

Synthetic image detection with CLIP: Understanding and assessing predictive cues[☆]

Marco Willi¹*, Melanie Mathys, Michael Graber

Institute for Data Science I4DS, University of Applied Sciences FHNW, Brugg-Windisch, AG, Switzerland

ARTICLE INFO

Dataset link: <https://github.com/marco-willi/clip-cues>

Keywords:

Synthetic image detection
CLIP
Deepfake detection
Multimedia forensics
Interpretability
Diffusion models
Generalization

ABSTRACT

Recent generative models produce near-photorealistic images, challenging the trustworthiness of photographs. Synthetic image detection (SID) methods, however, often struggle to generalize across datasets and generative models. CLIP, which embeds images and text in a shared semantic space, performs well at SID, but the cues underlying its decisions remain poorly understood. We therefore study CLIP-based SID as an empirical interpretability problem rather than proposing a new detector. We introduce *SynthCLIC*, which pairs real photographs with caption-matched, high-quality diffusion-generated counterparts. We evaluate CLIP-based detectors on *SynthCLIC*, a GAN-heavy benchmark, and a broad external benchmark, and compare them with a low-level forensic CNN, a broad-generator detector, and a text-grounded concept model. CLIP-based linear detectors reach 0.96 mAP on the GAN-heavy benchmark but 0.92 on *SynthCLIC*, while cross-family transfer to *CNNSpot* falls to 0.42 mAP. Within-class associations between detector scores and text-derived cue scores show that higher synthetic scores correspond to cleaner, more compositionally controlled, and technically polished images, whereas lower scores correspond to messier capture conditions and provenance cues characteristic of real photographs. These associations are distributed across many overlapping cues, and their profiles differ strongly across training datasets. CLIP-based and forensic detectors therefore fail in different ways and provide complementary evidence, while broad generator coverage appears important for robust SID.

1. Introduction

The rapid advancement in the quality of image generative models [1–5] has made it increasingly difficult to distinguish between real and synthetic photographs [6–8]. This enables nefarious actors to create convincing material to deceive and manipulate their targets, for example in disinformation campaigns or cyber influence operations [9–11]. Consequently, there is a growing body of research focused on detecting synthetic imagery [12–15] to mitigate such risks.

Synthetic image detection (SID) methods often exploit model-specific *fingerprints* [12] — subtle traces originating from the generative process — to identify synthetic images. Supervised learning algorithms can be trained to recognize these fingerprints with high accuracy [13]. Such approaches, however, suffer from two major limitations: (i) their accuracy is often tightly coupled to the specific generative model used for model training [16], and (ii) they are vulnerable to adversarial attacks that remove or obscure these low-level cues [17]. Consequently, SID methods require constant re-training and have unknown effectiveness in open-set conditions where test-time generative models

might differ from those seen during training [18] (referred to as generalization).

A recent line of work investigates the utility of foundational vision models — trained with self-supervised objectives on internet-scale datasets — for synthetic image detection. Particularly, the CLIP model [19] has been a focus of intensive research [20–27]. CLIP is trained on large-scale datasets of image–text pairs with the objective of maximizing similarity between matching pairs in a joint text–vision embedding space. The semantic richness of CLIP’s visual encoder, combined with its robustness to distribution shifts [19], makes it a promising candidate for detecting synthetic images. Prior results suggest that CLIP-based detectors do not rely on low-level forensic fingerprints [26]—which is plausible given they are trained to match semantic image descriptions.

Naturally, this raises a fundamental question: *Which photographic, semantic, and forensic information represented by CLIP supports SID decisions? Furthermore, how effective are CLIP-based methods considering generative models continue to improve in fidelity and diversity?*

Our goal is to identify which cue families in CLIP space support SID decisions and whether these cues are stable under changes in model

[☆] This article is part of a Special issue entitled: ‘ARRAY_InSeTMA’ published in Array.

* Corresponding author.

E-mail addresses: marco.willi@fhnw.ch (M. Willi), melanie.mathys@fhnw.ch (M. Mathys), michael.graber@fhnw.ch (M. Graber).

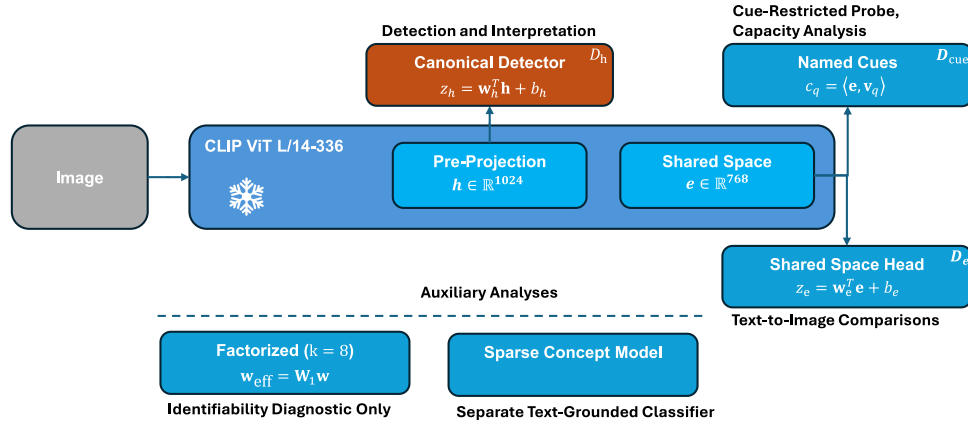


Fig. 1. Architecture overview. CLIP (frozen weights) computes the 1024-d pre-projection visual representation h for the canonical detector D_h . Projection yields the 768-d shared image-text representation e for text-coordinate analyses through the shared-space analysis head D_e . We evaluate a cue-restricted probe D_{cue} on 168 cue scores. The factorized $k=8$ head is a diagnostic attempt to isolate k interpretable projections, and the sparse concept model is a separate text-grounded classifier.

initialization and dataset. We treat SID performance as a prerequisite for interpretation, not as the main objective. Specifically, we address the following questions:

- Q1: How do CLIP-based SID detectors behave across paired diffusion datasets, GAN-heavy benchmarks, and cross-family evaluation settings?
- Q2: Which explanation units for CLIP-based SID are stable and interpretable: individual axes of a factorized linear head, the effective linear decision direction, or sparse text-grounded concepts?
- Q3: Which photographic, perceptual, or forensic cue families in CLIP space distinguish real from synthetic images, and how do these cues change across datasets and generator families?

We seek to answer these questions by systematically analyzing the behavior and capabilities of CLIP-based synthetic image detectors. We leverage CLIP's cross-modal nature to probe its visual representations with textual descriptions (cues), offering an interpretable window into what drives its predictions. Additionally, we evaluate CLIP-based detectors across datasets of varying generative quality to understand their performance boundaries and failure modes.

Our contributions are fourfold:

1. We introduce *SynthCLIC*, a paired dataset of real photographs from the *CLIC* dataset and high-quality synthetic counterparts from recent diffusion models, designed to support content-controlled analysis of CLIP-based SID cues.
2. We evaluate the performance of CLIP-based SID across paired diffusion datasets and a GAN-heavy benchmark. We compare with low-level forensic detectors and a broad-generator out-of-the-box detector for reference.
3. We show that explanations based on individual axes of an over-parameterized k -dimensional linear head are not constant across random initializations. We therefore use single linear decision directions and photographic cues as the primary interpretation units.
4. We analyze CLIP-based SID decisions with CLIP-IQA-style perceptual attributes, paired content-controlled cue shifts, and score-space alignment measured directly against the evaluated checkpoints, and separately evaluate text-grounded concept models as complementary vocabulary-conditioned classifiers rather than as explanations of the linear detector. Fig. 1 summarizes this detector and analysis hierarchy.

