (Wang et al., 2023b) for Stable Diffusion, JourneyDB (Sun et al., 2023) for Midjourney, and Pick-a-Pic (Kirstain et al., 2023) for user preferences. Evaluation now reaches well beyond Inception Score, FID, and CLIPScore (Salimans et al., 2016; Heusel et al., 2017; Hessel et al., 2021). One group of benchmarks asks how faithfully an image follows its prompt, through compositional and question-answering tests (Ghosh et al., 2023; Huang et al., 2023; Hu et al., 2023; Lin et al., 2024; Cho et al., 2024; Hu et al., 2024; Bakr et al., 2023; Fu et al., 2024); another predicts human preference (Wu et al., 2023; Xu et al., 2023; Zhang et al., 2024; Liang et al., 2024); and holistic suites and live arenas roll many axes into a single score (Lee et al., 2023; Jiang et al., 2024). Image sets aligned across generators already exist, including DRCT-2M's shared captions across sixteen Stable Diffusion variants with matched real images (Chen et al., 2024a) and the multi-model image releases of holistic benchmarks such as HEIM. They remain tied, though, to earlier open models and a narrow content distribution, whether MSCOCO captions, ImageNet classes, or fixed scenario lists, and each is usually aimed at a single purpose such as detection or preference scoring. GenSpectrum targets the gap they leave: it renders the same prompts through the newest commercial and open generators, spreads them across nine deployment-oriented domains, and links one source to the real photographs its prompts describe, under a schema meant to serve several downstream tasks (§4).

2.3 AI-Generated Image Detection

Telling generated images apart from real ones is the first task we build on top of the corpus. Detection research is organized largely around a handful of benchmarks (Wang et al., 2020; Zhu et al., 2023; Wang et al., 2023a; Chen et al., 2024a; Hong & Zhang, 2025; Park & Owens, 2025), many of them assembled from older open GAN and diffusion models. Discriminative detectors hunt for the low-level traces synthesis leaves behind: frequency and spectral artifacts (Durall et al., 2020; Frank et al.,

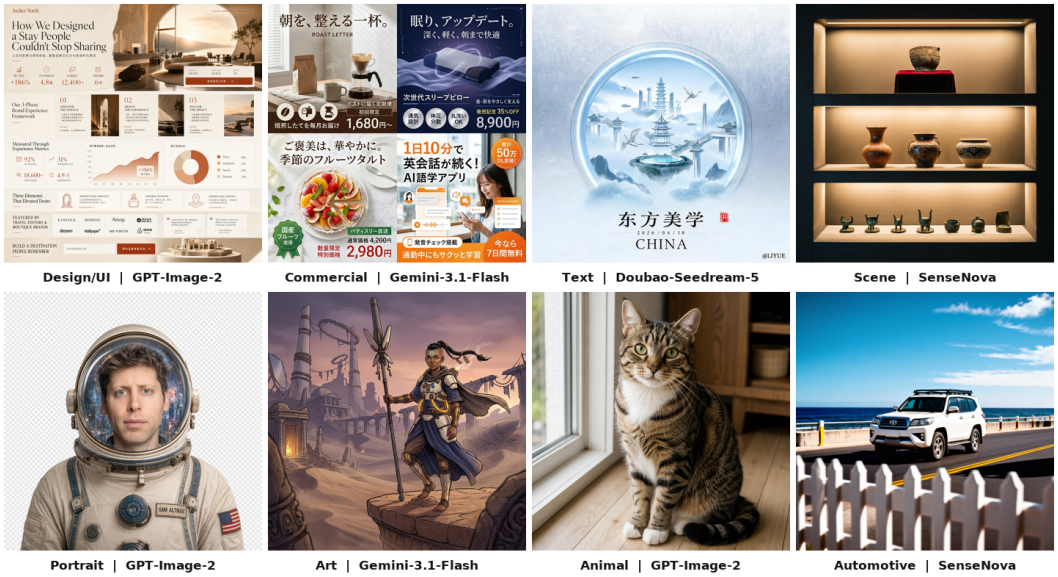

Figure 2: **High-quality samples across the eight main content domains.** One example per domain, labeled with the generator that produced it; GenSpectrum pairs broad domain coverage with strong sample quality. The full taxonomy is given in Appendix C.

2020; Qian et al., 2020; Tan et al., 2024b), texture and gradient statistics (Liu et al., 2020; Tan et al., 2023), learned noise or up-sampling fingerprints (Liu et al., 2022; Tan et al., 2024a), features taken from a frozen vision-language backbone (Ojha et al., 2023; Liu et al., 2024; Yan et al., 2025b; Tan et al., 2025a), augmentation-invariant training (Li et al., 2025a; Yan et al., 2025a), and reconstruction or alignment signals (Ricker et al., 2024; Chen et al., 2025a; Liu et al., 2026; Sundara Rajan et al., 2025). A newer line asks multimodal LLMs to flag an image and explain its artifacts at once (Wen et al., 2025; Jiang et al., 2025; Tan et al., 2025b; Xu et al., 2025; Huang et al., 2025; Li et al., 2025b; Chang et al., 2023b), with benchmarks built to test them in that role (Ye et al., 2025). One weakness recurs across these families: detectors tuned on older generators often transfer poorly to newer ones. Our matched real-and-generated subset is meant to probe exactly this failure, pairing genuine photographs with images from generators that today's detectors have never encountered (§4.1).

3 The GenSpectrum Corpus and Construction Workflow

GenSpectrum is organized around one principle: *the prompt is the unit of alignment*. Each entry stores a prompt together with every image rendered from it, so the same content can be compared across generators and, where available, against a real photograph. Concretely, every row follows a single 15-field schema: an identifier and `source`; a two-level domain label (`category`, `sub_category`); the `prompt` and its `original_prompt`; an optional `real_image`; and eight `generated_*` fields, one per generator, with all images stored inline as base64. The released corpus is 9,817 prompts and 72,523 images (∼279 GB) sharded into 42 Parquet files. Figure 2 shows representative high-quality samples across domains.

3.1 Prompt Sources

Three complementary pipelines feed the corpus; their distinct origins are what give GenSpectrum both breadth and the real/fake pairing.

`crawl` (3,818 prompts). Seed prompts are drawn from a public gallery of GPT-Image API cases[1] and from prompts scraped on the Xiaohongshu (RED) platform, covering seven practical case types

[1] https://github.com/EvoLinkAI/awesome-gpt-image-2-API-and-Prompts

Table 1: Composition of GenSpectrum. Each prompt is rendered by up to eight generators (seven evaluated); `pexels` additionally provides a matched real image. "Images" counts all generated images for that source.

