MIRROR: Manifold Ideal Reference ReconstructOR for AI-Generated Image Detection

Ruiqi Liu[1,2,*] **Manni Cui**[3,*] **Ziheng Qin**[1,*] **Zhiyuan Yan**[6] **Ruoxin Chen**[4]
Yi Han[5] **Zhiheng Li**[1] **Junkai Chen**[1] **ZhiJin Chen**[1] **Kaiqing Lin**[8]
Jialiang Shen[7] **Lubin Weng**[1] **Jing Dong**[1] **Yan Wang**[9,†] **Shu Wu**[1,†]

[1]Institute of Automation, CAS [2]School of Advanced Interdisciplinary Sciences, UCAS
[3]Huazhong University of Science and Technology [4]Tencent YouTu Lab
[5]Southwest University [6]Peking University [7]The University of Sydney
[8]Shenzhen University [9]Tsinghua University
[*]Equal contribution [†]Corresponding author

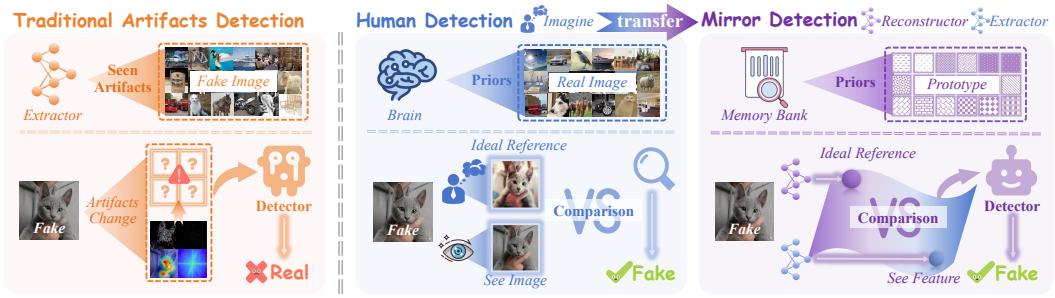

Figure 1: **AI-generated image detection paradigms.** Artifact cues vary across generators; real-image references aid human inspection. MIRROR reconstructs an Ideal Reference from a diversity-regularized real-image codebook, combining reconstruction residuals and retrieval uncertainty.

Abstract

High-fidelity generative models have narrowed the perceptual gap between synthetic and real images, putting media security under increasing pressure. Many AI-generated image (AIGI) detectors rely on artifact cues and struggle to generalise to new generators. We investigate whether a real-image reference, learned only from real images, provides a more transferable anchor than generator-specific forgery cues. We therefore cast AIGI detection as a Reference-Comparison problem that checks consistency against a learned real-domain feature representation. We propose **MIRROR** (**M**anifold **I**deal **R**eference **R**econstruct**OR**), which encodes real-image priors in a diversity-regularized discrete memory bank and uses Top-k attention to reconstruct an Ideal Reference. A supervised second phase combines reconstruction residuals and retrieval uncertainty while keeping the reference bank and projector fixed. To measure detector performance relative to human inspection, we build the Human-AIGI benchmark, which contains a psychophysically curated human-imperceptible subset. Across 14 benchmarks, MIRROR improves average benchmark scores over prior methods by 2.1 percentage points on six standard benchmarks and 8.1 points on seven in-the-wild benchmarks. On the original Human-AIGI set, DINOv3-L reaches 89.6% fake-image accuracy over 27 generators, exceeding the reported means of lay users and computer-vision experts. Scaling to DINOv3-H+ raises this score to 93.6%. Code: `https://github.com/handsome-rich/MIRROR`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

High-fidelity generative models have narrowed the perceptual gap between synthetic and real images, making generated content hard to distinguish from authentic photographs [3, 19, 22, 40]. This trend puts the reliability of online media, forensic analysis, and downstream vision systems under pressure, and motivates detection frameworks that remain effective when human inspection no longer suffices [26].

Many AIGI detectors treat real-versus-fake recognition as a binary classification problem, learning discriminative cues from labeled real and generated images [7, 18, 36, 47]. These cues can include visible artifacts and subtler feature-level traces, and work well when they transfer to the test distribution. Performance, however, can drop substantially on images from unseen generators or after post-processing [15, 42], which points to a generalization issue. This raises a more concrete question: as backbones grow and pretrained priors get stronger, do existing detection paradigms still keep gaining generalization?

We compare representative detectors against humans on unseen generated images and observe a counterintuitive trend: the benefit from a stronger pretrained prior plateaus for the evaluated baselines (Fig. 2). Pretraining-based methods [9, 34] outperform the evaluated training-from-scratch method [44], but larger backbones yield limited additional accuracy in these configurations. This motivates architectures that make better use of pretrained representations.

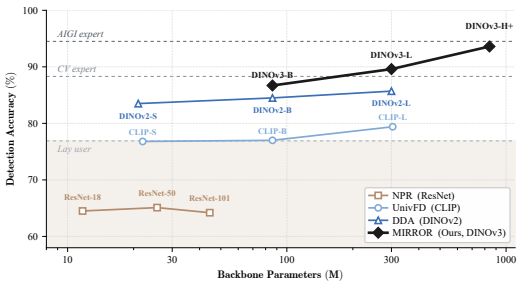

Figure 2: **Scalability on Human-AIGI.** The evaluated detectors show limited gains as their backbones scale (NPR [44], UnivFD [34], DDA [9]); MIRROR improves from Base to Huge.

We hypothesize that this saturation partly reflects how detection objectives use pretrained knowledge. In typical supervised pipelines, pretrained models act as feature extractors for cues present in the available generators (Fig. 1, left), without an explicit reference for the regularities of real images. The resulting decision boundary can depend on a forgery distribution that changes as new generators appear. Such dependence can limit transfer and the benefit of a larger backbone. Our approach introduces a fixed real-domain reference into this supervised setting, so that the classifier can compare its input with an anchor learned without generated images.

Human performance provides a complementary motivation for explicit references. In our psychophysical study, computer-vision experts with long exposure to natural images outperform non-experts, and lay users become more accurate when given real-image references. These observations suggest a role for learned or supplied real-image priors [11, 12]. As illustrated in Fig. 1 (Middle), they motivate an operational counterpart: anchor decisions on a fixed real-domain reference and learn which deviations are informative for detection.

This is the motivation for MIRROR (Manifold Ideal Reference ReconstructOR), a framework that casts forgery detection as a Reference-Comparison process. Following the manifold hypothesis [2, 5], we approximate structure in real-image features using a learnable discrete memory bank trained only on real images (Fig. 1, Right). A diversity regularizer discourages redundant prototypes in this overcomplete bank. Given an input, MIRROR retrieves a small set of prototypes and constructs an Ideal Reference, then measures both the reconstruction residual and retrieval uncertainty. Sparse retrieval controls reconstruction flexibility, while freezing the real-only bank and projector during supervised adaptation provides both evidence branches with a consistent real-domain reference.

Existing benchmarks emphasise generator coverage or semantic diversity but do not directly measure detector performance relative to human inspection. To make this measurable, we build the Human-AIGI Benchmark. Through psychophysical experiments with 50 participants we record accuracy, confidence, and response latency, and curate a Human-Imperceptible subset of samples that either deceive observers or trigger pronounced hesitation. We use "Superhuman Crossover" to describe benchmark performance above the reported mean of a specified observer group.

Across 14 benchmarks, MIRROR improves the average benchmark score over the strongest reported baseline by 2.1 percentage points on the six standard benchmarks and by 8.1 points on the seven in-the-wild benchmarks. On the Human-Imperceptible subset of Human-AIGI it reaches 89.5% fake-image accuracy, 2.6 points above the best prior detector in our comparison. Fig. 2 shows continued gains as the backbone scales: DINOv3-H+ reaches 93.6% on the original set, approaching the dedicated AIGI-expert mean of 94.5%.

Our contributions are: (i) investigating the limited scaling gains of representative artifact-driven detectors and formulating a Reference-Comparison architecture anchored on a real-only reference; (ii) proposing MIRROR, which combines a diversity-regularized prototype bank, Top-k feature reconstruction, a fixed reference during supervised adaptation, and joint residual–uncertainty evidence, with controlled ablations supporting the relative stability and transferability of real-domain prototypes; and (iii) introducing the Human-AIGI Benchmark with a psychophysically curated Human-Imperceptible subset spanning 27 modern generators, enabling comparison with lay users, computer-vision experts, and dedicated AIGI experts through generated-image detection accuracy (F.Acc.).

2 Related Work

2.1 AI-Generated Image Detection

AIGI detection research has largely focused on generalisation to unseen generators [34, 47]. Image forensics models sensor noise [61, 62], while early AIGI methods used low-level artifacts, including generator-specific fingerprints in the spatial or spectral domain [17, 18]. CNNSpot [47] showed cross-GAN transferability, and follow-up work has used frequency irregularities [36, 46, 49], upsampling traces [44], or selected local patches [54]. These texture-level cues are fragile when generators evolve or when images go through real-world post-processing [31].

A second line of work uses pretrained foundation models [14, 42, 50, 51] and semantic supervision [63, 64]. UnivFD [34] pairs CLIP [37] features with a lightweight classifier; FatFormer [30] and AIDE [52] adapt CLIP representations through task-specific modules; Effort [53] disentangles semantic and forgery-related subspaces. Selective representation use also appears in video readouts [16] and multimodal unlearning [28].

A third line of work attributes performance drops to data bias and reduces shortcut learning through reconstruction or alignment between real and synthetic data [32], including FakeInversion [6], SemGIR [55], DRCT [8], B-Free [20], AlignedForensics [38], and Dual Data Alignment [9]. REM [31] instead models a real-image envelope through perturbed self-reconstruction and cross-domain consistency.

Real-image modeling [58, 59, 65, 66], reconstruction, and memory-based detection provide precedents: DRAEM [56] learns restoration and defect segmentation, PatchCore [41] uses nominal patch neighbours, DIRE [48] classifies diffusion reconstruction errors, and AEROBLADE [39] uses autoencoder errors without detector training. He et al. [21] classify deviations in a learned CLIP reconstruction space. MIRROR instead reconstructs input-conditioned references through sparse patch-prototype retrieval and combines residual and retrieval-uncertainty evidence (App. H).

2.2 Detection Benchmarks

Early AIGI benchmarks targeted GAN imagery, e.g. the CNNSpot dataset [47]. Later datasets broaden generator coverage: GenImage [60] and AIGCDetect [57] provide large-scale samples from many models, and DRCT-2M [8] provides million-level data designed for diffusion reconstruction. UniversalFakeDetect [34] introduces a cross-model evaluation protocol on CLIP backbones, Synthbuster [1] a systematic diffusion evaluation, and EvalGEN [9] an evaluation focused on autoregressive generators. These benchmarks largely rely on images synthesised from simple prompts, which is not always representative of forgeries designed to deceive humans.

In-the-wild benchmarks address this gap by collecting web images or simulating realistic degradation. Chameleon [52] curates high-resolution samples that are difficult even for humans; SynthWildx [13], WildRF [4], and AIGIBench [27] broaden web coverage; RRDataset [23] and RealChain [31] model post-processing and online propagation. These benchmarks characterize transfer

across web sources and post-processing conditions. Human-AIGI complements this coverage by defining a perceptually challenging subset through measured human judgments and response times.