2. Previous work

2.1. Generative modeling

Generative models learn the distribution of a dataset and enable sampling of new data points. Often, such models allow for conditional generation, for example, to condition on class labels or textual descriptions. Recent advances in generative modeling have enabled near photo-realistic creation of images. There are different model types, such as generative adversarial networks (GANs) [28], variational autoencoders (VAEs) [29], normalizing flows [30], autoregressive models [31, 32], and diffusion models (DMs) [33]. In particular, DMs have demonstrated impressive generative quality, such as the stable diffusion family of models [1–3]. Furthermore, there are commercial providers which offer access to proprietary (and sometimes open-source) models, such as Midjourney [34], Google's Imagen3 [4], Black Forest's Flux models [35], Adobe's Firefly models [36], or OpenAI's GPT-Image [37]. The most recent advances include multi-modal large language models which can be used to generate images and text consistently and in conditional sequences [38,39], including image captioning.

2.2. Synthetic Image Detection (SID)

Generative models leave invisible traces — also referred to as *artificial fingerprints* — in generated images, which are tied to the generative process [12]. These fingerprints can be visualized by calculating the average frequency spectra of images from generative models [12,16]. This has been demonstrated for different models and architecture types, including diffusion models [40]. A common approach to SID is to model it as a binary classification task using a dataset of real and synthetic images. Several works [13,41–46] demonstrated that this approach can achieve high accuracy. Models trained for SID tend to generalize poorly to synthetic images from models not seen during model training [47]. Generalization improves with aggressive data augmentation during model training, such as post-processing operations like image compression [13]. However, models do not generalize across substantial architectural changes, e.g., from GANs to Diffusion models [18].

A complementary strategy is to broaden training coverage rather than change the feature space. Recent face-forgery research has addressed distribution shift through multi-domain learning and more realistic data generation. GM-DF [48] jointly models heterogeneous forgery datasets and uses CLIP to represent features common to several domains, while Agent4FaceForgery [49] generates ecologically diverse face-forgery data through multi-agent simulation. Although these works address face manipulation rather than fully synthetic images, they

reinforce the importance of heterogeneous training data. More directly related to our setting, CommunityForensics [50] trains on images from thousands of generators and demonstrates strong cross-generator transfer. We therefore include CommunityForensics-384 as a broad-exposure reference detector and evaluate all detectors on its evaluation set (CF-Eval) alongside our own test splits.

Additionally, SID models remain vulnerable to adversarial attacks, as shown in both white-box and black-box scenarios [17,51]. Furthermore, optimal detection thresholds vary for accurate detection of synthetic images from different generative models, making it necessary to calibrate model predictions when doing inference on synthetic images from novel sources.

2.3. CLIP-based Synthetic Image Detection

A growing body of work demonstrates that CLIP’s image representations are remarkably effective for SID [20–27]. A common paradigm is to extract features using CLIP’s frozen image encoder and train a lightweight classifier — typically logistic regression or a linear SVM — on top of those features. Across studies, CLIP-based representations consistently outperform features from supervised models in both accuracy and generalization.

A central insight from this line of work is the effectiveness of contrastive pre-training combined with the vision transformer (ViT) architecture [52], as opposed to CNN-based backbones. CLIP-ViT embeddings not only outperform supervised ImageNet features for SID but also naturally separate real from synthetic images in 2D t-SNE projections, even without fine-tuning [20]. These findings demonstrate the inherent discriminative capacity of CLIP representations for SID.

Another important thread investigates the layer-wise utility of CLIP representations. While most studies use (near) final image embeddings, several works show that intermediate layers contain richer low-level cues crucial for detecting artifacts and visual fingerprints. Fusing intermediate and final-layer features with a learnable module yields substantial gains [25]; intermediate features also outperform early or late layers in open-set and out-of-distribution scenarios, particularly with limited training data [21].

A third theme is robustness and data efficiency. CLIP-based detectors generalize to unseen generative models and remain stable under perturbations such as compression and resizing; they achieve strong performance with as few as 10 training examples, with diminishing returns beyond roughly 10,000 image pairs. Mitigating semantic biases via paired training — generating a synthetic counterpart from a real image’s caption — further improves reliability [26].

Finally, joint adaptation of CLIP’s visual and textual components has been explored. Comparisons of linear probing, fine-tuning, prompt tuning, and adapters identify prompt tuning as most effective in many settings; leveraging the text encoder through systematic input adaptation can further enhance SID, though results vary across datasets [23].

2.4. Why are CLIP-based detectors effective?

CLIP models are pre-trained on massive datasets of image–text pairs—400 million in the original release [19], and over 2 billion in recent variants [53,54]. This scale and diversity enable CLIP to learn rich, generalizable visual features by aligning images with textual descriptions. Models trained from scratch in a supervised setting may focus on the (easy) task of detecting generative fingerprints and thus ignore potential features related to real photographs [20]. This behavior could lead to poor generalization: in the absence of model-specific fingerprints everything gets classified as real. Models not explicitly trained for SID might provide a richer, and better suited feature-space [20].

Previous work [26] showed that CLIP-based detectors retain high performance under image manipulations such as resizing and compression, which often degrade the subtle traces exploited by conventional detectors. They found that supervised models (CNNs trained

from scratch) fail when (i) synthetic images are resized to suppress high-frequency artifacts, or (ii) real images are modified using an autoencoder to embed artificial fingerprints. In contrast, CLIP-based detectors remain effective, suggesting their decisions are based on features beyond low-level image statistics.

A study [27] examined the adversarial robustness of CLIP- and CNN-based SID models. Their Fourier-domain analysis revealed that adversarial perturbations targeting CLIP tend to affect a wider frequency range, particularly lower components, than those targeting CNNs. Consistently, they observed that low-pass filtering disproportionately harms CNN-based detectors. Overall, this is indicative of larger image structures being involved in CLIP-based SID approaches. Although CLIP is not inherently more robust to adversarial attacks, adversarial examples rarely transfer between CLIP and CNN architectures, underscoring fundamental differences in the features they exploit.

2.5. Interpreting CLIP-based models

CLIP embeds images and text in a shared space, enabling zero-shot classification and geometric analyses (e.g., linear directions aligned with textual prompts) [19]. For SID, [24] aligns linear-probe weights with text embeddings, prunes redundant features, and inspects attention heads. This reveals blur-related cues and dataset artifacts as being relevant features for CLIP-based SID; tokens like “blurry”/“blur” correlate with specific image generators, while “deeplearning”/“generative” associate with GANs—likely reflecting captions in CLIP’s pretraining data. Experiments span GAN-based datasets and paired images from several text-conditional diffusion models. Complementary research, unrelated to SID, reports head-level specialization in ViT-based CLIP and introduces *TextSpan*, which constructs a semantic basis for each attention head by matching activations to text embeddings [55]. A related but distinct line of work, concept discovery models [56], links a set of human-interpretable concepts to images by introducing per-example binary concept-selection latents (with a Bernoulli prior) and learning them end-to-end by maximizing an ELBO.

2.6. Relation to prior CLIP-based SID

Prior work has demonstrated the effectiveness of CLIP features for SID, typically by training linear classifiers or shallow heads on top of CLIP’s final image embedding and evaluating on specific benchmarks and generator families. In contrast, we (i) evaluate CLIP-based SID across three datasets including a new paired dataset (*SynthCLIC*) with high-quality diffusion-based images derived from CLIC, (ii) we analyze stable linear decision directions and the associated text-grounded cue profiles, and explicitly test the stability of different explanation units, showing that the effective ($k = 1$) decision direction is more reliable than individual axes of an over-parameterized factorized head, and (iii) adapt sparse linear concept discovery models to CLIP-based SID using photography-oriented vocabularies. This allows us not only to quantify detection performance but also to investigate which visual and photographic attributes are associated with CLIP-based decisions about whether an image is real or synthetic.