Source	Prompt origin	#Prompts	Real img.	Role
`crawl`	API gallery + Xiaohongshu (LLM-rewritten)	3,818	–	design / commercial breadth
`dreamina`	Dreamina community prompts (sampled)	4,000	–	typography / scene / portrait
`pexels`	Real photos → VLM captions	1,999	1,998	matched real/fake pairs
Total	three complementary pipelines	**9,817**	**1,998**	72,523 images, 9 domains

(poster, portrait, UI, comparison, character, e-commerce, ad-creative). Each seed case is rewritten by an LLM into several prompt variants (§3.2), expanding coverage while preserving intent.

`dreamina` (4,000 prompts). A large pool of community generation prompts from the Dreamina platform is classified into the domain taxonomy and sampled for balance, contributing strongly to the typography, scene, portrait, animal, and automotive domains.

`pexels` (1,999 prompts / 1,998 real images). Starting from real photographs on Pexels, a vision–language model produces a descriptive caption that serves as the prompt; feeding that prompt back to a generator yields an image aligned to a genuine reference photo. This source is the origin of all real images in GenSpectrum and enables real-versus-generated study, including the class-balanced detection protocol of §4.1 (Figure 4).

3.2 Prompt rewriting and domain taxonomy

For the `crawl` source, each seed prompt is expanded by an LLM into five variants under a per-type template, broadening stylistic and lexical diversity without departing from the original creative goal. Every prompt is then assigned a label from a nine-class domain taxonomy (Design/UI, Commercial/Ad, Text/Typography, Scene/Landscape, Person/Portrait, Art/Illustration, Animal, Automotive, Other) via an LLM classifier driven by a deterministic decision procedure over the *depicted image* rather than the prompt's surface form; the full label set is given in Appendix C.

3.3 Multi-model generation

Prompts are dispatched to eight generators through a single OpenAI-compatible gateway: GPT-Image-2, Gemini-3.1-Flash, Doubao-Seedream-5.0, Z-Image, Qwen-Image-2.0-Pro, and three SenseNova variants (base, MoT-SFT, and MoT-Infographic; abbreviated SN-MoT-* in figures). Each model keeps its native resolution and output format (e.g. 1024^2 for GPT-Image-2/Gemini/Z-Image, 1920^2 for Doubao-Seedream, 2048^2 for SenseNova); per-model settings are listed in Appendix A. Generation is text-only for all models, so even prompts that reference a source image are rendered purely from text. Coverage is deliberately uneven, mirroring real API access: GPT-Image-2, Gemini-3.1-Flash, Z-Image, and the SenseNova variants are rendered for nearly all prompts, whereas Qwen-Image-2.0-Pro covers only a small subset (3,100 images, 381 on `pexels`) and is kept in the corpus but excluded from the detection evaluation in §4.1 (Figure 3b).

3.4 Dataset statistics

Table 1 summarizes the corpus and Figure 3 shows its composition. The three sources are deliberately skewed toward different domains—`crawl` toward Design/UI and Commercial, `dreamina` toward Typography and Scenes, and `pexels` spread broadly across photographic domains—so that their union spans nine domains rather than a single distribution.

4 Downstream Tasks

Because GenSpectrum aligns every prompt across generators, and for the `pexels` source against a real photograph, it is a task-agnostic testbed rather than a single-purpose benchmark. The same corpus supports, among others: *cross-model comparison* of fidelity, aesthetics, and style; *prompt*

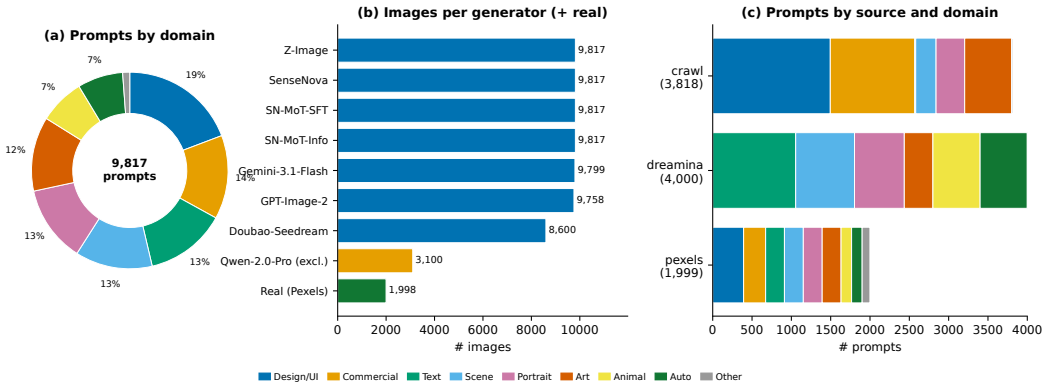

Figure 3: **Composition of GenSpectrum.** (a) Prompt distribution over the nine content domains. (b) Generated images per model plus matched real images; coverage is uneven by design, mirroring real API access, with Qwen-Image-2.0-Pro excluded from evaluation. (c) Prompts per source, broken down by domain, showing the three sources' complementary coverage.

following and instruction adherence; *in-image text rendering* (Appendix B); *generator attribution / fingerprinting*; *AI-generated image (AIGI) detection and forensics*; and robustness studies for watermarking and editing. This first report instantiates one of these, AIGI detection, as a case study; subsequent versions will add the others as the corpus and its annotations grow.

4.1 Case study: AI-generated image detection

We choose AIGI detection as the first task because it exercises exactly what GenSpectrum adds over prior resources: the newest generators, nine mixed content domains, and matched real images for a fair real-versus-generated test.