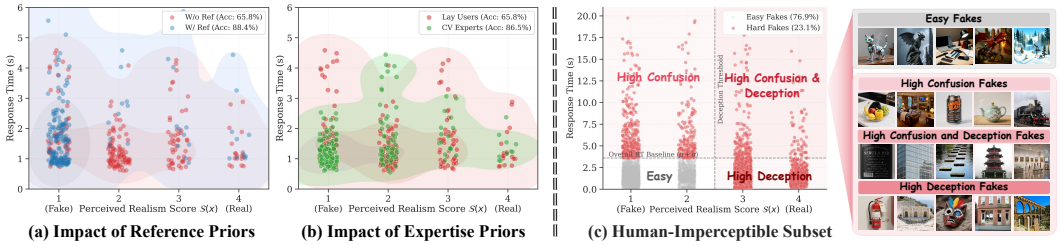

Figure 3: **Psychophysical Analysis on the Human-AIGI Benchmark.** Results are obtained from controlled psychophysical experiments on AI-generated images from 27 generators. **(a) Reference Priors.** Providing explicit reference images shifts hard samples (red) from high uncertainty with long response times to a higher-confidence regime (blue). **(b) Expertise Priors.** CV experts (green) consistently outperform lay users (red), consistent with a role for stronger internal perceptual priors. **(c) Human-Imperceptible Subset.** The *Hard* subset (red) selects generated images with a mean perceived-realism rating of at least 3 or a mean response time above $\mu_{rt} + \sigma_{rt}$; *Easy Fakes* (gray) satisfy neither condition.

3 A Stability Gap Between Real and Generated Distributions

A detector must contend with variation on both the real and generated sides. Within the natural-image domains assessed here, real-image features provide a relatively stable reference that can be learned without specifying a generator. Synthetic images come from an open, rapidly evolving family of generators, each with its own bias. A decision boundary fitted to the available generated samples can become brittle as that generator pool changes. This motivates examining a real-centered reference across the camera, content, and post-processing variations represented in these domains.

Our psychophysical experiments give behavioural evidence consistent with this view. Lay users who have not been systematically exposed to AIGI become substantially more accurate when given real-image references (Fig. 3(a)), and computer-vision experts, who have long exposure to natural-image statistics, outperform lay users without any external reference (Fig. 3(b)). These observations motivate grounding detection on real-domain regularities.

This view leads to MIRROR, which approximates the real-domain feature distribution with a learnable, diversity-regularized codebook and detects forgeries through reconstruction residuals and retrieval uncertainty. Sec. 6.3 shows higher detection performance with this real-domain anchor than with generated-only or mixed references.

4 Methodology

MIRROR is a two-stage learnable framework (Fig. 4). Phase 1 learns a real-only prototype bank and sparse reconstruction projector with a frozen feature encoder. Phase 2 freezes this reference and learns LoRA adapters and classification modules from real/fake supervision, combining complementary Comparison Residual and Reconstruction Perplexity evidence.

4.1 Phase 1: Encoding Reality Priors

Phase 1 aims to approximate structure in the real-image feature distribution $\mathcal{M}_{real}$. To retain local geometry and fine-grained texture information, we utilize a frozen DINO [43] encoder $\mathcal{E}$ to extract contextualized feature tokens:

$$F = \mathcal{E}(I) \in \mathbb{R}^{N \times D}, \tag{1}$$

where N counts non-CLS feature tokens and D is the feature dimension. We encode reality priors into a discrete memory bank $\mathbf{M} \in \mathbb{R}^{K \times D}$ containing K learnable prototypes. Each prototype is optimized through real-image reconstruction to capture recurring feature patterns. We construct the

Ideal Reference using multi-head cross-attention with a Top-k sparsity constraint. For attention head h, let Q_h, K_h, and V_h denote the learned query, key, and value projections of F and $\mathbf{M}$, where H is the number of attention heads and $d_h = D/H$. Reconstruction is

$$A_h = \mathrm{Softmax}_K\left(\mathrm{Mask}_k\left(\frac{Q_h K_h^\top}{\sqrt{d_h}}\right)\right), \quad \hat{F} = \mathrm{Proj}_{out}\left(\mathrm{Concat}_{h=1}^{H}[A_h V_h]\right), \tag{2}$$

where Mask_k retains the largest k logits along the prototype axis for every token and head, sets all other logits to $-\infty$, and is followed by softmax over that same axis. The attention module constructs the Ideal Reference as a learned sparse feature reconstruction.

The bank and all attention projections in Eq. 2 are trained exclusively on real images while the encoder stays frozen, using the following objective:

$$\mathcal{L}_{Phase1} = \frac{\|F - \hat{F}\|_F^2}{ND} + \lambda \|\bar{\mathbf{M}}\bar{\mathbf{M}}^\top - \mathbf{I}_K\|_F. \tag{3}$$

The first term is mean-squared real-feature reconstruction error; $\bar{\mathbf{M}}$ is the row-normalized bank, and λ controls the diversity penalty. For the overcomplete setting $K > D$, the second term encourages low prototype coherence; App. H derives the corresponding rank bound.

4.2 Phase 2: Reference-Comparison-Based Detection

In Phase 2, the memory bank $\mathbf{M}$ and all reconstruction projections in Eq. 2 remain frozen. We train encoder LoRA adapters and the detection head with real/fake labels. The measured mean pre-/post-LoRA feature cosine similarity is 0.991, showing that adaptation preserves substantial feature similarity.

Detection compares F with its **Ideal Reference** $\hat{F}$ through two signals: **Reconstruction Perplexity**, a learned uncertainty descriptor from attention maximum and entropy, and **Comparison Residual**, which encodes reconstruction error.

We compute attention maximum s_{max} and entropy s_{ent} over prototypes, then average over heads and tokens (App. H). For the residual, $\Delta F_i = F_i - \hat{F}_i$ is averaged over tokens before projection. The two image-level evidence vectors are

$$V_{per} = \mathrm{MLP}_{per}([s_{max}, s_{ent}]), \quad V_{res} = \mathrm{MLP}_{res}\left(\frac{1}{N}\sum_{i=1}^{N} \Delta F_i\right). \tag{4}$$

The classification head concatenates the encoder's global classification token f_{cls} with V_{per} and V_{res} to combine image-level context with both reference-comparison signals. The final forgery probability is the fake-class softmax probability of its two output logits:

$$y_{pred} = \mathrm{Softmax}_{fake}(\mathrm{MLP}_{\mathcal{C}}\left(\mathrm{Concat}[V_{per}, V_{res}, f_{cls}]\right)). \tag{5}$$

The framework is optimized by a binary cross-entropy loss:

$$\mathcal{L}_{Phase2} = -y \log y_{pred} - (1-y)\log(1 - y_{pred}), \tag{6}$$

where $y \in \{0, 1\}$ is the ground-truth label, with $y = 1$ for generated images. This is equivalently implemented as cross-entropy over the two logits.

5 The Human-AIGI Benchmark

5.1 Psychophysical Experimental Design

We conduct large-scale psychophysical experiments to probe the limits of human perception in AIGI detection. We recruit 50 participants, including lay users, computer vision experts, and AIGI detection experts, and evaluate them on a dataset $\mathcal{D}_{orig.}$ comprising over 30,000 images synthesized by 27 generators from T2I-COREBENCH [25]. Participants perform a binary forced-choice task with generated-image trials interleaved with real-image decoys. Reported accuracy uses the generated-image trials only. Human baselines use unaided judgments; the reference-assisted condition separately tests the effect of providing real-image examples. Beyond accuracy, we record perceived-realism ratings on a 4-point scale and response times for each trial. Every image receives at least

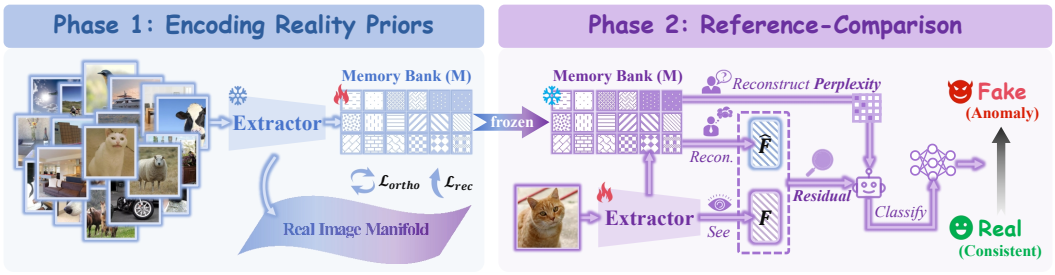

Figure 4: **Architecture of MIRROR.** In Phase 1, a frozen DINO encoder extracts real-image feature tokens to train a diversity-regularized prototype bank and sparse reconstruction projector. In Phase 2, the bank and projector stay fixed while LoRA adapters and the detection head are trained with real/fake supervision. The head combines Reconstruction Perplexity and Comparison Residual evidence, with global-token context, to detect deviations informative of generation. Sparse reconstruction supplies the learned real-domain reference used by both reference-comparison evidence branches.

Table 1: Performance comparison on **Standard Benchmarks**. We report Balanced Accuracy (B.Acc), JPEG Robustness (J.Rob), and the Resize Robustness (R.Rob) to comprehensively evaluate the detector's generalization and robustness performance. The best and second-best summary scores are highlighted in **bold** and underline, respectively. DRCT [8], Aligned [38], B-Free [20], and DDA [9] are evaluated using their officially released weights. Other methods based on architectural innovations are evaluated using the same training set as ours, and official-weight results for five baselines are additionally reported in the appendix.

Method	AIGCDetect			GenImage			DRCT-2M			UnivFakeDetect			Synthbuster			EvalGEN			Average		
	B.Acc	J.Rob	R.Rob	B.Acc	J.Rob	R.Rob	B.Acc	J.Rob	R.Rob	B.Acc	J.Rob	R.Rob	B.Acc	J.Rob	R.Rob	B.Acc	J.Rob	R.Rob	B.Acc	J.Rob	R.Rob
NPR [44]	63.9	64.0	48.2	73.7	73.7	46.6	59.5	58.3	37.4	51.1	51.1	49.2	64.6	64.0	34.4	66.1	65.7	64.2	63.2	62.8	46.7
UnivFD [34]	56.5	50.1	48.7	62.5	53.6	50.8	69.5	57.3	52.5	48.3	51.1	43.0	65.8	59.2	52.3	73.8	61.3	56.6	62.7	55.4	50.7
FatFormer [30]	82.1	71.5	80.0	71.5	58.4	68.4	53.9	52.1	52.4	86.7	68.9	85.2	69.2	57.1	65.6	56.7	53.4	57.5	70.0	60.2	68.2
SAFE [24]	48.4	57.8	49.4	47.7	64.4	49.4	50.3	62.4	49.2	49.7	48.5	49.8	51.6	53.8	49.9	50.2	55.4	49.1	49.7	57.1	49.5
C2P-CLIP [45]	80.0	65.7	77.6	71.1	55.6	67.8	54.4	54.0	55.7	87.3	68.8	85.2	69.4	57.3	66.4	56.8	52.1	65.8	69.8	58.9	69.8
AIDE [52]	84.0	53.4	77.0	88.6	55.3	76.3	59.2	54.6	72.5	77.9	51.5	70.8	75.4	76.8	62.5	59.5	58.8	76.4	74.1	58.4	72.6
DRCT [8]	67.1	67.1	67.5	78.8	78.5	80.3	97.0	93.0	98.4	66.8	69.5	69.7	81.0	80.6	83.3	81.9	85.5	82.4	78.8	79.0	80.3
Aligned [38]	55.7	54.5	55.3	57.5	55.9	57.0	54.9	54.7	54.3	60.5	59.7	61.0	54.5	53.0	53.8	65.8	65.2	67.5	58.2	57.2	58.2
B-Free [20]	84.7	85.0	82.7	89.6	90.2	87.9	99.2	99.0	98.2	87.8	87.0	86.2	95.7	95.9	94.8	94.6	93.7	92.5	91.9	91.8	90.4
DDA [9]	81.5	81.8	79.1	88.9	88.7	88.1	97.0	97.7	95.6	69.4	69.3	66.7	96.5	95.6	94.6	96.6	90.1	93.9	88.3	87.2	86.3
MIRROR (DINOv2)	90.5	90.0	88.6	91.3	94.1	93.7	92.8	88.9	94.2	84.6	85.8	85.0	97.0	96.4	97.6	98.1	97.1	96.6	92.4↑0.5	92.1↑0.3	92.6↑2.2
MIRROR (DINOv3)	91.7	91.5	91.5	94.2	96.7	94.9	93.0	90.4	93.4	88.2	88.3	87.5	98.1	97.5	97.6	99.0	98.5	98.6	**94.0**↑2.1	**93.8**↑2.0	**93.9**↑3.5

two independent annotations; $S(x)$ and $RT(x)$ denote the mean rating and response time across its annotators (App. D). These measurements capture decision outcomes and uncertainty in distinguishing generated images from real ones. Figure 3(a,b) illustrates the effects of reference assistance and expertise on observer responses. Both evaluation subsets considered below contain generated images; we therefore report fake-image accuracy (F.Acc.).