3. Method

3.1. Image data

We use three different datasets for our experiments. The first, *CNNSpot*, was introduced by [13] and is used in many subsequent works for SID, including CLIP-based approaches [20]. The dataset consists of a training set based on synthetic images generated solely by the ProGAN method, which was trained to generate 20 different classes from the LSUN [57] dataset. The test set of *CNNSpot* is more diverse and includes images from 11 different generative models, mostly based on GANs. For efficiency reasons and because previous work (e.g. [26])



Fig. 2. Examples from the *SynthBuster+* dataset. Different paired images are shown in each row. Each column depicts a different image source, starting with real photographs from the RAISE-1K dataset [59], followed by synthetic images from the *Synthbuster* dataset [58] and images added by us: Imagen 3 [4], FluxDev and FluxSchnell [35], and Stable Diffusion 3 Medium [3].

Table 1

Datasets and the number of real and synthetic images, including split sizes, used in subsequent experiments. For models trained on all datasets (*combined*) the size is the sum over the individual datasets.

	Train		Validation		Test	
	Real	Synthetic	Real	Synthetic	Real	Synthetic
CNNSpot	1000	1000	1000	1000	54970	52980
SynthBuster+	640	8319	160	2080	200	2600
SynthCLIC	1633	6532	102	408	428	1712

demonstrated low sensitivity to the number of training images, we only use 1k training images, instead of the full dataset with 720k images. Example images can be found in [Appendix A.1](#).

Additionally, we use the *SynthBuster* dataset [58] which is based on real photographs (RAISE-1K) [59] and corresponding, paired synthetic images which were generated with text-conditional generative models to match the content of the real photographs as closely as possible. The dataset includes images from nine different generative latent diffusion models. We augment the dataset with synthetic images from more recent models to reflect the current generative quality and name the dataset *SynthBuster+*. Qualitative examples illustrating the dataset are provided in [Fig. 2](#).

We observe that the real photographs used in *SynthBuster* are not quite recent and lack diversity, with a strong bias toward outdoor and street scenes. Moreover, many real images appear overly bright and of relatively low photographic quality. To mitigate the risk that detectors exploit these biases, we construct an additional dataset, which we call *SynthCLIC*. As a basis, we use real photographs from the 6th Challenge on Learned Image Compression (CLIC).¹ These images include both professionally captured photographs and images taken with mobile phones. We then use the same generative models as those used to extend the *SynthBuster* dataset to generate corresponding synthetic images. [Fig. 3](#) shows some example images.

We create images using Imagen 3 [4], FluxDev and FluxSchnell [35], and Stable Diffusion 3 Medium [3] (see [Appendix A.2](#) for exact model versions used). We aim at preserving the aspect ratio of each corresponding real image. We condition the generation process on descriptions derived from Gemini [60], a commercial visual large language model. [Table 1](#) provides an overview of the datasets, including the split sizes used for model training, validation (a subset of the training set, used for early stopping), and testing. In paired datasets any real image and its synthetic variants are assigned to the same split.



Fig. 3. Examples from the *SynthCLIC* dataset. Different paired images are shown in each row. Each column depicts a different image source, starting with real photographs from the CLIC dataset, followed by synthetic images generated with Imagen 3 [4], FluxDev and FluxSchnell [35], and Stable Diffusion 3 Medium [3].

To characterize and compare the synthetic images used in our experiments, and in particular to characterize potential perceptual differences and dataset biases, we employed a method from Image Quality Assessment (IQA). Such methods aim to quantify the *look* and *feel* of images from a human perspective. We used CLIP-IQA [61], a CLIP-based approach, to score several photographic attributes. The resulting score distributions show limited overlap for some attributes, such as *quality*, *beautiful*, *noisiness*, and *sharpness* (see [Appendix B.1](#)), suggesting that CLIP-based detectors can, to some extent, exploit these cues to distinguish real from synthetic images. At the same time, visual inspection of paired images reveals strong semantic similarity between real and synthetic versions ([Figs. 2](#) and [3](#)).

3.2. Evaluation settings and metrics

We consider three evaluation settings. (i) In-dataset (closed-set) detection: train and test on the same dataset, which measures performance when the detector is matched to both the real-image distribution and the set of generators used in that dataset. (ii) Cross-dataset generalization with overlapping generators: train on one diffusion-based dataset (*SynthBuster+* or *SynthCLIC*) and evaluate on the other. In this case, many underlying diffusion models overlap, but the real-image distributions, prompts, and generation pipelines differ, so the task mainly probes robustness to content and distribution shift rather than to completely unseen generators. (iii) Cross-family generalization: train

¹ <http://www.compression.cc/>.

on the GAN-heavy *CNNSpot* benchmark and evaluate on the diffusion-focused datasets, and vice versa, which combines both a change in real-image distribution and a shift in generator family.

For detector performance, we compute AP per generator using its synthetic images and corresponding real-image set and report the mean over generators (mAP), following common SID practice. Interpretation analyses instead use AUROC for matched comparisons among D_h , D_e , and D_{cue} , and correlations or paired effect sizes for cue-association and paired-shift analyses. Confidence intervals for the paired comparisons D_{cue} versus D_e and D_h versus D_e use a cluster bootstrap over source photographs, retaining all associated synthetic variants within each resampled cluster.

3.3. CLIP representations

We use a pre-trained CLIP model, the ViT-L/14-336 variant,² based on the vision transformer architecture [52]. Let L denote the final transformer layer of the CLIP vision encoder. For an image $\mathbf{x}^{(i)}$, we use its layer-normalized [CLS] representation

$$\mathbf{h}^{(i)} = \mathbf{a}^{L,(i)} \in \mathbb{R}^{1024}. \quad (1)$$

CLIP subsequently maps this representation through its learned image projection

$$\mathbf{P} \in \mathbb{R}^{768 \times 1024} \quad (2)$$

to the normalized shared image–text representation

$$\mathbf{e}^{(i)} = \frac{\mathbf{P}\mathbf{h}^{(i)}}{\|\mathbf{P}\mathbf{h}^{(i)}\|_2} \in \mathbb{R}^{768}. \quad (3)$$

We distinguish the 1024-dimensional pre-projection visual representation $\mathbf{h}$ from the 768-dimensional shared image–text representation $\mathbf{e}$. We do so because the CLIP projection is optimized for cross-modal alignment and may discard visual information relevant to synthetic image detection. Accordingly, our canonical detector operates directly on $\mathbf{h}$, while $\mathbf{e}$ is used for analyses that require direct comparison with text embeddings. All image representations are produced by the pre-trained CLIP encoder, which remains frozen throughout.

3.4. Canonical CLIP detector

The evaluated detector is a linear binary classifier on the pre-projection visual representations,

$$\mathbf{z}^{(i)} = \mathbf{w}^T \mathbf{h}^{(i)} + b, \quad \hat{p}^{(i)} = \sigma(\mathbf{z}^{(i)}), \quad (4)$$

with synthetic-class logit $\mathbf{z}^{(i)}$ and predicted probability $\hat{p}^{(i)}$. Over the N labeled training samples we minimize

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \text{BCE}(\hat{p}^{(i)}, y^{(i)}), \quad (5)$$

where $y^{(i)} \in \{0, 1\}$ is the label of sample i and $y^{(i)} = 1$ denotes synthetic. The weight vector $\mathbf{w}$ defines this detector’s decision boundary. This is the canonical CLIP detector used for all primary CLIP detector-performance results in the manuscript.

We use deterministic pre-processing, weight decay, and early stopping for regularization. We do not use random image augmentation in the primary experiments, so that the learned direction corresponds to a fixed CLIP representation of the evaluated image distribution. Some training splits are substantially imbalanced (Table 1); we do not apply class re-weighting or re-sampling, and therefore emphasize threshold-independent metrics such as mAP. See Appendix A.3 for implementation details.

Table 2

Example vocabulary antonym pairs used for constructing C_{antonyms} . Each pair yields one cue direction via Eq. (6).