Setup. An image is *real* (class 0) if it comes from the `pexels real_image` field and *fake* (class 1) otherwise. To avoid the source-distribution shortcut that a mixed split invites, every metric is computed on *prompt-matched* `pexels` pairs: the 1,998 real photographs against each generator's `pexels` fakes, one generator at a time. We report per-generator *balanced accuracy*, $\mathrm{BAcc} = \frac{1}{2}(\text{real-recall}+\text{fake-recall})$, and its mean over the seven evaluated generators; Qwen-Image-2.0-Pro is excluded for incomplete coverage. Because the full corpus is 96% fake, full-set accuracy is inflated, so we do not report it. We evaluate 16 detectors off-the-shelf (released checkpoints, native preprocessing, no fine-tuning on GenSpectrum): eight *discriminative* detectors (SAFE (Li et al., 2025a), DDA (Chen et al., 2025a), Effort (Yan et al., 2025b), MIRROR (Liu et al., 2026), DRCT (Chen et al., 2024a), FatFormer (Liu et al., 2024), AlignedForensics (Sundara Rajan et al., 2025), and UnivFD (Ojha et al., 2023)) and eight *generative* detectors: forensic multimodal LLMs (FakeVLM (Wen et al., 2025), Ivy-Fake (Jiang et al., 2025), Veritas (Tan et al., 2025b), and OmniVL-Guard at 2B and 8B) and general-purpose VLMs (Qwen3.5-9B, Qwen3-VL-8B, InternVL3.5-8B). Discriminative scores use the fake-class probability at a 0.5 threshold; generative answers are parsed from the model's text, and a parse failure counts as a miss.

Main results. Table 2 reports per-generator BAcc and Figure 5 ranks the detectors by their mean. The headline is that **no detector is robust across all seven generators**. DDA is the most consistent (mean 0.871, and still 0.71 on its hardest generator), holding up where the others break. SAFE shows the opposite pattern: it is the best detector on the older five generators (BAcc 0.96 to 0.98), yet its mean falls to 0.837 because the two newest generators push it to chance ($\mathrm{BAcc} \approx 0.50$) even as its real-recall stays at 0.98. A detector that looks solved on last year's generators, then, is not solved on this year's.

Which generators are hardest? Figure 6 shows per-generator fake-recall, the detection rate on each generator's images. Two recent models stand out, and they break *different* detectors. Gemini-3.1-Flash slips past the frequency- and reconstruction-style detectors: SAFE, FatFormer, and Uni-

Table 2: Per-generator balanced accuracy (BAcc) on prompt-matched `pexels` pairs, for 16 detectors over the seven evaluated generators (*disc* = discriminative, *gen* = generative), sorted by mean. The *real* column is the shared real-recall. Even the most consistent detector, DDA, drops to 0.71 on its hardest generator.

Detector	Type	real	GPT	Z	SN	SN-SFT	SN-Info	Gemini	Doubao	Mean ↑
DDA (Chen et al., 2025a)	disc	0.93	0.794	0.708	0.911	0.931	0.922	0.939	0.890	**0.871**
SAFE (Li et al., 2025a)	disc	0.98	0.967	0.983	0.975	0.961	0.974	0.500	0.496	0.837
OmniVL-Guard-2B	gen	0.80	0.812	0.748	0.892	0.878	0.883	0.770	0.865	0.835
Effort (Yan et al., 2025b)	disc	0.71	0.853	0.854	0.826	0.848	0.840	0.768	0.497	0.784
MIRROR (Liu et al., 2026)	disc	0.94	0.672	0.638	0.965	0.941	0.951	0.605	0.612	0.769
Veritas (Tan et al., 2025b)	gen	0.60	0.715	0.730	0.757	0.726	0.740	0.708	0.753	0.733
DRCT (Chen et al., 2024a)	disc	0.93	0.497	0.666	0.736	0.743	0.731	0.874	0.760	0.715
FakeVLM (Wen et al., 2025)	gen	0.48	0.724	0.730	0.740	0.733	0.740	0.724	0.562	0.708
AlignedForensics (Sundara Rajan et al., 2025)	disc	0.98	0.557	0.495	0.757	0.718	0.680	0.943	0.634	0.683
FatFormer (Liu et al., 2024)	disc	1.00	0.624	0.561	0.794	0.879	0.865	0.499	0.500	0.675
Qwen3-VL-8B	gen	0.93	0.572	0.548	0.740	0.670	0.669	0.540	0.797	0.648
Ivy-Fake (Jiang et al., 2025)	gen	0.15	0.551	0.564	0.570	0.571	0.568	0.560	0.565	0.564
InternVL3.5-8B	gen	0.94	0.528	0.509	0.548	0.553	0.545	0.517	0.629	0.547
OmniVL-Guard-8B	gen	0.17	0.496	0.508	0.538	0.547	0.536	0.503	0.544	0.525
Qwen3.5-9B	gen	0.14	0.513	0.494	0.553	0.527	0.539	0.504	0.536	0.524
UnivFD (Ojha et al., 2023)	disc	0.61	0.331	0.377	0.482	0.570	0.510	0.320	0.402	0.427

vFD each flag under 4% of its images, yet DDA (0.95), AlignedForensics (0.90), and DRCT (0.82) catch it readily. Doubao-Seedream-5.0 also defeats SAFE and FatFormer (under 1% recall) and weakens Effort, but DDA (0.85) and the forensic MLLM OmniVL-Guard-2B (0.93) remain strong. Almost no detector collapses on both, so the two generators expose complementary blind spots rather than one shared weakness. The SenseNova variants, in contrast, are caught reliably across the board.

Generative and forensic LLMs. The eight generative detectors span a wide bias range. OmniVL-Guard-2B is the strongest of them (mean BAcc 0.835, third overall) and stays balanced between real and fake, beating its own 8B sibling, which is heavily fake-biased (real-recall 0.17). Two failure modes recur. General-purpose VLMs (Qwen3-VL-8B, InternVL3.5-8B) default to calling images real, keeping high real-recall but missing most modern fakes; several others (Ivy-Fake, Qwen3.5-9B, OmniVL-Guard-8B) do the reverse, with strong fake-recall but labeling most real photographs fake (real-recall 0.14 to 0.17). Only OmniVL-Guard-2B, Veritas, and FakeVLM are reasonably balanced. Per-generator fake-recall for all 16 detectors is given in Appendix D.

A confound to control for. Real photographs are large, non-square camera images (5 to 30 MP), whereas the generators emit fixed square images (1024^2 for GPT-Image-2 and Gemini, 1920^2 for Doubao-Seedream, 2048^2 for SenseNova), so a detector could in principle key on resolution or compression rather than on generative artifacts. We partly control for this by downsampling every image to a 512-px long side before the generative VLMs, and by keeping each discriminative detector's native preprocessing (center-crop 224 or resize 256). A fully format-normalized protocol that re-encodes real and generated images to a common resolution and codec, following Chen et al. (2025a), is a planned addition. The numbers here should therefore be read as an off-the-shelf *cross-generator* benchmark rather than a fully controlled forensic study.