5.2 Definition of the Human-Imperceptible Subset

Based on these observations, we define a *human-imperceptible subset* consisting of generated samples that challenge human perception. Such samples either deceive observers into judging them as real or induce significant hesitation during decision-making. Formally, following psychophysical principles, we define the human-imperceptible subset $\mathcal{D}_{hard}$ as:

$$\mathcal{D}_{hard} = \{x \in \mathcal{X}_{gen} \mid S(x) \geq \tau_{real} \vee RT(x) > \mu_{rt} + \sigma_{rt}\}, \tag{7}$$

where $\mathcal{X}_{gen} = \mathcal{D}_{orig.}$, $S(x) \in [1, 4]$ is the image-level mean perceived-realism rating, τ_{real}=3 is the deception threshold, and μ_{rt}, σ_{rt} are the mean and standard deviation of image-level mean response times over $\mathcal{D}_{orig.}$. The latency rule selects images with mean response times more than one standard deviation above the benchmark mean, providing an empirical threshold for prolonged response times. Because the subset only contains images that look realistic or trigger pronounced hesitation in humans, it tests whether a detector can flag forgeries that challenge human inspection. Reporting F.Acc. jointly on $\mathcal{D}_{orig.}$ and $\mathcal{D}_{hard}$ measures detection performance on both the complete generator set and its perceptually challenging subset. The full participant breakdown, scale calibration, control procedure and human-baseline confidence intervals are deferred to App. D.

Table 2: Performance comparison on **In-the-wild Benchmarks**. These datasets include degradations caused by online propagation. The five CO-SPY columns report fake-image accuracy (F.Acc.); the other eleven columns report B.Acc. Avg.Score is the unweighted mean of all sixteen columns. The best and second-best summary scores are highlighted in **bold** and underline, respectively.

Method	Cham-eleon	SynthWildx			WildRF			AIGIBench		CO-SPY-Bench/in-the-wild					RR-Dataset	BFree-Online	Avg. Score
		DALLE3	Firefly	Midj.	FB	Reddit	Twitter	SocRF	ComAI	Civitai	DALLE3	instavibe.ai	Lexica	Midj.v6			
NPR [44]	55.2	62.9	53.8	63.0	53.8	53.7	57.7	55.8	52.0	51.3	74.4	6.0	3.9	80.7	48.3	40.5	50.8
UnivFD [34]	39.5	52.5	52.4	50.7	54.1	55.1	66.5	51.6	45.4	89.4	76.6	78.3	63.1	84.1	51.1	57.2	60.5
FatFormer [30]	57.8	52.6	56.6	50.0	52.5	65.4	39.7	55.7	50.4	0.4	2.2	0.0	0.3	15.2	50.4	32.7	36.4
SAFE [24]	56.8	48.8	46.6	48.7	49.7	49.2	34.7	49.2	49.3	0.1	0.9	0.5	0.3	2.4	49.3	32.6	32.4
C2P-CLIP [45]	57.6	49.6	57.9	49.6	51.9	67.6	40.4	58.1	50.4	0.0	4.5	0.2	0.2	23.2	50.0	32.7	37.1
AIDE [52]	65.7	66.4	48.2	66.4	61.6	66.3	52.5	59.2	62.2	15.8	34.0	7.2	19.5	64.5	57.6	52.1	50.0
DRCT [8]	79.8	85.9	58.9	90.5	90.3	66.8	79.6	71.3	84.6	99.1	82.1	22.0	55.8	94.8	58.2	77.1	74.8
Aligned [38]	61.3	49.6	53.9	52.4	48.4	54.0	40.6	51.0	60.2	7.5	8.4	6.2	11.5	9.7	47.7	38.1	37.5
B-Free [20]	78.3	96.1	92.3	95.3	95.6	85.5	96.7	84.9	79.7	98.8	97.4	5.9	53.6	98.6	69.5	87.1	82.2
DDA [9]	83.5	91.1	84.7	91.6	85.3	82.5	89.3	79.9	88.9	99.7	94.6	35.4	73.7	97.9	70.3	81.2	83.1
MIRROR (DINOv2)	85.4	91.2	86.1	89.5	93.8	90.2	92.7	82.3	88.9	94.6	93.4	72.5	91.1	85.6	76.8	84.3	87.4 ↑4.3
MIRROR (DINOv3)	90.7	95.9	88.4	94.9	97.1	96.6	96.4	87.6	93.4	99.2	98.1	75.0	94.2	89.8	78.9	83.0	**91.2** ↑8.1

Table 3: Performance comparison on the **Human-AIGI Benchmark**. We report fake-image accuracy (F.Acc., %) on the generated-only original test set (**Orig.**) and curated **Human-Imperceptible** subset (**Hard**). Average is the macro-average across the 27 generators.

Generator	NPR		UnivFD		C2P		FatFormer		SAFE		AIDE		DRCT		Aligned		B-Free		DDA		MIRROR (Ours)	
	Orig.	Hard	Orig.	Hard	Orig.	Hard	Orig.	Hard	Orig.	Hard	Orig.	Hard	Orig.	Hard	Orig.	Hard	Orig.	Hard	Orig.	Hard	Orig.	Hard
Diffusion Models																						
SD-3-Medium	61.2	52.4	80.3	76.2	1.2	0.0	0.8	0.0	32.0	42.9	34.4	28.6	66.2	76.2	12.2	0.0	93.9	100.0	94.5	95.2	91.1	100.0
SD-3.5-Medium	68.8	90.6	87.0	84.4	2.4	0.0	2.2	0.0	28.6	21.9	12.4	6.3	52.2	46.9	16.8	9.4	97.5	93.8	93.9	93.8	92.3	93.7
SD-3.5-Large	70.7	69.7	82.0	93.9	2.7	0.0	2.9	0.0	28.6	18.2	8.8	3.0	66.2	57.6	13.2	6.1	96.8	100.0	95.3	93.9	90.0	93.9
FLUX.1-schnell	67.6	69.7	77.3	79.8	0.7	0.0	0.4	0.0	29.0	28.3	16.4	10.1	53.8	48.0	21.0	16.2	78.6	74.2	93.2	95.5	83.7	91.4
FLUX.1-dev	68.3	65.2	80.5	81.9	1.5	0.7	0.4	0.7	35.2	36.2	23.0	21.7	51.1	50.7	21.2	21.0	54.3	39.1	87.2	89.1	86.5	86.9
FLUX.1-Krea-dev	65.3	63.3	86.5	89.3	0.8	0.4	0.6	0.4	42.2	47.4	8.4	3.7	20.3	23.0	12.8	9.6	24.1	18.9	23.1	18.5	64.2	62.5
PixArt-α	66.1	57.0	81.2	83.0	5.9	5.0	5.8	8.0	58.8	59.0	52.2	49.0	93.5	95.0	12.9	10.0	99.8	99.0	99.0	99.0	99.5	100.0
PixArt-Σ	64.7	61.9	89.9	94.3	3.0	1.0	2.0	0.0	44.3	43.8	30.5	28.6	88.7	87.6	14.4	14.3	99.8	99.1	97.4	97.1	99.7	100.0
HiDream-I1	74.0	72.8	76.2	85.5	0.7	0.9	0.4	0.0	25.8	31.1	14.8	15.3	77.6	82.1	11.2	11.9	81.1	79.6	81.6	82.1	88.0	84.6
Qwen-Image	53.1	42.3	87.5	94.2	0.2	0.0	0.5	0.0	37.5	48.1	1.0	1.9	48.3	38.5	19.3	17.3	57.9	40.4	90.8	94.2	86.2	86.5
Autoregressive Models																						
Infinity-8B	67.3	67.0	66.7	71.0	34.3	20.0	33.0	20.0	39.7	43.0	6.2	1.0	57.0	66.0	21.8	16.0	99.3	96.0	98.8	98.0	99.8	100.0
GoT-R1-7B	57.6	75.5	84.5	91.8	84.1	67.4	49.1	28.6	59.1	61.2	29.2	16.3	15.5	6.1	46.0	34.7	96.3	95.9	90.8	85.7	99.9	100.0
Unified Models																						
BAGEL	71.7	69.0	80.1	83.0	2.7	2.0	1.5	1.0	39.7	36.0	16.1	18.0	57.5	57.0	19.9	24.0	88.1	89.0	93.6	97.0	90.3	92.0
BAGEL w/Think	67.9	67.8	84.1	90.1	3.0	0.8	0.9	0.8	37.0	30.6	21.5	22.3	61.1	59.5	21.8	28.9	90.8	87.6	96.3	98.4	96.7	98.3
show-o2-1.5B	56.4	71.4	85.0	100.0	22.4	0.0	15.4	14.3	20.6	42.9	14.7	42.9	88.8	85.7	35.0	14.3	100.0	100.0	97.4	100.0	99.9	100.0
show-o2-7B	39.2	53.3	87.9	86.7	86.2	76.7	53.6	46.7	38.4	43.3	24.5	13.3	92.6	96.7	56.8	46.7	100.0	100.0	96.8	96.7	99.9	100.0
Janus-Pro-1B	67.9	76.2	86.0	85.7	72.0	66.7	44.9	47.6	47.3	76.2	38.3	19.1	14.9	9.5	55.6	9.5	91.0	100.0	82.1	85.7	89.3	95.2
Janus-Pro-7B	69.2	63.3	84.4	93.3	71.3	46.7	43.4	33.3	49.5	73.3	27.4	20.0	15.9	13.3	52.6	43.3	93.4	90.0	88.4	93.3	97.5	93.3
BLIP3o-4B	60.5	54.5	73.8	72.3	31.3	12.9	46.8	42.6	64.1	62.4	16.9	16.8	83.5	80.2	16.6	16.8	99.3	100.0	99.6	100.0	97.8	98.0
BLIP3o-8B	65.3	65.1	73.5	78.6	39.4	21.4	51.4	49.5	70.0	78.6	19.8	20.4	76.8	70.9	15.3	11.7	99.4	100.0	99.1	99.0	97.4	96.1
OmniGen2-7B	66.8	55.2	76.4	82.8	7.0	6.9	5.8	5.2	33.4	37.9	34.8	27.6	65.7	77.6	35.4	32.8	97.4	94.8	93.1	96.6	84.1	98.2
Closed-Source Models																						
Seedream 3.0	63.0	50.0	76.0	77.3	2.9	0.0	7.1	4.6	35.5	47.7	40.9	64.8	92.4	92.1	19.6	10.2	81.9	75.0	94.3	92.1	84.1	78.4
Gemini 2.0 Flash	45.6	50.4	85.5	86.9	11.0	7.3	7.7	6.9	39.4	36.5	25.8	15.0	78.5	77.4	24.3	22.6	90.1	83.9	86.6	83.2	79.2	79.5
Nano Banana	73.9	80.9	72.3	78.3	0.5	0.7	0.4	1.3	25.5	20.4	2.3	1.3	83.3	81.6	14.4	18.4	50.8	41.5	61.9	57.9	70.7	64.4
Imagen 4	73.2	78.3	70.8	74.6	0.9	0.0	0.2	0.0	33.6	33.3	11.4	6.7	75.7	70.4	13.3	22.1	60.1	64.2	80.8	62.9	63.5	52.5
HunyuanImage-3.0	82.8	82.7	70.8	70.2	6.0	1.4	10.2	5.3	15.1	22.1	0.1	0.0	93.3	82.7	3.8	5.8	99.2	97.1	98.9	95.7	97.0	89.4
GPT-Image	75.3	81.7	74.2	71.7	0.3	0.4	0.1	0.0	45.3	55.8	0.1	0.0	2.5	1.3	3.6	7.9	18.1	22.1	63.1	45.8	92.1	81.5
Average	65.3	66.2	80.0	83.6	18.3	12.6	14.4	11.7	39.1	43.6	19.7	17.5	62.0	60.5	22.6	17.8	82.9	80.8	88.1	86.9	**89.6**	**89.5**