Attribute	Positive	Negative	Category
Saturation	Balanced saturation	Oversaturation	Color
Vibrance	High vibrance	Muted palette	Color
Vignetting	Subtle vignetting	Heavy vignetting	Lens
Glitch artifacts	No glitch artifacts	Glitch artifacts	Other

We also experimented with a factorized, two-layer linear head whose activation-level penalty encourages orthogonality over the observed CLIP features (Appendix A.4, Eq. (17)). Without a nonlinearity it spans the same class of linear decision boundaries as Eq. (4). Refits show that its individual axes depend on initialization and are unstable for interpretation (Section 4.2), so we treat it as a diagnostic rather than the canonical detector.

The approach is closely related to prior CLIP-based SID work: [20] used a functionally equivalent single linear layer, and [26] used the same frozen CLIP vision features with an SVM.

3.5. Text-grounded vocabularies

We use two vocabularies to interpret SID with CLIP: (i) *TextSpan* (3498 descriptions, C_{textspan}) from [55]; and (ii) a photography-focused antonym vocabulary (168 attributes, C_{antonyms}) consisting of paired attribute descriptors (e.g., *sharp detail* $\leftrightarrow$ *blurry detail*) with corresponding prompt templates. For each attribute, we embed the positive and negative prompts with CLIP, subtract the normalized negative embedding from the normalized positive embedding, and apply ℓ_2 normalization. Candidate attributes and antonym pairs were generated using ChatGPT with a prompt requesting photographic and perceptual cues (e.g., color, lens effects, technique), while discouraging object semantics. See Table 2 for examples (prompts not shown, only attributes).

We refer to each signed attribute direction $\mathbf{v}_q$ as a *cue*, and to its projection on an image embedding $c_q = \langle \mathbf{e}, \mathbf{v}_q \rangle$ as the corresponding *cue score*. Writing $\mathbf{t}_q^\pm$ for the normalized positive- and negative-prompt embeddings of attribute q , this yields the cue direction

$$\mathbf{v}_q = \frac{\mathbf{t}_q^+ - \mathbf{t}_q^-}{\|\mathbf{t}_q^+ - \mathbf{t}_q^-\|}. \quad (6)$$

Both vocabularies are fixed in advance and not optimized for SID.

3.6. Detector interpretation

We interpret and analyze CLIP detectors using (i) *score association* to show how strongly named cues covary with the detector’s logit, (ii) *shared-space cue similarity* to compare detector decision directions with cue directions in the cross-modal space, and (iii) *paired cue shifts* to analyze cue differences in content-matched real-synthetic pairs. We also report *cue-span capacity* to assess how much predictive information might be lost by restricting the analysis on the cues.

Score association. This uses the canonical detector’s logit directly. Sample i ’s score on cue q is its projection onto the cue direction of Eq. (6),

$$c_q^{(i)} = \langle \mathbf{e}^{(i)}, \mathbf{v}_q \rangle, \quad (7)$$

so positive values point toward the positive prompt pole. Score association is

$$r_q^{\text{within}} = \frac{1}{2} \sum_{y \in \{0,1\}} \text{Corr}_{\text{pearson}}(z^{(i)}, c_q^{(i)} \mid y^{(i)} = y), \quad (8)$$

each correlation over the samples of that class. A positive value means that, averaged over the two subsets, larger cue scores associate with

² <https://huggingface.co/openai/clip-vit-large-patch14-336>.

larger synthetic-class logits; near zero means no consistent within-class covariation. Averaging within-class correlations prevents class-mean separation alone from producing a high association. Additional diagnostics and uncertainty details are reported in [Appendix B.3](#).

Shared-space cue similarity. The canonical detector D_h operates on the pre-projection visual representation which cannot be directly compared to cue embeddings. Therefore, we train a shared-space analysis head D_e which uses the same matched training recipe

$$z_e^{(i)} = \mathbf{w}_e^\top \mathbf{e}^{(i)} + b_e, \quad (9)$$

D_e is not the benchmark detector, it exists only because a 1024-d pre-projection direction cannot be compared directly with 768-d text directions.

Cue-span capacity. Sample i 's cue scores $\mathbf{c}^{(i)} = (c_1^{(i)}, \dots, c_{168}^{(i)})$ feed a classifier restricted to the vocabulary (D_{cue}),

$$z_{\text{cue}}^{(i)} = \mathbf{u}^\top \mathbf{c}^{(i)} + b_{\text{cue}}, \quad (10)$$

with coefficient vector $\mathbf{u}$ and intercept b_{cue} . Its AUROC measures the label-predictive information accessible from the named cue scores, not agreement with the detector. The cue-restricted probe D_{cue} uses the same recipe on the 168 cue scores themselves. Taken together, $D_h \rightarrow D_e \rightarrow D_{\text{cue}}$ form a nested restriction: the first is the detector, the second asks how much remains after projection into CLIP's shared space, and the third asks how much information remains under the vocabulary restriction.

Paired cue shifts. The paired datasets let us compare each real image $\mathbf{x}_r^{(i)}$ with synthetic images $\mathbf{x}_s^{(i,g)}$, g indexing the generator, from the same source photograph or caption. These give descriptive within-pair changes in cue scores while approximately controlling content (see [Appendix B.2](#) for details).

Cue results depend on the vocabulary and how much of the representation it covers. We read them as descriptive profiles rather than causal explanations, emphasizing recurring cue families over single directions.

[Fig. 1](#) summarizes the detector and analysis hierarchy. The same optimizer, loss, and schedule are used for all three heads.

3.7. Concept modeling

Our concept model follows the Sparse Linear Concept Discovery Models (CDMs) [56], which use CLIP image–text similarities and a variational Bernoulli prior to obtain sparse linear concept activations. We adopt this framework for binary SID on *CNNSpot*, *SynthBuster+*, and *SynthCLIC* using the two fixed vocabularies of Section 3.5.

Let $\mathbf{C} \in \mathbb{R}^{n \times p}$ be the row-normalized concept embeddings with n terms, $\mathbf{I} \in \mathbb{R}^{m \times p}$ the row-normalized image embeddings for m samples, and $\mathbf{S} = \mathbf{I}\mathbf{C}^\top \in \mathbb{R}^{m \times n}$ their cosine similarities. Following [56], we use a binary relevance mask $\mathbf{Z} \in \{0, 1\}^{m \times n}$ to flexibly select concepts per sample. $\mathbf{W}_c \in \mathbb{R}^{n \times 1}$ contains the classifier weights that map the gated concept similarities to the synthetic-class logit; the per-sample selection is performed by $\mathbf{Z}$. The class logits $\mathbf{a} \in \mathbb{R}^{m \times 1}$ are thus:

$$\mathbf{a} = (\mathbf{S} \odot \mathbf{Z}) \mathbf{W}_c \quad (11)$$

During training, we treat the relevance mask $\mathbf{Z}$ as a latent variable and infer it with a factorized Bernoulli variational posterior. For each sample i (row $\mathbf{z}_i$ of $\mathbf{Z}$), we set

$$q(\mathbf{z}_i) = \text{Bernoulli}(\sigma(\mathbf{I}_i \mathbf{W}_s^\top)), \quad (12)$$

$$p(\mathbf{z}_i) = \text{Bernoulli}(\alpha), \quad (13)$$

where $\mathbf{W}_s \in \mathbb{R}^{n \times p}$ is a learned projection from image-embedding space to per-concept logits (with p the embedding dimension), $\sigma(\cdot)$ is applied element-wise, and the scalar α is fixed across concepts to encourage sparsity. We optimize the negative ELBO:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{y}, \mathbf{z})} [\text{BCEWithLogits}(\mathbf{a}, \mathbf{y}) + \beta \text{KL}(q(\mathbf{z}) \parallel p(\mathbf{z}))] \quad (14)$$

where the first term is the binary cross-entropy applied to the logits, KL regularizes the inferred masks toward the sparse prior, and β controls the sparsity–accuracy trade-off.