5 Conclusion and Roadmap

We introduced GenSpectrum, a unified, prompt-aligned corpus of 9,817 prompts rendered by up to eight state-of-the-art generators (72,523 images across nine domains), with matched real photographs for a subset, together with a reproducible construction workflow. By aligning many generators under one schema, GenSpectrum is a general testbed for studying modern T2I systems rather than a single-task benchmark. As a first case study we benchmarked 16 AI-generated image detectors, already surfacing findings that single-generator datasets cannot: no detector is robust across all generators, and the two newest, Gemini-3.1-Flash and Doubao-Seedream-5.0, defeat detectors that look solved on older ones while breaking different detector families.

This is a preliminary report, and the resource is designed to grow, in both generators and, especially, *tasks*. Planned downstream tasks include generator attribution and fingerprinting, prompt-following and aesthetic / human-preference evaluation, in-image text-rendering analysis, and edit-

ing/watermarking robustness; on the detection side we will complete coverage of Qwen-Image-2.0-Pro and add a format-normalized protocol that removes the resolution and codec confound. Alongside these, we will continuously fold in new models as they are released. By aligning many generators under one schema, GenSpectrum aims to be a moving-target resource that keeps pace with T2I progress and serves many downstream studies, of which detection is only the first.

Data availability. GenSpectrum (prompts, generated images, matched real images, and the unified Parquet schema) and the construction scripts will be released to support reproducible research; a versioned snapshot accompanies this report.

Ethics and licensing. Real photographs originate from Pexels and are used under its license; community prompts and gallery cases are aggregated from public sources with provenance recorded per entry. The corpus is intended for research on generative-model analysis, detection, attribution, and model comparison; it is not intended for training systems to evade detection or to impersonate individuals. Generated content may reflect biases of the underlying models.

Use of AI tools. Large language models were used as components of the pipeline (prompt rewriting and taxonomy classification) and as assistive tools during manuscript preparation; all quantitative results and claims were verified by the authors.

6 ROADMAP TO FUTURE

Model coverage The first direction is to keep GenSpectrum synchronized with the frontier models that are actually used by researchers, creators, and online platforms. We will continue adding both proprietary API systems and strong open-weight models, including future releases from the GPT-Image, Gemini, Seedream, Qwen-Image, Z-Image, SenseNova, and Nano Banana families. For each model update, we will preserve versioned metadata such as generation date, API version when available, decoding settings, safety-filter behavior, output resolution, and failure cases. This is important because modern generators are not static artifacts: API models may change silently, and benchmark conclusions can quickly become outdated if the underlying generator version is not tracked. A versioned release structure will allow the community to compare not only different model families, but also different generations of the same model over time.

Scale and domain expansion The second direction is to enlarge the prompt and image coverage while keeping the dataset balanced and interpretable. We will increase the number of prompts, add more real-image matched samples, and refine the domain taxonomy beyond the current nine categories. Future versions will place more emphasis on high-value deployment scenarios, including product photography, advertisements, posters, UI and web mockups, infographics, documents, social-media images, stylized illustrations, human portraits, indoor and outdoor scenes, and safety-sensitive or misinformation-prone content. Rather than simply scaling the dataset uniformly, we will use targeted sampling to cover known hard cases such as dense text, fine-grained object attributes, spatial relations, hands and faces, reflections, small objects, and multi-object interactions.

Beyond text-to-image generation The third direction is to extend GenSpectrum from static T2I generation to broader visual generation tasks. We plan to add image-editing pairs, multi-turn editing trajectories, and video-generation samples. For image editing, the dataset will support evaluation of instruction fidelity, identity and subject preservation, background consistency, local edit precision, and unintended changes outside the target region. For multi-turn editing, we will study whether models can maintain visual state across a sequence of user instructions. For video generation, we will introduce prompts and reference frames that enable evaluation of temporal consistency, motion realism, object permanence, text consistency across frames, and the accumulation of artifacts over time. These extensions will make the benchmark useful not only for image synthesis, but also for the broader ecosystem of controllable visual generation.

Multi-dimensional evaluation The fourth direction is to broaden the benchmark from real-versus-generated detection to a suite of complementary evaluation tasks. We will include automatic and human-aligned protocols for visual quality, aesthetics, realism, prompt following, object and attribute binding, spatial and counting reasoning, compositionality, in-image text rendering, style

controllability, generator attribution, and domain-wise failure analysis. These tasks are designed to provide different views of model behavior: some measure whether an image looks good, some measure whether it follows the prompt, some measure whether it resembles a real photograph, and some measure whether its artifacts reveal the source generator. Combining these dimensions will help avoid a single-score view of image generation quality and make the benchmark more useful for diagnosing model strengths and weaknesses.

Mechanistic and interpretable analysis The fifth direction is to move from aggregate model rankings toward explanations of why a model performs better or worse. For example, when GPT-Image-2 or another frontier model outperforms alternatives, we want to identify whether the advantage comes from better semantic grounding, stronger layout planning, more reliable text rendering, fewer local artifacts, better face and hand synthesis, improved style control, or more realistic lighting and material rendering. We will therefore report results by prompt type, visual domain, object category, reasoning requirement, and artifact class. We also plan to connect detector failures with generator properties, asking which visual cues detectors rely on and why these cues disappear or become unreliable for newer models. This analysis will make GenSpectrum useful not only as a leaderboard, but also as a diagnostic tool for improving generation and forensic methods.

Data quality for T2I training The sixth direction is to study how high-quality frontier-model data affects downstream T2I training, adaptation, and evaluation. Because GenSpectrum contains prompt-aligned outputs from multiple strong generators, it can support controlled studies of data selection, prompt rewriting, synthetic-real mixing, domain balancing, preference filtering, and model-specific data transfer. We will investigate whether images from stronger generators provide better supervision, whether mixing real and synthetic images improves robustness, whether certain prompt domains are more valuable for training, and whether synthetic data can transfer improvements to models with different architectures or training recipes. These experiments will help clarify the role of curated frontier-model data as a training resource rather than only as an evaluation set.

Release and maintenance The final direction is to make GenSpectrum a living but reproducible community resource. Future releases will include standardized metadata, evaluation scripts, data cards, model-version records, and benchmark reports for each major update. We will separate stable benchmark splits from expandable data pools, so that researchers can compare methods on fixed versions while still benefiting from newly added models and tasks. We also plan to release additional annotations, including prompt taxonomy labels, generation failure tags, detector predictions, and human or model-based preference judgments when available. This release strategy is intended to make GenSpectrum useful both as a static reference dataset and as a continuously updated measurement platform for the evolution of visual generation.