6 Experiments

6.1 Experimental Setup

Datasets. We evaluate the out-of-distribution (OOD) detection performance of MIRROR on a comprehensive set of benchmarks. To avoid shortcut learning caused by formatting discrepancies, all datasets are verified following the DDA protocol [9], ensuring no systematic differences in image format between real and fake samples. **(1) Standard Benchmarks:** We evaluate on *AIGCDetect* [57], *GenImage* [60], *DRCT-2M* [8], *UnivFakeDetect* [34], *Synthbuster* [1], and *EvalGEN* [9]. These benchmarks cover dozens of generative models spanning GANs, diffusion models, and autoregressive architectures. **(2) In-the-wild Benchmarks:** To assess real-world generalization, we use *Chameleon* [52], *SynthWildx* [13], *WildRF* [4], *AIGIBench* [27], *CO-SPY-Bench* [10], *BFree-Online* [20], and *RRDataset* [23]. These datasets are collected from complex network environments and include realistic degradations introduced during online propagation. **(3) Human-AIGI Benchmark:** Crucially, we evaluate MIRROR on $\mathcal{D}_{orig.}$ and $\mathcal{D}_{hard}$, defined in Sec. 5, to measure performance in high-risk scenarios where human visual inspection fails.

Implementation Details. In Phase 1, we adopt DINOv3-large [43] as the feature extractor and collect 200k real images from MSCOCO [29] for self-supervised training. The memory bank size is set to $K = 4096$. We also evaluate a DINOv2 [35] variant. In Phase 2, we use a LoRA-finetuned DINOv3 as the feature extractor and finetune on SDv1.4 images from GenImage [60],

JPEG-reencoded at QF 96 to align their format with real images. Training is performed with the AdamW optimizer [33], an initial learning rate of 1×10^{-4}, for 5 epochs, using a cosine annealing learning rate schedule. The final input size is 224×224. The released implementation uses mixed crop/resize modes during training and conditional short-side downsampling to 512 followed by a center crop at inference; full preprocessing details are given in App. G.

Evaluation Metrics and Comparative Methods. We report Balanced Accuracy (B.Acc), JPEG Robustness (J.Rob) under QF=90 compression, and Resize Robustness (R.Rob) under resize=0.9 on the standard benchmarks. B.Acc is the mean of real- and fake-image accuracies. Synthbuster and EvalGEN are paired with the corresponding MSCOCO evaluation images for B.Acc.; the five generated-only CO-SPY columns and Human-AIGI use F.Acc. In-the-wild Avg.Score is the unweighted mean of the sixteen displayed evaluation columns across seven datasets. We compare MIRROR against a broad set of state-of-the-art detectors. Artifact- and pretraining-based methods include NPR [44], UnivFD [34], FatFormer [30], C2P-CLIP [45], SAFE [24], and AIDE [52]. These methods capture generic forgery traces through architectural design or specialized training. Methods in this group are retrained using the same data configuration as MIRROR, including the real samples used in Phase 1, with their respective backbones. We also evaluate dataset-alignment-based detectors, including DRCT [8], Aligned [38], B-Free [20], and DDA [9]. These approaches mitigate semantic or formatting biases to isolate forgery-related cues. For methods that rely on alignment datasets, we evaluate them using their official pretrained weights.

6.2 Comparison with State-of-the-Art

Generalisation on standard benchmarks. On six standard datasets spanning a wide range of generative architectures (Table 1), MIRROR achieves an average Balanced Accuracy of 94.0%, above AIDE [52] and 2.1 percentage points above the strongest average baseline, B-Free [20]. The gain suggests that a real-image prior helps the detector avoid overfitting to the training distribution. Alignment-based methods such as B-Free are competitive under both clean and mildly compressed conditions, retaining 91.9% and 91.8%, respectively. MIRROR gives higher absolute robustness scores, outperforming B-Free by 2.0 points on JPEG Robustness (J.Rob) and 3.5 points on Resize Robustness (R.Rob).

Generalisation on in-the-wild benchmarks. In real-world deployment, images undergo compression, rescaling and other post-processing during online propagation. We evaluate on seven in-the-wild benchmarks (Table 2). The evaluated NPR [44] and SAFE [24] configurations transfer poorly in aggregate, while pretraining- or alignment-based methods achieve higher scores. MIRROR maintains an Avg.Score of 91.2% across the sixteen evaluation columns, 8.1 percentage points above DDA [9]; this combines paired B.Acc and generated-only F.Acc. Whereas DDA uses a large pool of real and VAE-reconstructed images, MIRROR learns its reference from readily available real images and uses SDv1.4-generated images for supervised fine-tuning.

Detection of human-imperceptible images. We finally compare MIRROR with human observers on the Human-AIGI benchmark (Sec. 5). Table 3 shows low aggregate performance for several low-level-cue detectors, while pretraining-based detectors improve substantially but still struggle on some human-imperceptible generators. MIRROR with DINOv3-L reaches 89.6% F.Acc on the original set, above the reported means of lay users (76.9%) and computer-vision experts (88.3%). Scaling to DINOv3-H+ raises original-set F.Acc to 93.6%, approaching the AIGI-detection-expert mean of 94.5%. DINOv3-L maintains 89.5% F.Acc on the human-imperceptible subset, supporting its use as an assistive tool for challenging forgeries. Human-baseline confidence intervals and the participant breakdown are reported in App. D.

Five independently trained MIRROR runs yield $94.1 \pm 0.8\%$ standard B.Acc, $90.9 \pm 0.7\%$ in-the-wild Avg.Score, and $89.5 \pm 1.2\%$ original Human-AIGI F.Acc (mean$\pm$standard deviation). These runs assess training variability across the three benchmark groups.

6.3 Ablation Study and Analysis

Capacity K and sparsity top-k. Figure 5 reports group scores across the codebook-capacity and projection-sparsity sweeps. The average follows an inverted-U in codebook capacity K and peaks at $K{=}4096$. The individual groups exhibit different capacity preferences. When K is too small, the

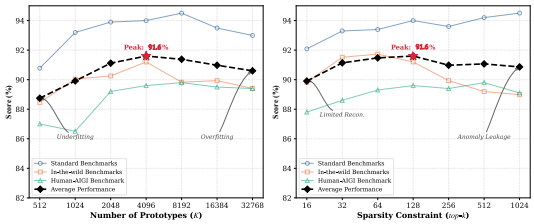

Figure 5: **Sensitivity of K and top-k.** Mean scores follow a broad inverted-U around K=4096 and top-k=128. Small values limit real-image reference coverage, while large values increase reconstruction flexibility and can let generative anomalies leak through prototype combinations.

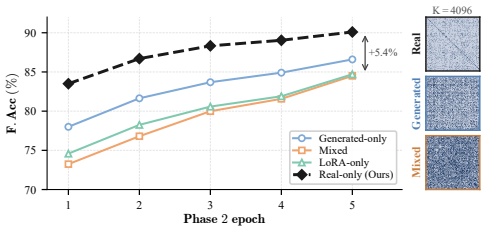

Figure 6: **Stability of the real anchor.** (Left) Phase-2 F.Acc on $\mathcal{D}_{orig.}$ over five epochs; real-only stays above the alternatives and mixed stays below LoRA-only. (Right) $|\mathbf{M}\mathbf{M}^\top|$ at K=4096: off-diagonal responses are weakest for the real-only bank.

prototypes cannot cover the diversity of real-image features and the residual gap driving detection shrinks; when K is too large, additional reconstruction flexibility can absorb generator-side variation despite the diversity regularizer. The projection sparsity top-k shows a similar aggregate trade-off: small k under-reconstructs complex textures, while large k can explain more generated-side deviations through prototype combinations. The two parameters therefore provide complementary control over reference coverage and reconstruction flexibility. The neighborhood of K=4096, top-k=128 is comparatively stable, and we use this configuration throughout.

Stability of the real-domain anchor. To isolate the role of the memory bank, we vary the source of the prototypes while keeping the LoRA backbone, classifier head, and training data identical. Figure 6 (left) tracks F.Acc on $\mathcal{D}_{orig.}$ over five Phase-2 training epochs for real-only, generated-only, mixed, and LoRA-only configurations. The real-only trajectory remains above the alternatives, and generated-only also stays above LoRA-only throughout the plotted epochs. Mixed remains slightly below LoRA-only and is the lowest curve from epoch 1 onward. Its recovery toward the no-bank baseline highlights the influence of prototype composition: the mixed reference yields lower scores than direct classification. The Gram maps in Figure 6 (right) show weak off-diagonal responses for real-only prototypes, clustered responses for generated-only prototypes, and more widespread responses for mixed prototypes. Thus generated-only references also improve detection, and the real-only bank provides the more stable and transferable anchor across the evaluated configurations.

Table 4: **Reference-Comparison signals and LoRA.** Single-branch rows augment LoRA-only; w/o LoRA freezes the backbone. Standard, Wild, and Human use B.Acc, Avg.Score, and F.Acc. Average is the unweighted mean of the three group scores.

Configuration	Standard	In-the-wild	Human-AIGI	Average
LoRA-only	88.4	82.2	83.7	84.8
$+ V_{per}$	90.1	84.3	85.0	86.5
$+ V_{res}$	92.5	85.1	88.5	88.7
w/o LoRA	91.8	88.4	87.1	89.1
MIRROR	**94.0**	**91.2**	**89.6**	**91.6**

Reference-Comparison signals and LoRA adaptation. Table 4 isolates the roles of the comparison head and backbone adaptation. LoRA-only classifies encoded features directly and reaches an average of 84.8%, 6.8 percentage points below MIRROR. Adding Reconstruction Perplexity or Comparison Residual individually improves this baseline, showing that the two signals provide complementary evidence. Keeping the encoder frozen during Phase 2 while training the complete Reference-Comparison head (w/o LoRA) reaches 89.1%, 4.3 points above LoRA-only. LoRA adaptation adds a further 2.5 points, bringing the full model to 91.6%. A separate parameter-matched DINOv3-L/LoRA control reaches 85.2% standard B.Acc versus MIRROR's 94.0%, supporting the design beyond head capacity (App. G).