We use the concept model as a complementary, intrinsically text-grounded SID classifier rather than as an explanation of the canonical linear detector: it is trained directly on class labels and is not fitted to reproduce the linear classifier's scores. Its concept contributions show how this model uses the supplied vocabulary to predict the SID label. We therefore evaluate both its detection performance and its sensitivity to the vocabulary and the sparsity weight β , and we report how stable its concept selection is across random seeds ([Appendix B.5.2](#)).

4. Experiments and results

We first characterize detection performance (Section 4.1) and establish which explanation unit is stable enough to interpret (Section 4.2). The remaining subsections then analyze what the evaluated detectors' cues are, how they differ between real and synthetic images, why cross-family generalization fails, and how the separately trained sparse concept model behaves.

4.1. Detection performance under distribution shift

We report and compare detection performance of our CLIP-based detectors with a CNN-based forensic detector, trained using the protocol described in [13]. Additionally, we use a more recent detector (CommunityForensics-384 [50]), trained on approximately 5.4M images: 2.7M synthetic images generated by 4803 models and a corresponding set of 2.7M real images. We evaluate the models on the test splits shown in [Table 1](#) and *CF-Eval* from [50]. Results are shown in [Table 3](#).

Low-level forensic CNNs are highly effective when training and test distributions match (e.g. 0.98 mAP in-domain on *SynthCLIC*), but their performance drops sharply under dataset and generator-family shift. The picture is similar for the CLIP linear detector: strong in-distribution results (0.96 mAP on *CNNSpot*, 0.92 on *SynthCLIC*) but weak transfer, falling to 0.42 mAP from *SynthCLIC* to *CNNSpot*. Training on the combined dataset improves performance on *CF-Eval* substantially (0.93 mAP). The broad-generator CommunityForensics-384 detector transfers well across *SynthCLIC*, *SynthBuster+*, and *CNNSpot*, indicating that diverse generator exposure improves generalization. Notably, its performance on *SynthCLIC* (0.87) is markedly lower than on *CF-Eval* (0.99), suggesting *SynthCLIC* is a comparatively difficult distribution even for a broadly trained detector. These results do not position our CLIP linear classifier as a state-of-the-art detector. Rather, the three families fail in different places: on *CF-Eval* the *SynthCLIC*-trained CLIP head transfers better overall and on latent-diffusion and commercial generators, whereas the ProGAN-trained forensic CNN is stronger on GANs, and the broad-exposure detector is strongest everywhere (Section 4.6). The canonical detector and the factorized $k = 8$ head achieve similar detection performance in a matched cached-feature comparison, and the same qualitative pattern holds across the three OpenAI CLIP ViT backbones we evaluate: ViT-L/14-336, ViT-B/16, and ViT-B/32 achieve comparable in-domain mAP on *SynthCLIC* ([Appendix B.6.2](#)). Full cross-dataset and per-generator results are reported in [Appendix B.6](#).

4.2. Interpretation target and stability

Before interpreting learned directions, we first assess whether they are stable across refits. We compare decision directions using the covariance-weighted cosine $\cos_\Sigma$ (Eq. (19)) and their 168-dimensional cue-association profiles using Spearman correlation ρ ([Appendix B.3](#)).

Across five seeds, the canonical detector D_h is highly stable: it reaches AUROC 0.921 on the *SynthCLIC* test split, with $\cos_\Sigma = 0.989$ between refit directions and $\rho = 0.991$ between their cue-association

Table 3

Detectors trained on different datasets are evaluated on *SynthCLIC*, *SynthBuster+*, *CNNSpot*, and the CommunityForensics-Eval (CF-Eval) dataset. Shown is mean AP across the generators in each test set.

Detector	Train	SynthCLIC	SynthBuster+	CNNSpot	CF-Eval
CLIP (linear head)	SynthCLIC	0.92	0.79	0.42	0.72
	SynthBuster+	0.61	1.00	0.67	0.79
	CNNSpot	0.52	0.50	0.96	0.66
	Combined	0.87	0.98	0.88	0.93
Forensic CNN	SynthCLIC	0.98	0.80	0.49	0.58
	SynthBuster+	0.76	0.99	0.52	0.73
	CNNSpot	0.49	0.52	0.94	0.53
	Combined	0.98	1.00	0.76	0.80
CommunityForensics-384	Broad (off-the-shelf)	0.87	0.91	0.97	0.99

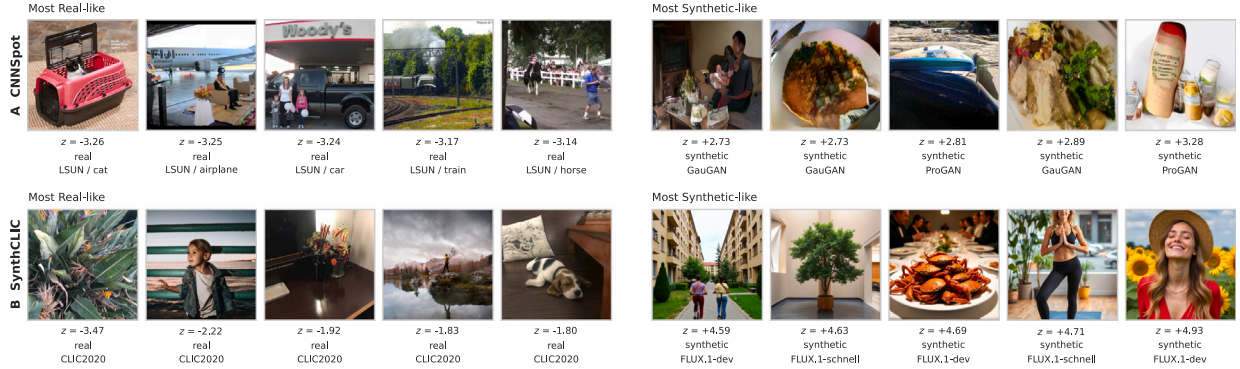


Fig. 4. Test-set images with the five lowest and five highest canonical-detector logits z_h for *CNNSpot* (A) and *SynthCLIC* (B). Beneath each image the logit, the class and the image's source are shown.

profiles. A matched factorized $k = 8$ head achieves similar detection performance (AUROC 0.915 versus 0.921) and its effective direction is likewise stable ($\cos_{\Sigma} = 0.963$, $\rho = 0.979$). In contrast, the individual factorized axes are unstable: after Hungarian matching between refits, their mean absolute cosine is 0.289 and their cue-profile correlation is $\rho = 0.119$. We therefore interpret the effective decision boundary rather than individual axes of the factorization (Appendix A.5).

Fig. 4 illustrates the images ranked most strongly along the canonical decision direction. The extremes differ substantially between datasets: the most synthetic-scoring *CNNSpot* images are low-resolution GAN images with visible artifacts, whereas the corresponding *SynthCLIC* images are high-quality diffusion-generated photographs. Because the exact extreme samples vary more across refits than the decision direction itself, the montage is intended as a qualitative illustration rather than a stable characterization of individual images.

4.3. How much predictive information is captured by named cues?

We next assess how much SID-relevant predictive information the fixed photography-oriented antonym vocabulary captures, by following the nested restriction $D_h \rightarrow D_e \rightarrow D_{\text{cue}}$ on the *SynthCLIC* test split. We report AUROC here because it compares the ranking of matched heads on the same imbalanced split independently of class prevalence. Table 4 reports each step as a signed change relative to the row above. Projecting into CLIP's shared image-text space, where textual cues are defined, moves AUROC from 0.921 for the canonical detector D_h to 0.893 for the shared-space analysis head D_e , a change of -0.029 (95% CI $[-0.037, -0.021]$). Projecting to a cue-restricted space with D_{cue} gives an AUROC of 0.834, a further -0.059 (95% CI $[-0.075, -0.043]$). The named cue span therefore carries a substantial but incomplete portion of the linearly accessible SID signal, and most of what is lost to the vocabulary rather than to the projection. This measures what the named vocabulary *can predict*, not what the detector uses.