REFERENCES

Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. HRS-Bench: Holistic, reliable and scalable benchmark for text-to-image models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2304.05390.

Black Forest Labs. FLUX.1. https://blackforestlabs.ai, 2024. Accessed 2026.

ByteDance Seed Team. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025a.

ByteDance Seed Team. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025b.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning (ICML)*, 2023a. arXiv:2301.00704.

You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. AntifakePrompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023b.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. arXiv:2102.08981.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *International Conference on Machine Learning (ICML)*, 2024a.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-Σ: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. In *European Conference on Computer Vision (ECCV)*, 2024b. arXiv:2403.04692.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024c. arXiv:2310.00426.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes AI-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025a. DDA.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.18235.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2105.13290.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. arXiv:2003.01826.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2403.03206.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning (ICML)*, 2020. arXiv:2003.08685.

Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-T2I challenge: Can text-to-image generation models understand commonsense? In *Conference on Language Modeling (COLM)*, 2024. arXiv:2406.07546.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Google DeepMind. Gemini flash image generation. https://ai.google.dev/gemini-api/docs/image-generation, 2026. Gemini 3.1 Flash Image (Preview); accessed 2026.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.04431.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. arXiv:2104.08718.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1706.08500.

Yan Hong and Jianfu Zhang. WildFake: A large-scale and hierarchical dataset for AI-generated images detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025. arXiv:2402.11843.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2303.11897.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.04292.

Imagen-Team-Google. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Chenyang Si, et al. Ivy-Fake: A unified explainable framework and benchmark for image and video AIGC detection. *arXiv preprint arXiv:2506.00979*, 2025.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. GenAI arena: An open evaluation platform for generative models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024. arXiv:2406.04485.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2303.05511.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.01569.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023. arXiv:2311.04287.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025a. SAFE; arXiv:2408.06741.

Yixuan Li, Xuelin Liu, Xiaoyang Wang, Bu Sung Lee, Shiqi Wang, Anderson Rocha, and Weisi Lin. FakeBench: Probing explainable fake image detection via large multimodal models. *IEEE Transactions on Information Forensics and Security (TIFS)*, 2025b. arXiv:2404.13306.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2312.10240.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2404.01291.

Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision (ECCV)*, 2022.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ruiqi Liu, Manni Cui, Ziheng Qin, Zhiyuan Yan, Ruoxin Chen, Yi Han, Zhiheng Li, Junkai Chen, ZhiJin Chen, Kaiqing Lin, Jialiang Shen, Lubin Weng, Jing Dong, Yan Wang, and Shu Wu. MIRROR: Manifold ideal reference ReconstructOR for generalizable AI-generated image detection. *arXiv preprint arXiv:2602.02222*, 2026.

Zhengzhe Liu, Xiaojuan Qi, and Philip H. S. Torr. Global texture enhancement for fake face detection in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2112.10741.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2302.10174.

OpenAI. Introducing our latest image generation model in the API (GPT-Image). https://openai.com/index/image-generation-api/, 2025. Accessed 2026.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2411.04125.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2212.09748.

Pablo Pernias, Dominic Rampas, Mats L. Richter, Christopher J. Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.00637.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2307.01952.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision (ECCV)*, 2020. arXiv:2007.09355.

Qwen Team. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2102.12092.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2401.17879.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2112.10752.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2205.11487.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. arXiv:1606.03498.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022. arXiv:2210.08402.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. JourneyDB: A benchmark for generative image understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2307.00716.

Anirudh Sundara Rajan et al. Aligned datasets improve detection of latent diffusion-generated images. In *International Conference on Learning Representations (ICLR)*, 2025. AlignedForensics; arXiv:2410.11835.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for GAN-generated images detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a. arXiv:2312.10461.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024b. arXiv:2403.07240.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2P-CLIP: Injecting category common prompt in CLIP to enhance generalization in deepfake detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025a. arXiv:2408.09647.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. *arXiv preprint arXiv:2508.21048*, 2025b.

Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. DF-GAN: A simple and effective baseline for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2008.05865.

Tongyi-MAI Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. arXiv:1912.11035.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023a. arXiv:2303.09295.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023b. arXiv:2210.14896.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. FakeVLM; arXiv:2503.14905.

Chenfei Wu et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2409.11340.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. In *International Conference on Learning Representations (ICLR)*, 2025a. arXiv:2410.10629.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2025b. arXiv:2408.12528.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2304.05977.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. arXiv:1711.10485.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.02761.

Shilin Yan et al. A sanity check for AI-generated image detection. In *International Conference on Learning Representations (ICLR)*, 2025a. AIDE; arXiv:2406.19435.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, et al. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *International Conference on Machine Learning (ICML)*, 2025b. Effort; arXiv:2411.15633.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, Zhizheng Wu, Yiping Chen, Dahua Lin, Conghui He, and Weijia Li. LOKI: A comprehensive synthetic data detection benchmark using large multimodal models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.09732.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research (TMLR)*, 2022. arXiv:2206.10789.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. arXiv:1710.10916.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2405.14705.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.11039.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-Next: Making Lumina-T2X stronger and faster with Next-DiT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.18583.

A Generation Model Settings

Table 3 lists the eight generators and their configuration as used in GenSpectrum; seven are used in the detection evaluation, and Qwen-Image-2.0-Pro is incomplete and excluded. All models are accessed through a single OpenAI-compatible gateway; each keeps its native output resolution and format, and images are normalized to inline base64 in the unified schema. Prompt rewriting and the nine-class taxonomy classification are performed by a GPT-5.4 chat model accessed through an OpenAI-compatible gateway; the `pexels` captions are produced by a vision–language model (specified in the released code). Each `crawl` seed prompt is expanded into five variants. Generation is text-only for all models. Per-model coverage (Figure 3b) is governed by API access and rate limits rather than by a sampling design; Qwen-Image-2.0-Pro, in particular, covers only a small subset and is excluded from the detection evaluation in §4.1.

Table 3: Generators in GenSpectrum and their native settings.