Computational efficiency. The discrete memory bank and reference reconstructor augment the DINOv3-large feature extractor without running a separate image-generation model at inference. On a single NVIDIA Tesla V100, MIRROR runs at approximately 20 FPS and processes 1,000 samples

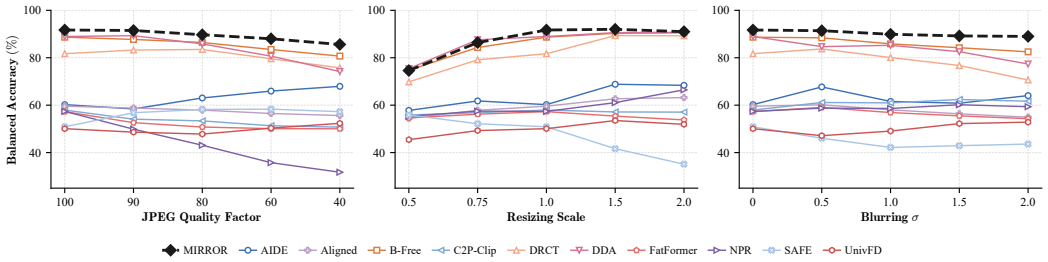

Figure 7: **Robustness to common image degradations.** Balanced accuracy under (left) JPEG quality from 100 to 40, (middle) resize scales from $0.5\times$ to $2.0\times$, and (right) Gaussian blur with $\sigma \in [0, 2]$. MIRROR (black, dashed) remains competitive across the sweep, with a larger reduction under strong downsampling than under JPEG compression or blur.

in 49.91 seconds, which is competitive with existing detectors (Fig. 8). It is faster than several higher-cost baselines while achieving higher standard and in-the-wild benchmark summary scores.

Robustness to common corruptions. Real-world images undergo compression, rescaling, and blurring during online propagation, which can alter high-frequency cues used by artifact-driven detectors. Figure 7 sweeps the three corruptions one at a time. Under JPEG quality from 100 down to 40, MIRROR decreases from 91.7% to 85.6%; FatFormer and NPR end below 60%, whereas AIDE reaches about 68%. Across rescaling from $0.5\times$ to $2.0\times$, MIRROR ranges from approximately 75% under strong downsampling to 92% near $1.5\times$, showing that resolution reduction remains consequential. Under Gaussian blur up to σ=2.0, MIRROR retains 89.0%, while B-Free, DDA, and DRCT decrease more strongly. Feature-level reference comparison remains informative under these perturbations, consistent with the cross-generator results.

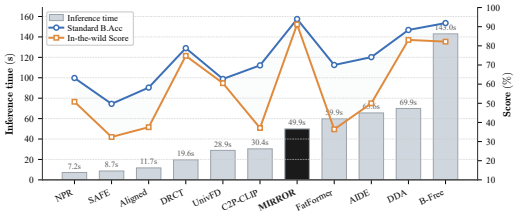

Figure 8: **Inference efficiency.** Time per 1,000 samples on V100. MIRROR offers competitive speed and achieves the highest score on both standard and in-the-wild benchmarks.

7 Conclusion

MIRROR addresses cross-generator AIGI detection by anchoring supervised decisions on a real-centered reference. A diversity-regularized prototype bank and sparse reconstruction projector are learned only from real images, then frozen while LoRA adapters and the detection head learn complementary residual and retrieval-uncertainty evidence. Across 14 benchmarks, including the proposed Human-AIGI benchmark, this design improves transfer over the evaluated baselines and benefits from larger pretrained backbones. On Human-AIGI's original generated-image set, MIRROR exceeds the reported means of lay users and computer-vision experts and approaches the dedicated AIGI-expert mean as the backbone scales. It maintains strong detection on the Human-Imperceptible subset. Reference coverage and the overlap between real and generated features determine the effectiveness of this comparison, motivating broader real-domain coverage and further study of fixed-reference adaptation.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62372454). The authors declare no competing interests.

References

[1] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2023.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2012. URL https://arxiv.org/abs/1206.5538.

[3] J. et al. Betker. Dall-e 3. https://openai.com/dall-e-3, 2023.

[4] Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world, 2024. URL https://arxiv.org/abs/2406.09398.

[5] Lawrence Cayton. Algorithms for manifold learning. Technical report, University of California, San Diego, 2005. URL https://www.lcayton.com/resexam.pdf.

[6] George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion, 2024. URL https://arxiv.org/abs/2406.08603.

[7] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize, 2020. URL https://arxiv.org/abs/2008.10588.

[8] B. et al. Chen. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024.

[9] 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, 2025. URL https://arxiv.org/abs/2505.14359.

[10] Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai, 2025. URL https://arxiv.org/abs/2503.18286.

[11] Cristiano Costa, Cristina Scarpazza, and Nicola Filippini. The anterior insula engages in feature-and context-level predictive coding processes for recognition judgments. *Journal of Neuroscience*, 45(5), 2025.

[12] David Daniel Cox and Thomas Dean. Neural networks and neuroscience-inspired computer vision. *Current Biology*, 24(18):R921–R929, 2014.

[13] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip, 2023. URL https://arxiv.org/abs/2312.00195.

[14] Manni Cui, Ruiqi Liu, Zijian Yu, Hao Tan, Zibo Wei, Zian Wang, Ziheng Qin, Huijia Zhu, Weiqiang Wang, Jun Lan, and Shu Wu. Learning continuous source responses for generalizable ai-generated image detection, 2026. URL https://arxiv.org/abs/2609.14316.

[15] Manni Cui, Ruiqi Liu, Dianyuan Zou, Ziheng Qin, Jingrui Xu, ZiAn Wang, Jianglan Wei, Han Zhou, Yu Liu, Yan Wang, and Shu Wu. Globalforge: Towards robust ai-generated image detection, 2026. URL https://arxiv.org/abs/2607.14684.

[16] Manni Cui, Ziheng Qin, ZiAn Wang, Ruiqi Liu, Dianyuan Zou, Jianglan Wei, Han Zhou, Yu Liu, Jingrui Xu, Wenhao Wang, and Zhenyu Zhang. Rethinking the readout: Unlocking video backbones for ai-generated video detection, 2026. URL https://arxiv.org/abs/2607.15321.

[17] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions, 2020. URL https://arxiv.org/abs/2003.01826.

[18] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition, 2020. URL https://arxiv.org/abs/2003.08685.

[19] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

[20] Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection, 2024. URL https://arxiv.org/abs/2412.17671.

[21] Qinghui He, Haifeng Zhang, Xiuli Bi, Bo Liu, Chi-Man Pun, and Bin Xiao. Leveraging arbitrary data sources for ai-generated image detection without sacrificing generalization, 2026. URL https://arxiv.org/abs/2603.00717.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

[23] Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation: Benchmarking ai-generated image detection in challenging scenarios, 2025. URL https://arxiv.org/abs/2509.09172.

[24] Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective, 2024. URL https://arxiv.org/abs/2408.06741.

[25] Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play?, 2025. URL https://arxiv.org/abs/2509.03516.

[26] Zhiheng Li, Zongyang Ma, Yuntong Pan, Ziqi Zhang, Xiaolei Lv, Bo Li, Jun Gao, Jianing Zhang, Chunfeng Yuan, Bing Li, and Weiming Hu. Making mllms blind: Adversarial smuggling attacks in mllm content moderation, 2026. URL https://arxiv.org/abs/2604.06950.

[27] Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem?, 2025. URL https://arxiv.org/abs/2505.12335.

[28] Junkai Lin, Junkai Chen, Siqi Hou, Yuhao He, Ruiqi Liu, Chenhan Jin, Shengze Xu, Tieyong Zeng, and Xiangyuan Lan. Toward fine-grained forgetting:attribute unlearning for multimodal large language models, 2026. URL https://arxiv.org/abs/2608.01008.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. URL https://arxiv.org/abs/1405.0312.

[30] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Yao Zhao, and Jingdong Wang. Forgery-aware adaptive transformer for generalizable synthetic image detection, 2023. URL https://arxiv.org/abs/2312.16649.

[31] Ruiqi Liu, Yi Han, Zhengbo Zhang, Liwei Yao, Zhiyuan Yan, Jialiang Shen, ZhiJin Chen, Boyi Sun, Lubin Weng, Jing Dong, Yan Wang, and Shu Wu. Beyond artifacts: Real-centric envelope modeling for reliable ai-generated image detection, 2025. URL https://arxiv.org/abs/2512.20937.

[32] Ruiqi Liu, Yi Han, Boyi Sun, Houxin He, Yuxuan Zhou, and Yan Wang. M^2EA: Multi-vae manifold envelope alignment for challenging AI-generated image detection. In *ICASSP 2026*, pages 9147–9151. IEEE, 2026.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. URL https://arxiv.org/abs/1711.05101.

[34] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models, 2023. URL https://arxiv.org/abs/2302.10174.

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

[36] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues, 2020. URL https://arxiv.org/abs/2007.09355.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

[38] Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images, 2024. URL https://arxiv.org/abs/2410.11835.

[39] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error, 2024. URL https://arxiv.org/abs/2401.17879.

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. URL https://arxiv.org/abs/2112.10752.

[41] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection, 2021. URL https://arxiv.org/abs/2106.08265.

[42] Jialiang Shen, Jiyang Zheng, Yunqi Xue, Huajie Chen, Yu Yao, Hui Kang, Ruiqi Liu, Helin Gong, Yang Yang, Dadong Wang, and Tongliang Liu. Dino-detect: A simple yet effective framework for blur-robust ai-generated image detection, 2025. URL https://arxiv.org/abs/2511.12511.

[43] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

[44] 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, 2023. URL https://arxiv.org/abs/2312.10461.

[45] 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, 2024. URL https://arxiv.org/abs/2408.09647.

[46] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space learning, 2024. URL https://arxiv.org/abs/2403.07240.

[47] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now, 2019. URL https://arxiv.org/abs/1912.11035.

[48] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection, 2023. URL https://arxiv.org/abs/2303.09295.

[49] Jiazhen Yan, Ziqiang Li, Fan Wang, Ziwen He, and Zhangjie Fu. Dual frequency branch framework with reconstructed sliding windows attention for ai-generated image detection, 2025. URL https://arxiv.org/abs/2501.15253.

[50] Jiazhen Yan, Ziqiang Li, Fan Wang, Boyu Wang, Ziwen He, and Zhangjie Fu. Dgs-net: Distillation-guided gradient surgery for clip fine-tuning in ai-generated image detection, 2025. URL https://arxiv.org/abs/2511.13108.

[51] Jiazhen Yan, Fan Wang, Weiwei Jiang, Ziqiang Li, and Zhangjie Fu. Ns-net: Decoupling clip semantic information through null-space for generalizable ai-generated image detection, 2025. URL https://arxiv.org/abs/2508.01248.

[52] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection, 2024. URL https://arxiv.org/abs/2406.19435.

[53] Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection, 2024. URL https://arxiv.org/abs/2411.15633.

[54] Liwei Yao, Sen Niu, Xiaomei Feng, Bofeng Zhang, Ruiqi Liu, and Kefeng Li. All patches are not equal: Focusing on a single exposed patch for AI-generated image detection. *Information Processing & Management*, 64(1):105058, 2027.

[55] Xiao Yu, Kejiang Chen, Kai Zeng, Han Fang, Zijin Yang, Xiuwei Shang, Yuang Qi, Weiming Zhang, and Nenghai Yu. Semgir: Semantic-guided image regeneration based method for ai-generated image detection and attribution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8480–8488, 2024.