Table 4

Successive information restriction on the *SynthCLIC* test split (mean over 5 seeds). Each Δ is the change in AUROC relative to the row above, with a 95% cluster bootstrap confidence interval (2000 resamples, clustered on source photo); negative values are losses. Projection into the shared image-text space costs little, restriction to the named cue span costs roughly twice as much.

Stage	AUROC	Δ vs. previous row [95% CI]
D_h (full pooler, 1024-d)	0.921	–
D_e (shared space, 768-d)	0.893	-0.029 $[-0.037, -0.021]$
D_{cue} (168 named cues)	0.834	-0.059 $[-0.075, -0.043]$

4.4. Which cues covary with detector scores?

We correlate the evaluated detector's score with each textual cue score. The correlation is computed separately within real and synthetic images and then averaged, so Panel A of Fig. 6 identifies cues that track detector confidence among images of the same class. It does not identify the largest differences between real and synthetic images, and it does not establish that the detector causally uses a cue. Appendix B.3 reports secondary diagnostics and uncertainty details.

Higher synthetic-class scores covary with clean framing ($|r_q^{\text{within}}| = 0.333$), clean low-light rendering (0.306), high clarity (0.257) and smooth long-exposure motion (0.235). Lower scores covary with capture-provenance cues: an instant-film texture (0.302), a documentary look (0.216) and legible text (0.200). The profile is distributed across many overlapping cue directions rather than carried by any one of them, so no individual cue is a complete description of the score.

4.5. Which cues change between real and synthetic images?

The paired design lets us measure cue changes while approximately controlling image content. Panel B of Fig. 6 reports paired effect sizes over the 1712 *SynthCLIC* test pairs (Appendix B.2). Synthetic variants shift toward streaked-background panning ($d_q = +1.03$),

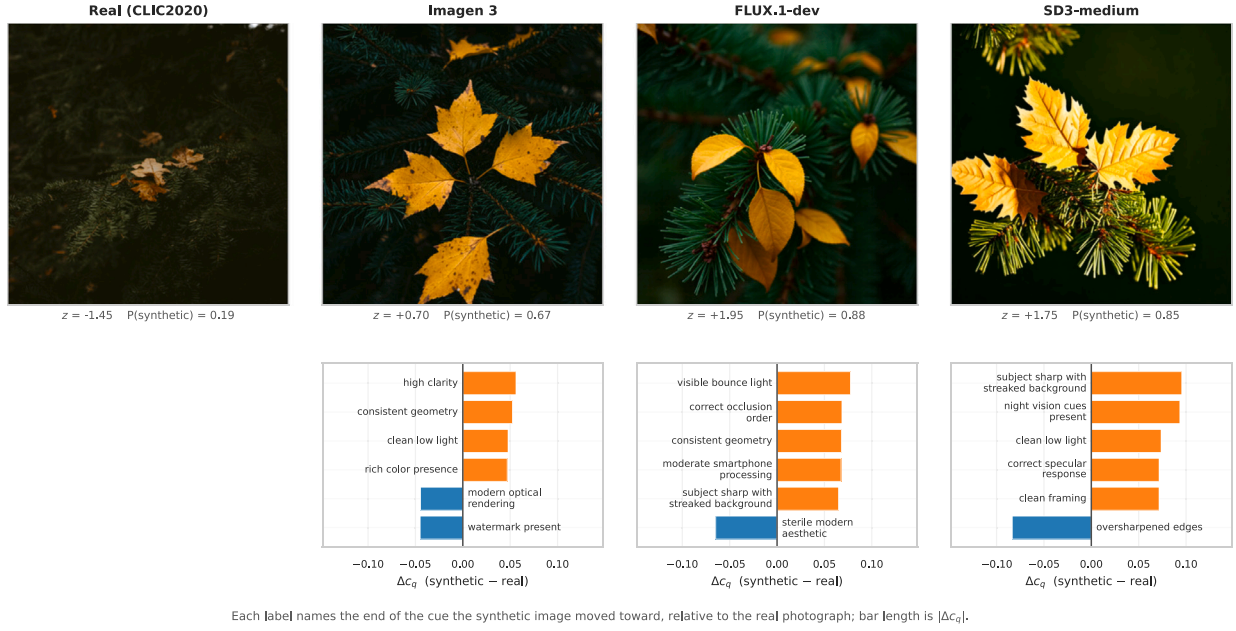


Fig. 5. One CLIC photograph and three caption-conditioned synthetic counterparts. Bars show the largest cue-score changes $\Delta c_q = c_q(x_{\text{synth}}) - c_q(x_{\text{real}})$ relative to the real reference. Positive and negative bars indicate movement toward the corresponding named cue pole. Canonical-detector logits are shown below each image. FLUX.1-schnell not shown for legibility.

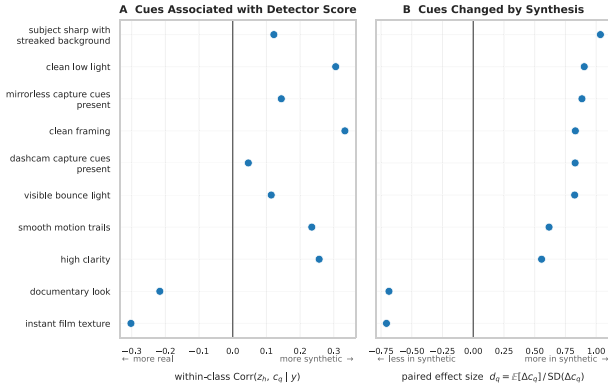


Fig. 6. Population-level cue interpretation on the *SynthCLIC* test split. (A) Within-class association between each named cue and the canonical detector's logit. (B) Paired effect size of synthesis, $d_q = \mathbb{E}[\Delta c_q]/\text{SD}(\Delta c_q)$, where Δc_q is the change in cue score between a synthetic image and its own real reference. Rows carry each cue's positive phrase and share a y-axis, and each panel's poles are annotated, so a row reads across both panels.

cleaner low-light rendering (+0.90) and framing (+0.83), and away from instant-film texture (−0.71) and a documentary look (−0.69). Together, these distributed changes indicate a shift toward a more polished, camera-ideal photographic appearance rather than a single dominant artifact.

Fig. 5 illustrates these cue shifts for one real photograph and its caption-conditioned synthetic counterparts.

4.6. Why does cross-family generalization fail?

The weakest transfer occurs when the *SynthCLIC*-trained detector is applied to GAN-heavy data. On CF-Eval, its family-level mean AP is lowest for GAN (0.36) and pixel-space diffusion (0.41) generators, compared with 0.87 for latent diffusion and 0.73 for commercial generators (see Table 5). The two weakest families contain only two and one

Table 5

Per-architecture mean AP on CommunityForensics-Eval. Each value is the unweighted mean of per-generator AP within the family; mAP is the unweighted mean across all 21 generators. Family sizes are strongly unbalanced: Commercial 11, LatDiff 6, GAN 2, PixDiff 1, Other 1.

Detector	mAP	GAN	LatDiff	PixDiff	Commercial	Other
CLIP linear probe (SynthCLIC)	0.72	0.36	0.87	0.41	0.73	0.98
Forensic CNN (SynthCLIC)	0.58	0.43	0.64	0.50	0.58	0.73
Forensic CNN (ProGAN/CNNSpot)	0.53	0.80	0.39	0.59	0.57	0.44
CommunityForensics-384 (out-of-the-box)	0.99	1.00	1.00	0.93	0.99	1.00

generators, respectively, but the pattern is consistent with the detector's low mAP on *CNNSpot* (0.42).

The learned boundaries provide a geometric explanation. The *SynthCLIC*- and *CNNSpot*-trained shared-space heads have a covariance-weighted cosine of $\cos_{\Sigma} = -0.207$. Their signed difference decomposes into processing- and artifact-related cues on the *CNNSpot* side and photographic, optical, and provenance-related cues on the *SynthCLIC* side (see Fig. 7). Thus, the two training distributions induce substantially different decision rules rather than different strengths of one common synthetic-image signature.