Model	Provider	Resolution	Format	Access
GPT-Image-2	OpenAI	1024×1024	PNG (b64)	API
Gemini-3.1-Flash	Google	1024×1024	PNG (b64)	API
Doubao-Seedream-5.0	ByteDance	1920×1920	JPEG (url)	API
Z-Image	Alibaba Tongyi	1024^2/2048^2	PNG (url)	API
Qwen-Image-2.0-Pro	Alibaba	1024×1024	PNG (url)	API (limited; excl.)
SenseNova	SenseTime	2048×2048	PNG (url)	API
SenseNova MoT-SFT	SenseTime	2048×2048	PNG (url)	API
SenseNova MoT-Info	SenseTime	2048×2048	PNG (url)	API
Real (Pexels)	—	native	JPEG	photograph

B Sample Gallery

Figure 7 shows one representative GenSpectrum prompt per domain, rendered across five generators, illustrating the corpus's domain breadth and the stylistic spread between models. Figure 8 highlights a text-heavy advertising prompt, a setting where in-image text rendering differs sharply across generators.

C Domain Taxonomy

Every prompt is assigned a two-level domain label by an LLM classifier that reasons about the *depicted image* (not the prompt's surface form) following a fixed decision procedure. The nine top-level domains and their sub-categories are listed below, with the corpus-wide count of prompts per domain.

- **Design/UI** (1,888): product design drawing, interior design, furniture design, logo design, graphic design, poster layout, UI design, brand-identity design, concept design, packaging design, clothing design, tattoo design, jewelry design, infographic, before/after collage, other design.
- **Commercial/Ad** (1,353): product photography, food photography, fashion advertising, product advertising, automotive advertising, movie poster, e-commerce hero image, promotional banner, advertising poster, other advertising.
- **Text/Typography** (1,307): title design, typeface design, calligraphy art, handwriting design, type poster, other typography.
- **Scene/Landscape** (1,247): natural landscape, nature photography, architectural photography, indoor scene, indoor photography, cityscape, city night, cinematic scene, street scene, still-life scene, historical scene, sci-fi scene, wedding scene, festival/celebration, other scene.
- **Person/Portrait** (1,237): portrait, fashion photography, fashion portrait, black-and-white portrait, female portrait, still-life portrait, people photography, close-up portrait, traditional-costume figure, child portrait, group portrait, sports figure, other person.
- **Art/Illustration** (1,202): digital illustration, digital painting, traditional ink painting, traditional Chinese painting, fantasy illustration, concept art, anime-style illustration, 3D illustration, 3D animated character, pixel art, abstract art, children's illustration, character illustration, comic, other illustration.
- **Animal** (735): cat, dog, bird, insect, wildlife photography, wildlife, animal portrait, animal illustration, aquatic animal, other animal.
- **Automotive** (734): sports car, SUV, sedan, concept car, vintage car, other vehicle.
- **Other** (114): parse failure, cross-category composite, unclassifiable.

D Detailed Detection Results

Detector configurations. Each discriminative detector runs with its released backbone and native preprocessing (Table 4); scores are the model's fake-class probability at a 0.5 threshold. The eight generative detectors are served through vLLM with a uniform 512-px long-side downsampling: the forensic MLLMs are FakeVLM (LLaVA-1.5), Ivy-Fake (Qwen2.5-VL), Veritas (InternVL-Chat), and OmniVL-Guard-2B/8B (Qwen3-VL), and the general-purpose VLMs are Qwen3.5-9B, Qwen3-VL-8B-Instruct, and InternVL3.5-8B. For chain-of-thought models the answer is parsed from an `<answer>` tag; an unparseable or truncated answer counts as a miss.

Table 4: Discriminative detectors: backbone / training variant and input preprocessing.

Detector	Backbone / training	Preprocessing
SAFE	ResNet-50	CenterCrop 256
DDA	DINOv2 + LoRA (MSCOCO + aligned)	CenterCrop 336 + CLIP norm
Effort	CLIP ViT-L/14 + SVD (SDv1.4)	CenterCrop 224 + CLIP norm
MIRROR	DINOv3-H + memory bank	Resize 512 $\rightarrow$ CenterCrop 224
DRCT	ConvNeXt-B (GenImage)	Pad + CenterCrop 224 + ImageNet norm
FatFormer	CLIP (ProGAN)	224 + ImageNet norm
AlignedForensics	aligned real/fake (sync)	variable size (no resize) + ImageNet norm
UnivFD	CLIP ViT-L/14 + linear probe	CenterCrop 224 + CLIP norm

Per-generator fake-recall. Table 5 reports each detector's fake-recall (detection rate) on every generator's `pexels` images, the raw signal behind the balanced accuracies in Table 2 and the heatmap in Figure 6. The near-zero entries for Gemini-3.1-Flash and Doubao-Seedream-5.0 under SAFE, FatFormer, and UnivFD are the clearest sign of the cross-generator gap; conversely, the fake-biased generative detectors (Ivy-Fake, Qwen3.5-9B, OmniVL-Guard-8B) reach high fake-recall everywhere but pay for it with low real-recall.

Table 5: Per-generator fake-recall on prompt-matched `pexels` images, for all 16 detectors (*disc* = discriminative, *gen* = generative). *real* = real-recall (shared). Sample sizes: real 1,998; GPT 1,986; Z / SenseNova ($\times$3) 1,999; Gemini 1,998; Doubao 1,990.

Detector	Type	real	GPT	Z	SN	SN-SFT	SN-Info	Gemini	Doubao
SAFE	disc	0.984	0.950	0.981	0.966	0.938	0.964	0.015	0.008
DDA	disc	0.932	0.657	0.485	0.891	0.930	0.913	0.947	0.849
Effort	disc	0.711	0.994	0.997	0.941	0.985	0.969	0.824	0.282
MIRROR	disc	0.943	0.400	0.333	0.987	0.939	0.960	0.267	0.280
DRCT	disc	0.930	0.063	0.402	0.542	0.555	0.531	0.818	0.589
FatFormer	disc	0.996	0.251	0.126	0.591	0.762	0.733	0.003	0.003
AlignedForensics	disc	0.984	0.130	0.006	0.529	0.451	0.375	0.901	0.283
UnivFD	disc	0.610	0.053	0.144	0.355	0.531	0.411	0.031	0.193
FakeVLM	gen	0.482	0.966	0.978	0.999	0.984	0.999	0.966	0.643
Ivy-Fake	gen	0.152	0.950	0.975	0.987	0.990	0.983	0.968	0.977
Veritas	gen	0.601	0.830	0.859	0.914	0.850	0.879	0.816	0.905
OmniVL-Guard-2B	gen	0.805	0.820	0.691	0.978	0.951	0.961	0.734	0.926
OmniVL-Guard-8B	gen	0.165	0.827	0.850	0.911	0.929	0.907	0.840	0.923
Qwen3.5-9B	gen	0.139	0.887	0.848	0.966	0.915	0.938	0.869	0.933
Qwen3-VL-8B	gen	0.926	0.217	0.170	0.554	0.414	0.412	0.153	0.668
InternVL3.5-8B	gen	0.936	0.120	0.082	0.159	0.169	0.153	0.097	0.322

Generative VLM notes. General-purpose VLMs default to predicting *real*, which keeps real-recall high (Qwen3-VL-8B 0.93, InternVL3.5-8B 0.94) but leaves fake-recall low on the newest generators. The forensic MLLM OmniVL-Guard-2B is the most balanced generative detector and the third-best detector overall (Table 2). FakeVLM is evaluated through its native transformers path; its vLLM path degenerates to an empty `[IMAGE]` response and is not used.