[56] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem – a discriminatively trained reconstruction embedding for surface anomaly detection, 2021. URL https://arxiv.org/abs/2108.07610.

[57] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection, 2023. URL https://arxiv.org/abs/2311.12397.

[58] Nan Zhong, Mian Zou, Yiran Xu, Zhenxing Qian, Xinpeng Zhang, Baoyuan Wu, and Kede Ma. Self-supervised ai-generated image detection: A camera metadata perspective, 2025. URL https://arxiv.org/abs/2512.05651.

[59] Nan Zhong, Yiran Xu, and Mian Zou. Color matters: Demosaicing-guided color correlation training for generalizable ai-generated image detection, 2026. URL https://arxiv.org/abs/2601.22778.

[60] 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, 2023. URL https://arxiv.org/abs/2306.08571.

[61] Mian Zou and Heng Yao. Sensor-aware blind image splicing detection by noise probabilistic modeling. In De-Shuang Huang, Yijie Pan, Wei Chen, and Bo Li, editors, *Advanced Intelligent Computing Technology and Applications*, pages 3–14, Singapore, 2025. Springer Nature Singapore. ISBN 978-981-96-9815-8.

[62] Mian Zou, Heng Yao, Chuan Qin, and Xinpeng Zhang. Statistical analysis of signal-dependent noise: Application in blind localization of image splicing forgery, 2020. URL https://arxiv.org/abs/2010.16211.

[63] Mian Zou, Baosheng Yu, Yibing Zhan, Siwei Lyu, and Kede Ma. Semantic contextualization of face forgery: A new definition, dataset, and detection method, 2024. URL https://arxiv.org/abs/2405.08487.

[64] Mian Zou, Baosheng Yu, Yibing Zhan, Siwei Lyu, and Kede Ma. Semantics-oriented multitask learning for deepfake detection: A joint embedding approach, 2024. URL https://arxiv.org/abs/2408.16305.

[65] Mian Zou, Baosheng Yu, Yibing Zhan, and Kede Ma. Self-supervised learning for detecting ai-generated faces as anomalies, 2025. URL https://arxiv.org/abs/2501.02207.

[66] Mian Zou, Nan Zhong, Baosheng Yu, Yibing Zhan, and Kede Ma. Bi-level optimization for self-supervised ai-generated face detection, 2025. URL https://arxiv.org/abs/2507.22824.

A Detailed Benchmark Results and Official Weight Evaluation

A.1 Detailed Breakdown of Benchmark Performance

This section reports per-benchmark numbers that supplement the main benchmark tables. Tables 7 and 8 give the results on *AIGCDetect* and *DRCT-2M*, covering architectures from early GANs (ProGAN, StyleGAN) to modern diffusion models (SDXL). Table 9 reports *UniversalFakeDetect*, which includes diverse unseen generation pipelines. Tables 10 and 11 report fake-class accuracy (F.Acc.) for the generated-image collections *Synthbuster* (including Midjourney and DALL-E 3) and *EvalGEN*. Main-table B.Acc averages F.Acc. and paired MSCOCO real-image accuracy. Across these benchmarks, MIRROR stays consistent on diverse generative semantics and on high-fidelity samples where baseline methods often vary significantly between generators.

A.2 Evaluation of Official Weights

We evaluate the five baselines in Table 12 with both their officially released weights ("Official") and weights retrained on our format-aligned SDv1.4 set ("Ours"). The retrained configurations share our training source and format-alignment procedure, enabling comparison under a common data configuration. The official-weight evaluations characterize each method in its released configuration. Together, the two settings characterize transfer across released and format-aligned baseline training configurations.

B Heatmap Visualization

To show how MIRROR distinguishes real from synthetic content, we visualise the reconstruction residual across generators. Unlike Grad-CAM, which uses backpropagated logits, our heatmaps are read directly from the patch residual map $R_i = \|F_i - \hat{F}_i\|_2$, where F_i is a contextualized input feature and $\hat{F}_i$ is its sparse reconstruction from the prototypes. Larger residuals identify regions with weaker reconstruction under the learned real-image reference, exposing the spatial distribution of reference-comparison evidence.

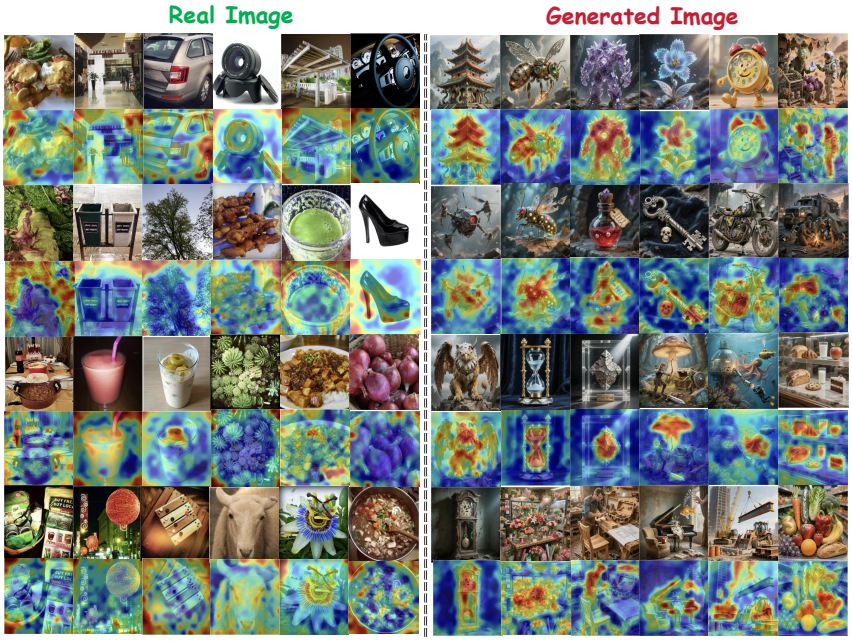

Figure 9: **Reconstruction residual heatmaps for authentic and synthetic images.** Authentic examples (left) generally have lower residuals, whereas generated examples (right) contain stronger localized responses. Red regions mark larger feature-reconstruction errors, highlighting local deviations from the learned real-image reference representation.

In Figure 9, the displayed real images generally have lower residuals, consistent with features that are well reconstructed by the learned real-image reference. Generated examples contain localized high-residual regions around textures, boundaries, or structures. The backbone contextualizes each patch feature, combining local appearance with information from the surrounding image. These maps make the comparison branch's spatial responses visible even when artifacts are subtle, illustrating how reference reconstruction supplies evidence for generated-image detection decisions.

C Visual Showcase of the Human-AIGI Benchmark

This appendix shows representative samples from the Human-AIGI Benchmark (Fig. 10). Compared with prior benchmarks that rely on a single generator or on easily distinguishable artifacts, our dataset emphasises cross-generator coverage and samples that are visually hard to tell apart from real images. It contains outputs from a range of recent generative models and diverse content domains, which is why we use it as the testbed for robustness, generalisation, and behaviour against forgeries that resist human inspection.



Figure 10: **Visual showcase of the Human-AIGI Benchmark.** (Left) **Representative Samples:** A diverse collection encompassing portraits, objects, and complex scenes curated from 27 state-of-the-art generators in T2I-COREBENCH (e.g., Flux, Hunyuan 3.0, Nanobanana). These samples reflect the broad semantic coverage and high fidelity of current commercial-grade AIGI. (Right) **The Human-Imperceptible Subset ($\mathcal{D}_{hard}$):** A rigorous subset consisting of "perceptual corner cases" identified through psychophysical trials with 50 participants. Following the criteria $S(x) \geq \tau_{real} \vee RT(x) > \mu_{rt} + \sigma_{rt}$, these images either reach a mean perceived-realism rating of at least 3 (the likely-real level) or exceed the mean-response-time threshold $\mu_{rt} + \sigma_{rt}$. This subset serves as a stringent testbed for evaluating whether a detector can identify generative threats that evade human inspection.

D Human-AIGI Benchmark: Protocol and Statistical Description

This appendix complements the qualitative description in Sec. 5 with protocol details and a statistical description of the psychophysical study.

Confidence scale. Each trial records a 4-point Likert rating that indicates how strongly the participant believes the image to be real, where 1 corresponds to *clearly fake*, 2 to *likely fake*, 3 to *likely real*, and 4 to *clearly real*. For each image, $S(x)$ is the mean of these ratings across annotators and lies in $[1, 4]$. The deception threshold in Eq. 7 is τ_{real}=3, selecting generated images whose mean rating reaches the likely-real level or above.

Response-time threshold. Trial response times are recorded in milliseconds from image onset to button press, and $RT(x)$ is their mean across the image's annotators. We identify pronounced

hesitation using $RT(x) > \mu_{rt} + \sigma_{rt}$, where μ_{rt} and σ_{rt} are the mean and standard deviation of $RT(x)$ over $\mathcal{D}_{orig.}$. The subset $\mathcal{D}_{hard}$ in Eq. 7 is the union of the deception and hesitation conditions: a generated image is included if its mean realism rating reaches 3, its mean response time exceeds the hesitation threshold, or it meets both conditions.

Participants. We recruited $N{=}50$ adult participants who were partitioned into three groups: 35 lay users without specialised perceptual training, 10 computer-vision experts who are graduate students or researchers in vision-related areas, and 5 AIGI-detection experts who actively work on synthetic-image detection. All participants provided informed consent before the trial and were free to withdraw at any time. Participation was voluntary, and no monetary compensation was offered.

Procedure and controls. The original set $\mathcal{D}_{orig.}$ contains more than $30{,}000$ images drawn from 27 generators. Each image was randomly assigned to a subset of participants and received at least two independent annotations; the per-image $S(x)$ and $RT(x)$ are means across those annotators. Participants viewed images one at a time in a randomised order, performed a binary forced choice (real or AI-generated) and rated realism on the 4-point scale; response times were logged automatically. Real-image decoys were interleaved with generated-image trials, and reported human accuracy uses $\mathcal{D}_{orig.}$ only. To control for fatigue and learning effects, each session began with 20 practice images that were excluded from the analysis, participants took a mandatory rest break every 50 images, and no per-trial feedback was provided. Display resolution, viewing distance, and image pixel size were uniform across participants.

Table 5: Human F.Acc. on $\mathcal{D}_{orig.}$ and participant-bootstrap 95% confidence intervals.

Group	n	**F.Acc.**	95% **CI**
Lay users	35	76.9%	$[74.2,\ 79.6]$
CV experts	10	88.3%	$[85.6,\ 91.0]$
AIGI experts	5	94.5%	$[92.0,\ 97.0]$
MIRROR (DINOv3-L)	—	$\mathbf{89.6}\%$	—
MIRROR (DINOv3-H+)	—	$\mathbf{93.6}\%$	—

Human baseline accuracy. Table 5 reports per-group fake-image accuracy on $\mathcal{D}_{orig.}$ with participant-bootstrap 95% confidence intervals. Dedicated AIGI-detection experts achieve the highest mean (94.5%), followed by computer-vision experts (88.3%) and lay users (76.9%). MIRROR reaches 89.6% on this set with DINOv3-L and 93.6% with DINOv3-H+, exceeding the lay-user and computer-vision-expert means and approaching the dedicated-expert reference. With DINOv3-L, it achieves 89.5% F.Acc. on the Human-Imperceptible subset, as reported in Table 3. Figure 2 presents the corresponding backbone-scaling comparison against these human reference groups.