Score association gives a consistent result: on *CNNSpot*, the fixed *SynthCLIC* detector retains a moderately similar cue profile (Spearman $\rho = 0.638$) even though its label ranking reverses (AUROC 0.094), while the independently *CNNSpot*- and *SynthCLIC*-trained cue profiles are essentially unrelated ($\rho = 0.008$). Cross-family failure therefore reflects distribution-specific decision boundaries rather than uniformly weak CLIP features, motivating broad generator exposure and complementary forensic cues.

4.7. The sparse concept model

The sparse concept model is an independently trained, text-grounded SID classifier rather than an approximation of the linear

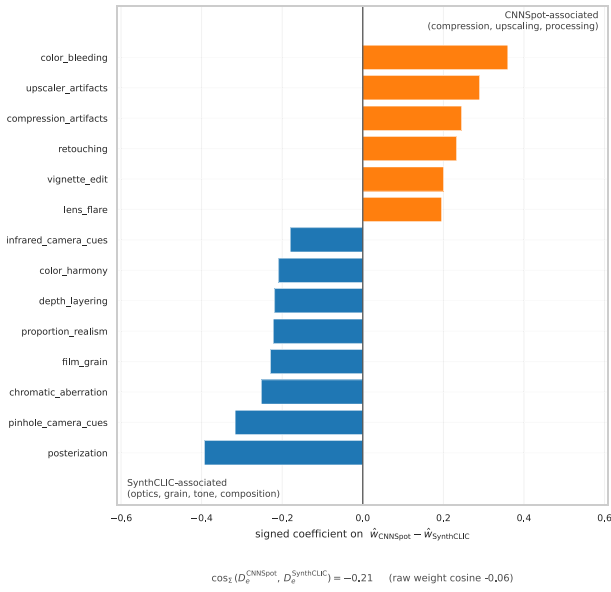


Fig. 7. Signed decomposition of the difference between the *CNNSpot*- and *SynthCLIC*-trained shared-space decision directions onto the named cue axes (14 largest coefficients; all selected across bootstrap refits). Positive values indicate relatively stronger synthetic-class weighting by the *CNNSpot* boundary. Processing and compression cues dominate the *CNNSpot* side, whereas photographic, optical, and composition-related cues dominate the *SynthCLIC* side.

Table 6

mAP of sparse concept models across training sets, test sets, and vocabularies.

Test set	Vocabulary	Training set			
		CNNSpot	SynthBuster+	SynthCLIC	Combined
CNNSpot	Antonyms	0.96	0.62	0.47	0.90
	TextSpan	0.96	0.64	0.40	0.89
SynthBuster+	Antonyms	0.53	0.99	0.79	0.99
	TextSpan	0.51	0.99	0.78	0.99
SynthCLIC	Antonyms	0.54	0.48	0.85	0.87
	TextSpan	0.53	0.53	0.88	0.88

detector. It therefore asks how well the supplied vocabulary supports detection on its own.

Table 6 shows strong in-dataset performance but substantially weaker cross-dataset transfer. Compared with the linear CLIP detector (Table 3), performance is competitive overall, with the largest drop on *SynthCLIC*; neither vocabulary is consistently superior.

Fig. 8 shows representative concept profiles for two trained models (*SynthCLIC* and *CNNSpot*). The highest-ranked concepts vary across random seeds (Appendix B.5.2), so we interpret recurring cue families rather than individual terms. On *CNNSpot*, the strongest concepts are synthetic-side artifact cues such as *moire patterns* and *tiling artifacts*. On *SynthCLIC*, the top concepts are less discriminative and often correspond to capture artifacts that occur more frequently in real images, such as *glitch artifacts*, *scan artifacts*, and *print texture*. This suggests a weaker and more dataset-specific basis for classification than on *CNNSpot*.

Increasing the sparsity weight β reduces the mean number of active concepts from about 73 to nearly zero, while mAP varies only moderately and peaks at $\beta = 10^{-4}$; the useful range is approximately 10^{-4} – 3×10^{-4} (Appendix B.5.1). At stronger regularization, larger remaining weights partly compensate for fewer active gates, so active-concept count is only an approximate measure of explanation complexity.

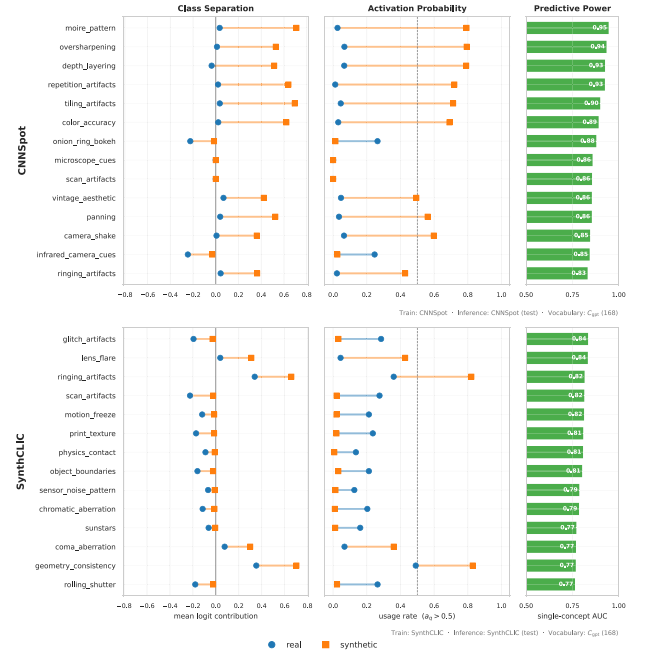


Fig. 8. Representative concept profiles for models trained and evaluated on *CNNSpot* (top) and *SynthCLIC* (bottom). Shown are the 14 concepts with highest single-concept AUROC, together with class-wise mean logit contribution, usage rate (gate probability > 0.5), and single-concept AUROC. Concepts are from the antonym vocabulary C_{antonyms} .

5. Discussion & conclusion

We revisit the three research questions formulated in Section 1.

Q1: How do CLIP-based SID detectors behave across paired diffusion datasets, GAN-heavy benchmarks, and cross-family evaluation settings? Our experiments show that CLIP-based detectors perform well when evaluated in-distribution. Cross-dataset evaluations reveal that detectors trained on *CNNSpot* generalize poorly to *SynthBuster+* and *SynthCLIC*, and vice versa, while pooled training on multiple datasets improves but does not fully close the gap. On the external CF-Eval set the same pattern resolves by generator family: the *SynthCLIC*-trained detector remains strong on latent-diffusion and commercial generators but is weak on the small GAN and pixel-space diffusion subsets, while a broad-exposure detector transfers across all families. These results indicate that a frozen CLIP representation does not by itself yield a universally generalizing SID detector: the discriminative structure in CLIP space depends strongly on both the real-image distribution and the family of generative models seen during training.

Q2: Which explanation units for CLIP-based SID are stable and interpretable: the canonical pre-projection detector, the matched shared-space analysis head, the cue-restricted probe, or sparse text-grounded concepts? The actual canonical detector D_h is stable: its boundary and cue-association profile are reproducible across refits, while the factorized $k = 8$ head has a stable effective direction but unstable individual axes. The shared-space analysis head D_e carries a small but measurable cost relative to D_h (+0.029 AUROC, 95% CI [+0.021, +0.037]), and the cue-restricted probe D_{cue} loses another 0.059 AUROC relative to D_e , so the information available in the named vocabulary is substantial but incomplete. Sparse concept selections are also unstable in identity across seeds and vocabularies, so the exact concept list should not be treated as a unique decomposition of the detector.