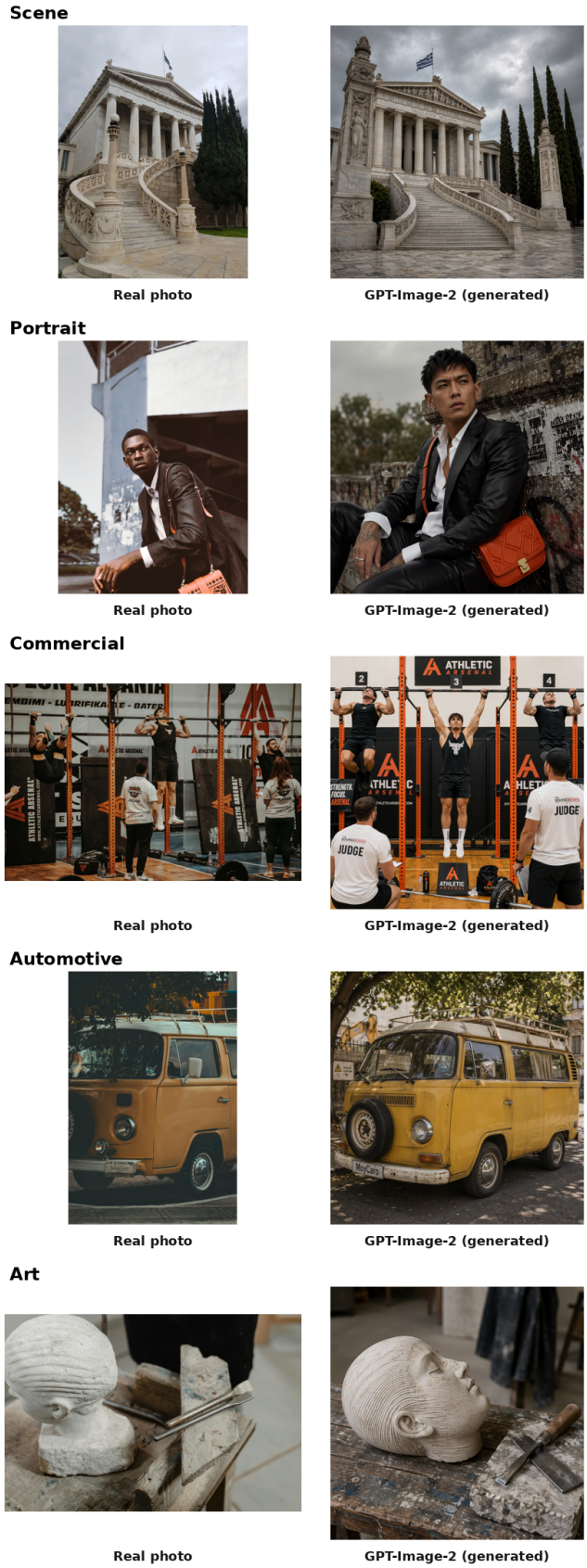

Figure 4: **Matched real/fake pairs from the `pexels` source.** Each row shows a real Pexels photograph (left) and the GPT-Image-2 image generated from its caption (right), across five photographic domains. These aligned pairs enable the class-balanced detection protocol in §4.1.

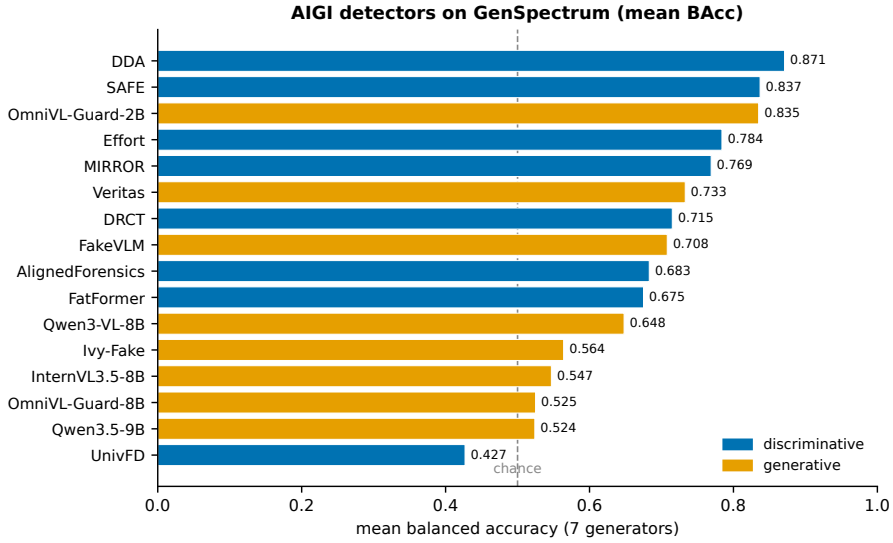

Figure 5: **Mean balanced accuracy over the seven evaluated generators**, for all 16 detectors (discriminative vs. generative). Even the best detector stays well below ceiling and the spread is wide; the dashed line marks chance.