E Limitations

MIRROR's real-image reference reflects the size and coverage of its training set. The 200k MSCOCO images provide a useful reference for the evaluated natural-image benchmarks, while specialized domains such as medical, satellite, or artistic imagery may require broader reference data and domain adaptation. The bank stays fixed during supervised adaptation, so reference coverage remains a distinct design choice. Detection also depends on separation between real and generated feature distributions: domain shifts can increase real-image residuals, while realistic generated features can closely match the reference. The remaining gap to dedicated AIGI experts motivates further improvements in representation and comparison. Our robustness evaluation addresses common JPEG, resizing, and blur perturbations; adaptive attacks against the detector are another direction for evaluation. Finally, contextualized features encode image context implicitly, and incorporating explicit global physical and semantic consistency offers a complementary direction for future research.

F Broader Impacts

MIRROR can assist journalists, platforms, and forensic analysts in verifying synthetic media, supporting responses to misinformation, impersonation, fabricated evidence, and non-consensual imagery. Human-AIGI provides a benchmark for comparing detector performance with human judgments on perceptually challenging generated images.

Public detector access can also lower the cost of adaptive forgeries through adversarial probing. False positives can cast doubt on authentic material, while false negatives can admit generated content. Deployment should combine detector predictions with contextual evidence and independent verification, including provenance, watermarking, and expert human review.

G Implementation and Additional Controls

Preprocessing. Our preprocessing preserves JPEG alignment and produces 224×224 inputs. Phase 2 training selects among center crop, direct resize, random crop, and random-short-side resize followed by random crop, with probabilities $(0.3, 0.1, 0.3, 0.3)$; the last mode samples a short side from 224 to 1024. Random cropping first resizes images that are too small. Generated training images receive JPEG QF 96 alignment after spatial preprocessing, and both classes then receive random JPEG recompression at an integer quality from 50 through 99. At inference, an image is resized to short side 512 only if its original short side exceeds 512, followed by a center 224×224 crop. PNG and BMP inputs are JPEG-reencoded at QF 96 after cropping; existing JPEG inputs are not reencoded by this format-alignment step. Its inference decision uses fake-class probability greater than 0.5. Phase 1 uses only real images, the same four spatial modes, a center crop, and random JPEG quality sampled from 50 through 100, followed by ImageNet normalization.

Architecture. The released DINOv3-L detector uses $H = 8$ attention heads with $d_h = 128$. LoRA adapters have rank 8 and scaling parameter $\alpha = 16$ and act on the query, key, and value projections in each transformer block. The uncertainty branch maps its two retrieval statistics to 64 features. The residual branch uses layer widths $D \to 512 \to 256$, and the classifier maps the concatenated $(D + 320)$-dimensional vector through widths $512 \to 2$. These modules use ReLU activations and dropout 0.3 after their hidden-layer activation functions.

Matched-backbone controls. Table 6 evaluates reference comparison using matched backbones. The DINOv3-L baseline pairs the same LoRA backbone with a parameter-matched MLP, controlling for backbone choice and head capacity. The CLIP pairs test transfer to two additional pretrained representations. MIRROR improves all three evaluation groups in every pair, supporting the contribution of reference comparison across the assessed backbones.

Table 6: Additional matched-backbone controls (%). Standard reports B.Acc.; Wild reports the mixed benchmark Avg.Score defined in Table 2; Human reports F.Acc. on Human-AIGI Orig. Each pair uses the same backbone and LoRA setting.

Backbone	Method	Standard	Wild	Human
DINOv3-L	LoRA + parameter-matched MLP	85.2	80.3	81.4
	MIRROR	94.0	91.2	89.6
CLIP-ViT-B/16	LoRA + classification	79.7	77.1	69.9
	MIRROR	86.4	83.8	79.2
CLIP-ViT-L/14	LoRA + classification	81.3	77.8	78.5
	MIRROR	87.5	84.1	83.7

H Reference Reconstruction and Retrieval Details

Diversity regularization for an overcomplete bank. The diversity penalty in Phase 1 uses the row-normalized prototype bank $\bar{\mathbf{M}}$. For an overcomplete bank with $K > D$ (here, 4096 prototypes in 1024 dimensions), $\operatorname{rank}(\bar{\mathbf{M}}\bar{\mathbf{M}}^\top) \le D$. Consequently, at least $K - D$ eigenvalues of its Gram matrix are zero, giving

$$\|\bar{\mathbf{M}}\bar{\mathbf{M}}^\top - \mathbf{I}_K\|_F \ge \sqrt{K - D}. \tag{8}$$

The penalty regularizes prototype correlations in this overcomplete bank. Top-k attention combines a small set of prototype values to form an input-conditioned reference, balancing real-feature coverage with sparse feature reconstruction capacity.

Image-level retrieval statistics. For a batch of B images, the masked and normalized attention tensor is $\mathbf{A} \in \mathbb{R}^{B\times H\times N\times K}$. The reconstruction sequence contains every non-CLS backbone token, including auxiliary tokens when present; reconstruction and pooling use this sequence. For each image b, we compute the maximum and entropy over prototypes first, then average over heads and tokens:

$$\begin{aligned} s_{max}^{(b)} &= \frac{1}{HN}\sum_{h=1}^{H}\sum_{i=1}^{N}\max_{m\in\{1,\dots,K\}} A_{bhim}, \\ s_{ent}^{(b)} &= -\frac{1}{HN}\sum_{h=1}^{H}\sum_{i=1}^{N}\sum_{m=1}^{K} A_{bhim}\log(A_{bhim}+\epsilon), \end{aligned} \tag{9}$$

where $\epsilon > 0$ ensures numerical stability. These statistics use post-softmax attention, including zeros at masked entries. For the residual, we form $\Delta F_i = F_i - \hat{F}_i$ and average over tokens before its learned projection. These statistics are the two inputs to MLP_{per} in Sec. 4.2; the residual average is the input to MLP_{res}. Reconstruction Perplexity denotes the descriptor $\mathrm{MLP}_{per}([s_{max}, s_{ent}])$ derived from these two retrieval statistics.

Reference construction and supervised adaptation. Phase 1 constructs the reference from real images. During Phase 2, the bank and reconstruction projections remain fixed while LoRA adapters and classification modules learn from real/fake labels. The adapted feature tokens query the same reference, and the two comparison signals and global context jointly predict generation. The pre-/post-LoRA cosine similarity quantifies the representation change during this adaptation, providing an empirical measure of encoder feature stability.

Relation to reconstruction and memory-based detection. Reconstruction and memory-based anomaly detection provide important precedents for MIRROR. DRAEM [56] jointly learns RGB restoration and synthetic-defect segmentation, while PatchCore [41] detects anomalies using nearest-neighbour distances to a coreset of nominal patch features. In AIGI detection, DIRE [48] trains a real/fake classifier on diffusion reconstruction-error maps, whereas AEROBLADE [39] uses pre-trained autoencoder reconstruction error without detector training. He et al. [21] learn an encoder–decoder attribution space from either real or generated frozen CLIP image embeddings and classify feature-reconstruction deviations. MIRROR uses learned low-coherence feature prototypes and Top-k retrieval to construct an input-conditioned Ideal Reference. This combines learned prototype coverage with sparse local reconstruction. Comparison Residual and Reconstruction Perplexity then capture reconstruction error and retrieval uncertainty. Its real-only bank and projector stay frozen during real/fake-supervised LoRA and head adaptation.

Sparse reconstruction configuration and terminology. K denotes the prototype count, whereas K_h denotes a key matrix. Each head combines at most k prototype values per token before the output projection. We use $K = 4096$ prototypes for real-feature coverage and retrieve $k = 128$ entries per head to limit reconstruction flexibility. Residual and retrieval statistics characterize how each input uses the real-image reference and provide complementary evidence for classification.

Table 7: Performance comparison on the *AIGCDetect* benchmark. The test benchmark uses balanced accuracy (B.Acc). The MIRROR row is shaded; bold marks the best average score.

Method	ADM	DALLE2	GLIDE	Midj.	VQDM	Big-GAN	Cycle-GAN	Gau-GAN	Pro-GAN	SDXL	SD1.4	SD1.5	Star-GAN	Style-GAN	Style-GAN2	WFR	Wukong	**Avg B.Acc**
NPR [44]	91.4	43.8	64.0	83.8	84.1	50.6	47.7	55.4	52.1	63.1	77.5	77.5	52.8	53.4	61.2	51.8	75.8	63.9
UnivFD [34]	68.5	75.0	69.2	71.0	55.5	39.5	43.5	36.6	42.1	73.9	69.2	69.4	47.1	43.5	39.2	51.2	65.6	56.5
FatFormer [30]	94.7	52.4	75.7	52.4	77.5	99.0	99.5	99.7	99.7	56.7	65.4	65.8	99.9	94.6	91.9	99.5	71.8	82.1
SAFE [24]	46.8	46.6	46.7	47.2	48.7	50.1	48.3	49.7	50.0	46.9	47.0	47.0	50.0	50.0	49.6	50.0	47.6	48.4
C2P-CLIP [45]	87.9	50.8	71.2	55.1	74.7	97.4	99.5	99.7	99.8	56.6	68.8	69.4	99.8	91.5	80.6	84.1	73.3	80.0
AIDE [52]	98.9	61.1	75.6	82.7	96.3	91.1	89.1	95.9	80.2	78.4	93.0	93.2	75.2	76.6	76.6	74.3	89.7	84.0
DRCT [8]	53.6	83.0	61.6	97.7	64.5	48.8	48.8	49.2	50.2	96.8	98.6	98.5	43.2	48.5	49.4	50.0	98.0	67.1
Aligned [38]	60.4	60.9	65.7	58.2	60.5	53.0	50.5	53.1	51.2	63.2	54.3	54.8	49.1	56.7	50.1	48.9	56.5	55.7
B-Free [20]	76.4	74.9	70.8	94.3	89.2	91.5	66.7	96.1	95.4	99.6	99.4	99.3	81.0	75.4	71.0	58.9	99.4	84.7
DDA [9]	90.7	93.9	88.8	93.9	67.8	81.1	64.8	81.6	73.1	98.0	96.7	96.7	67.9	67.8	76.7	49.8	96.8	81.5
MIRROR	75.6	88.9	93.1	92.5	95.9	97.2	73.3	95.8	96.8	99.8	99.9	99.8	87.6	91.9	91.1	78.9	99.8	**91.7**

Table 8: Performance comparison on the *DRCT-2M* benchmark.

Method	LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	SDv1-DR	SDv2-DR	SDXL-DR	**Avg B.Acc**
NPR [44]	48.8	59.9	59.6	62.3	68.7	74.6	72.7	63.7	62.5	73.3	55.8	53.0	46.0	50.9	50.1	50.4	59.5
UnivFD [34]	66.9	73.0	73.0	73.9	72.4	73.4	73.7	74.3	74.4	73.3	68.5	72.4	70.2	53.3	58.3	61.2	69.5
FatFormer [30]	54.6	50.3	50.1	50.0	50.5	49.9	50.0	49.9	50.6	53.2	52.3	53.0	65.4	70.8	59.6	51.9	53.9
SAFE [24]	50.0	50.0	50.0	49.9	49.9	49.9	49.9	49.9	50.1	49.9	51.1	50.1	50.2	52.1	50.3	51.6	50.3
C2P-CLIP [45]	59.6	51.1	50.9	50.5	52.2	50.8	50.6	50.3	50.4	53.1	55.6	53.2	68.6	66.6	54.9	52.8	54.4
AIDE [52]	58.7	60.0	60.0	56.8	51.9	66.0	51.3	50.7	61.8	56.6	55.3	52.2	51.5	83.1	72.1	60.1	59.2
DRCT [8]	99.8	97.3	97.8	99.3	96.8	96.8	99.3	87.3	99.3	95.3	95.3	99.8	96.8	96.3	99.8	94.3	97.0
Aligned [38]	51.1	53.7	53.8	51.9	57.9	63.1	51.7	57.2	58.1	69.1	50.8	52.5	50.8	52.9	51.9	51.3	54.9
B-Free [20]	99.7	99.7	99.7	99.6	99.1	99.4	99.2	98.3	99.6	99.6	99.8	99.8	99.8	99.5	99.6	95.4	**99.2**
DDA [9]	97.6	97.4	97.4	97.3	97.2	94.4	97.4	94.1	95.5	97.4	97.7	97.9	97.9	97.9	98.1	97.5	97.0
MIRROR	99.8	99.6	99.6	99.5	91.7	99.5	99.4	96.9	99.9	97.0	99.8	99.0	98.7	80.9	67.9	59.4	93.0

Table 9: Performance comparison on the *UniversalFakeDetect* benchmark.