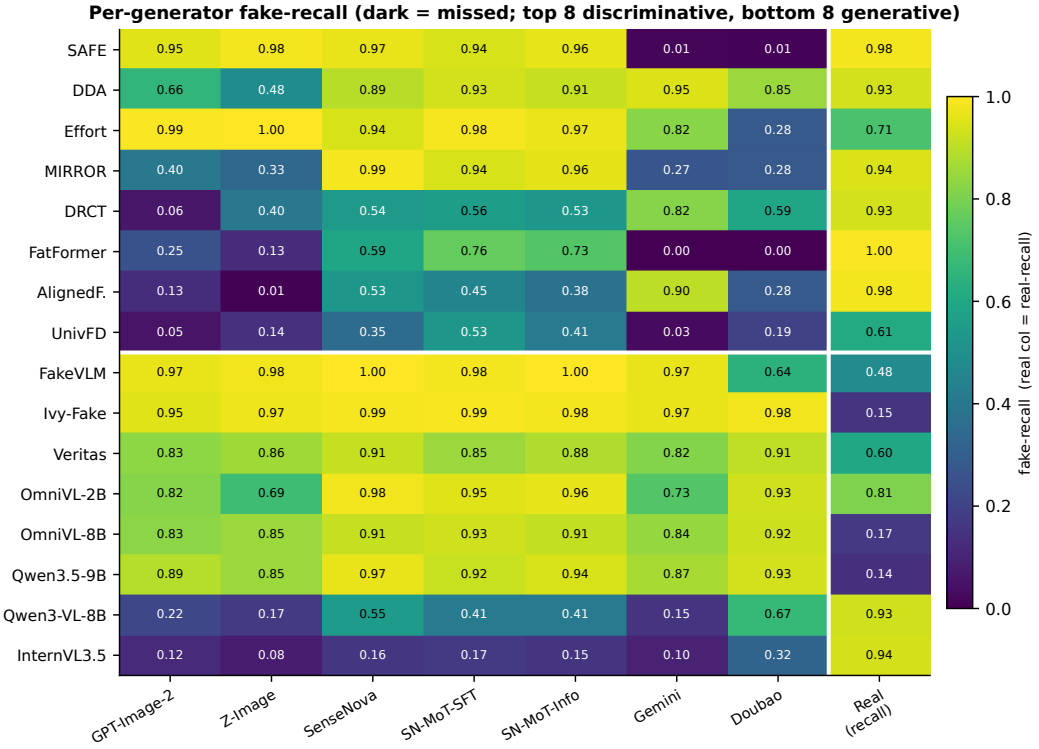

Figure 6: **Per-generator fake-recall** (detection rate) for all 16 detectors. Dark = missed. Gemini-3.1-Flash and Doubao-Seedream-5.0 form two dark columns that hit *different* detectors, exposing complementary blind spots.

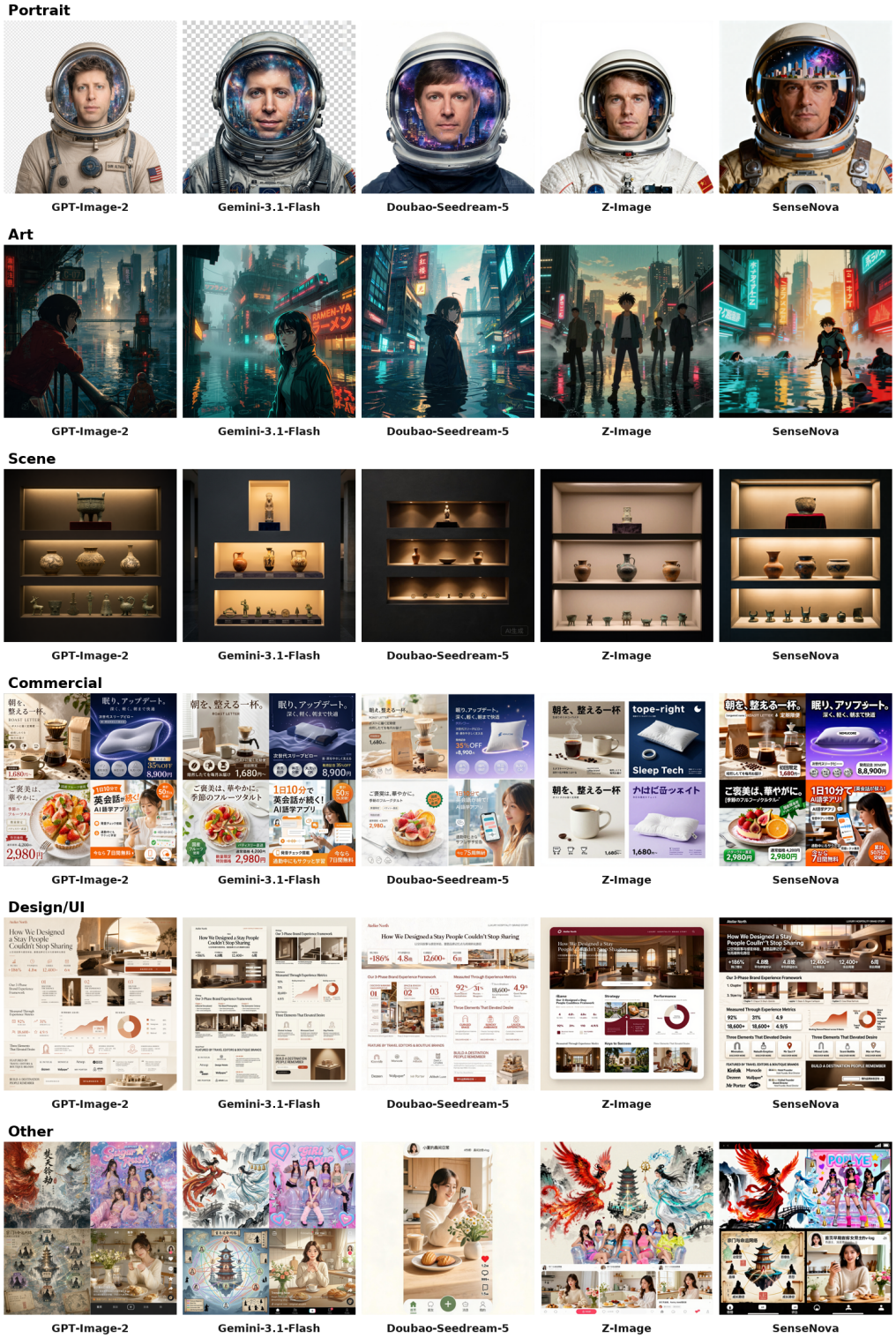

Figure 7: **Per-domain gallery.** One prompt per domain (rows) rendered by five generators (columns). Domains shown: Portrait, Art, Scene, Commercial, Design/UI, and Other.



Figure 8: **In-image text rendering.** A single advertising-banner prompt rendered by seven of the generators; models differ markedly in typography fidelity and layout adherence.