Method	SITD	SAN	CRN	IMLE	Guided	LDM-200	LDM-200_cfg	LDM-100	Glide-100_27	Glide-50_27	Glide-100_10	Dalle	**Avg B.Acc**
NPR [44]	57.8	36.8	52.7	52.6	78.5	50.4	50.4	49.6	46.9	50.9	44.2	43.3	51.1
UnivFD [34]	54.7	46.8	23.0	36.6	53.5	44.0	45.3	43.7	61.6	62.2	62.1	46.1	48.3
FatFormer [30]	90.6	75.3	93.4	93.4	65.7	95.8	83.4	95.8	84.8	83.6	83.2	95.8	86.7
SAFE [24]	50.0	47.0	50.0	50.0	48.2	50.7	50.3	50.3	49.9	49.9	49.9	50.4	49.7
C2P-CLIP [45]	91.7	80.8	95.1	95.1	60.2	97.0	84.0	96.8	84.2	83.4	84.9	94.4	87.3
AIDE [52]	55.0	60.5	53.1	76.4	87.4	93.5	85.6	93.7	78.3	78.1	81.1	91.7	77.9
DRCT [8]	78.3	82.0	31.1	48.2	62.5	90.6	91.4	90.5	55.6	54.5	58.8	57.9	66.8
Aligned [38]	63.9	54.1	54.5	68.9	57.1	51.0	55.0	51.4	68.8	70.5	68.4	62.3	60.5
DDA [9]	48.9	87.0	60.8	67.5	88.1	73.7	73.5	72.9	69.1	67.9	71.6	51.7	69.4
B-Free [20]	76.4	92.9	90.4	94.2	75.5	97.1	97.1	97.2	77.9	79.2	82.0	93.6	87.8
MIRROR	63.3	80.6	76.4	76.9	79.8	98.9	98.8	99.0	95.1	94.7	95.6	98.6	**88.2**

Table 10: Fake-image accuracy (F.Acc., %) on the *Synthbuster* generated-image collection. The final column reports the unweighted mean of scores across all generators.

Method	DALLE2	DALLE3	Firefly	GLIDE	Midj.	SD1.3	SD1.4	SD2	SDXL	**Avg F.Acc.**
NPR [44]	72.3	70.2	74.0	26.8	77.9	68.3	69.0	87.7	77.5	69.3
UnivFD [34]	88.7	75.9	74.9	89.1	84.2	71.2	66.2	75.4	87.0	79.2
FatFormer [30]	32.5	0.4	80.7	43.9	5.9	57.6	56.7	27.9	40.7	38.5
SAFE [24]	0.5	2.5	0.6	0.0	6.4	9.5	7.6	1.8	0.4	3.3
C2P-CLIP [45]	25.2	0.6	78.1	39.4	10.0	54.7	54.5	44.5	44.4	39.0
AIDE [52]	16.5	23.6	0.3	42.1	51.5	88.7	90.2	77.4	67.3	50.8
DRCT [8]	3.9	36.8	13.5	21.6	98.9	94.9	93.8	99.9	96.4	62.2
Aligned [38]	23.7	2.3	28.0	22.0	13.1	4.2	3.4	4.6	14.9	12.9
DDA [9]	80.7	92.5	95.3	84.3	100.0	99.5	99.5	99.8	100.0	94.6
B-Free [20]	89.9	93.7	99.2	45.8	98.8	100.0	99.8	99.5	99.9	91.8
MIRROR	98.0	99.9	91.6	96.5	98.3	99.8	100.0	99.2	99.5	**98.1**

Table 11: Fake-image accuracy (F.Acc., %) on the *EvalGEN* generated-image collection. The final column reports the unweighted mean of scores across all generators.

Method	Flux	GoT	Infinity	NOVA	OmniGen	**Avg F.Acc.**
NPR [44]	74.4	70.5	69.7	76.0	70.7	72.3
UnivFD [34]	94.3	96.6	91.9	95.5	98.1	95.3
FatFormer [30]	0.4	2.6	8.6	54.2	1.7	13.5
SAFE [24]	0.3	0.3	0.9	0.2	0.6	0.5
C2P-CLIP [45]	1.8	6.7	11.4	46.6	2.7	13.8
AIDE [52]	17.9	24.7	3.4	16.3	33.4	19.1
DRCT [8]	52.2	71.2	58.3	67.9	70.5	64.0
Aligned [38]	30.0	46.2	20.8	47.0	32.9	35.4
DDA [9]	89.1	94.9	95.0	98.3	96.2	94.7
B-Free [20]	52.1	99.5	97.1	99.9	99.4	89.6
MIRROR	99.9	99.8	99.5	99.9	95.7	**99.0**

Table 12: Performance comparison between **Official Weights** and **Our Retrained Weights**. We report Balanced Accuracy (B.Acc), JPEG Robustness (J.Rob), and Resize Robustness (R.Rob). For **Synthbuster** and **EvalGEN**, B.Acc averages the fake-image accuracy of each collection and the real-image accuracy of its paired **MSCOCO** evaluation images. "Official" denotes results using officially released weights, and "Ours" denotes results obtained by training on our dataset.

Method	**Weights**	**AIGCDetect**			**GenImage**			**DRCT-2M**			**UnivFakeDetect**			**Synthbuster**			**EvalGEN**			**Average**		
		B.Acc	J.Rob	R.Rob	B.Acc	J.Rob	R.Rob	B.Acc	J.Rob	R.Rob	B.Acc	J.Rob	R.Rob	B.Acc	J.Rob	R.Rob	B.Acc	J.Rob	R.Rob	B.Acc	J.Rob	R.Rob
AIDE [52]	Official	81.8	53.7	75.5	88.4	57.3	77.3	66.7	59.8	63.6	81.1	51.8	72.8	80.6	54.0	78.3	59.7	59.9	69.8	76.4	56.1	72.9
	Ours	84.0	53.4	77.0	88.6	55.3	76.3	59.2	54.6	72.5	77.9	51.5	70.8	75.4	76.8	62.5	59.5	58.8	76.4	74.1	58.4	72.6
FatFormer [30]	Official	51.4	51.1	52.0	52.2	51.3	52.6	53.0	51.6	53.5	52.5	52.1	52.5	53.9	52.5	54.6	54.3	52.8	54.9	52.9	51.9	53.4
	Ours	82.1	71.5	80.0	71.5	58.4	68.4	53.9	52.1	52.4	86.7	68.9	85.2	69.2	57.1	65.6	56.7	53.4	57.5	70.0	60.2	68.2
NPR [44]	Official	72.4	60.0	68.8	84.7	67.9	74.9	70.3	64.7	63.1	64.2	56.2	58.0	82.6	63.3	76.2	82.9	80.5	77.9	76.2	65.4	69.8
	Ours	63.9	64.0	48.2	73.7	73.7	46.6	59.5	58.3	37.4	51.1	51.1	49.2	64.6	64.0	34.4	66.1	65.7	64.2	63.2	62.8	46.7
SAFE [24]	Official	49.4	50.1	35.3	50.6	50.1	34.4	49.0	49.9	44.8	61.4	49.8	46.5	48.9	50.0	25.7	51.4	50.0	43.9	51.8	50.0	38.4
	Ours	48.4	57.8	49.4	47.7	64.4	49.4	50.3	62.4	49.2	49.7	48.5	49.8	51.6	53.8	49.9	50.2	55.4	49.1	49.7	57.1	49.5
UnivFD [34]	Official	76.3	71.0	73.6	67.2	62.5	65.2	62.1	61.9	58.7	76.0	70.4	72.7	68.5	64.6	65.0	58.2	57.7	55.9	68.1	64.7	65.2
	Ours	56.5	50.1	48.7	62.5	53.6	50.8	69.5	57.3	52.5	48.3	51.1	43.0	65.8	59.2	52.3	73.8	61.3	56.6	62.7	55.4	50.7

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Sec. 1 state the reference architecture, Human-AIGI benchmark, and cross-generator results. Secs. 4–6, Tables 1–3, and App. A.1 define the metrics and support the reported experimental comparisons.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: App. E discusses reference coverage, domain shifts, feature-distribution overlap, the gap to dedicated experts, and adaptive attacks. App. F describes deployment risks from prediction errors and adversarial detector probing.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: App. H states the row-normalized bank and $K > D$ assumptions and derives the Gram-penalty lower bound using rank and eigenvalue properties.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: Secs. 4 and 6 specify the architecture and core training setup; App. G gives architecture, preprocessing, and inference details. Code and checkpoint links are public. The Human-AIGI images and psychophysical metadata are pending release with documentation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Training and inference code, setup instructions, and checkpoint links are available at https://github.com/handsome-rich/MIRROR. Human-AIGI images and per-image psychophysical metadata are pending release with documentation.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 6 gives the phase-wise training data and Phase 2 optimizer, learning rate, schedule, epochs, and input size. Sec. 4 and App. G define the architecture, frozen components, token pooling, architecture dimensions, preprocessing, and the inference decision rule.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main benchmark tables report single-run scores. Sec. 6 gives mean and standard deviation across five independent MIRROR training runs. App. D reports participant-bootstrap 95% confidence intervals for human group means.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Sec. 6.3 and Fig. 8 report V100 inference throughput and latency. Per-phase training wall time, peak memory, and total development compute are not fully documented.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [No]

   Justification: Human-AIGI participants volunteered without monetary compensation. The study used publicly released real images and AI-generated images from publicly available models. Participants gave informed consent, could withdraw at any time, and provided no personally identifiable information.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: App. F discusses media-verification benefits, false-positive and false-negative harms, contextual evidence and human review, and risks from adversarial probing.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The public assets provide image-detection code and weights. App. F discusses detector misuse and adversarial probing; benchmark redistribution follows the source assets' terms.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original dataset and pretrained-model publications are cited. DINOv2 uses Apache-2.0, DINOv3 the DINOv3 License, CLIP MIT, and Stable Diffusion v1.4 CreativeML Open RAIL-M. COCO annotations use CC-BY-4.0; image rights and benchmark redistribution follow their individual release terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: Sec. 5 and App. D document the benchmark scope, confidence and latency measurements, and subset criterion. The repository documents code setup, training, and inference. Human-AIGI metadata and distribution documentation are pending release alongside images.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

Justification: Sec. 5 and App. D describe the task, participants, scale, timing, practice and rest controls, informed consent, and voluntary participation without payment. The full participant-facing instructions and consent form are not included here.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: No formal IRB approval was obtained. Participants received a written protocol description, gave informed consent, and could withdraw at any time. The study collected no personally identifiable information.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core pipeline consists of a visual encoder, prototype memory, sparse reconstruction, and comparison-based image classification heads.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.