Q3: Which photographic, perceptual, and forensic cue families in CLIP space distinguish real from synthetic images, and how do these cues change across datasets and generator families? Our

analyses suggest that CLIP-based SID decisions are associated with a distributed mixture of photographic, processing, and perceptual cues, and that the relevant cue profile is distribution-specific rather than a universal synthetic-image signature. Measured directly against the evaluated detector, higher synthetic scores track composition, craft and technical execution, whereas lower scores track the messiness of real capture. The signed cross-dataset decomposition shows that diffusion-era boundaries emphasize photographic, optical, and grain-related cues, whereas GAN-heavy boundaries emphasize processing and artifact cues. The two families are therefore separated by more than just a single low-level fingerprint. Content-controlled comparisons shift toward a more polished, camera-ideal photographic appearance, including cleaner framing and low-light rendering and weaker documentary/instant-film cues. CLIP-IQA attributes show systematic shifts in perceived *quality*, *noisiness*, and *sharpness*—perceptual attributes that are not reducible to a single low-level forensic statistic. On older GAN-based data, these cues often coincide with visible artifacts and compositional issues, whereas on high-quality diffusion images they manifest as more subtle stylistic and aesthetic differences. The exact sparse concept identities remain unstable, but the aggregate evidence is robust: CLIP-based SID is associated with distributed cue families whose relative emphasis shifts between generator families and datasets.

Limitations and future work. Our study is limited to three training datasets and one broader external evaluation benchmark, all dominated by photographic imagery, and to OpenAI CLIP variants sharing the same broad pre-training approach (ViT-L/14-336, ViT-B/16, ViT-B/32; Appendix B.6.2). The cross-family decomposition is informative but not definitive, especially for the GAN and pixel-space diffusion families, which are represented by only two and one generators in CF-Eval. The canonical detector (D_h) exceeds the shared-space analysis head (D_e) by 0.029 AUROC, and the cue-restricted probe D_{cue} is 0.059 below D_e , so the named vocabulary remains incomplete. Sparse concept identities are stable only at the aggregate level, not in the exact concept lists selected. We therefore present the conclusions as distribution-specific and model-family-dependent rather than as a universal explanation of synthetic imagery. As a result, our conclusions may not fully transfer to non-photographic domains (e.g., medical or scientific imaging) or to future generators with qualitatively different failure modes. Our analysis focuses on photographic cues and does not explicitly combine them with low-level fingerprint-based features. Future work could explore hybrid detectors that jointly exploit semantic CLIP embeddings and forensic cues, extend our analysis to additional backbones and modalities, and investigate continual or online adaptation as new generators appear.

In summary, our results show that CLIP-based methods are powerful but not universal tools for synthetic image detection. They perform strongly on the older GAN-heavy *CNNSpot* benchmark, struggle more with high-quality diffusion images and cross-dataset generalization, and their decisions reflect a distributed mixture of photographic and perceptual cues — not reducible to a single stable generator fingerprint — whose relative importance shifts across datasets, making CLIP-based and low-level forensic detectors complementary. Text-grounded concept models provide useful human-readable summaries, but their interpretation depends on the vocabulary, the sparsity regularization, and the random initialization, and should not be treated as a unique or controllable decomposition of the detector. We hope that our datasets, models, and analyses provide a useful basis for developing more robust and interpretable SID systems that can keep pace with rapidly evolving generative models.

CRedit authorship contribution statement

Marco Willi: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Melanie Mathys:** Writing – review & editing, Software, Methodology, Conceptualization. **Michael Graber:** Writing – review & editing, Supervision, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT to edit the manuscript and used Claude Code to create code for visualizing the results. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Methodological details

A.1. Datasets - *CNNSpot* samples

Fig. 9 shows a few example images from the *CNNSpot* test set.

A.2. Datasets—generative models used

For the datasets *SynthBuster+* and *SynthCLIC*, we generated synthetic images and captions using the following model versions: Imagen3 [4], ³ FluxDev, ⁴ FluxSchnell, ⁵ Stable Diffusion 3 Medium, ⁶ and Gemini⁷ (for caption generation).

A.3. CLIP classifier—implementation details

Table 7 summarizes the canonical detector and the factorized $k = 8$ configuration. The stability analysis instead refits both heads on identical cached, unaugmented features so that they differ only in factorization. When using the shared embedding space $\mathbf{e}$, we rescale it and train on $\tilde{\mathbf{e}}^{(i)}$. A single global scale factor equalizes the representation scales of $\mathbf{h}^{(i)}$ and $\mathbf{e}^{(i)}$, so that the fixed ℓ_2 weight decay of the shared recipe acts as the same amount of regularization in both spaces while the angular geometry is left unchanged.

A.4. Factorized orthogonal head

As an auxiliary interpretability experiment, we consider a factorized linear head with $k = 8$ dimensions. For a pre-projection visual representation $\mathbf{h}^{(i)} \in \mathbb{R}^{1024}$, the logit is

$$z^{(i)} = \mathbf{h}^{(i)\top} \mathbf{W}_1 \mathbf{w}_2 + b, \quad \mathbf{W}_1 \in \mathbb{R}^{1024 \times k}, \quad \mathbf{w}_2 \in \mathbb{R}^k. \quad (15)$$

Because no non-linearity is applied, the head is equivalent to a single linear classifier with effective direction

$$\mathbf{w}_{\text{eff}} = \mathbf{W}_1 \mathbf{w}_2. \quad (16)$$

To encourage distinct intermediate activation dimensions, let $\mathbf{A} = \mathbf{H} \mathbf{W}_1 \in \mathbb{R}^{n \times k}$ denote the mini-batch activations and let $\bar{\mathbf{A}}$ denote $\mathbf{A}$ after ℓ_2 -normalizing each column. We optimize

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda \left\| \mathbf{I} - \bar{\mathbf{A}}^\top \bar{\mathbf{A}} \right\|_F^2, \quad \lambda = 0.33. \quad (17)$$

The regularizer therefore acts directly on the activation dimensions used for interpretation.

³ *imagen-3.0-generate-002*.

⁴ <https://huggingface.co/black-forest-labs/FLUX.1-dev>.

⁵ <https://huggingface.co/black-forest-labs/FLUX.1-schnell>.

⁶ <https://huggingface.co/stabilityai/stable-diffusion-3-medium>.

⁷ *gemini-2.0-flash-001*.



Fig. 9. Example images from the *CNNSpot* test set, showing real and synthetic images from representative generator groups. Real images originate from the benchmark-specific source datasets indicated below each panel. Generator labels refer to the corresponding synthetic subset. Examples are selected near the median canonical-detector score within each group.

Table 7

Implementation details for CLIP classifier.

Category	Parameter	Value	Notes
Training	Max epochs	200	With early stopping
	Batch size	64	
	Early stopping patience	5 epochs	Based on validation loss
	Validation frequency	Every epoch	
Optimizer	Accelerator	gpu (single)	
	Type	Adam	torch.optim.Adam
	Learning Rate	1e-3 (default)	
	Weight decay	0.01	
Model architecture	Beta parameters	(0.9, 0.999)	Defaults
	Epsilon	1e-8	Default
	Feature extractor	clip-large-patch14	openai/clip-vit-large-patch14-336
	Feature extraction layer	Pooler output	Layer-normalized final-layer [CLS], $d = 1024$
Loss function	Frozen backbone	Yes	No gradient updates to CLIP
	Primary loss	Binary cross-entropy with logits	
	Label smoothing	$y = y * (1 - \epsilon) + (1 - y) * \epsilon$	$\epsilon = 0.1$
<i>Canonical linear classifier ($k = 1$; all main-text CLIP detector-performance results)</i>			
Data Aug. (Train)	Augmentations	None	Deterministic pre-processing only (Section 3.4)
Model architecture	Classification head	single linear layer	$\mathbf{w} \in \mathbb{R}^d$, no hidden layer
<i>Factorized diagnostic head ($k = 8$; Appendix A.4 only)</i>			
Data Aug. (Train)	RandomResizedCrop	scale = (0.5, 1.0), size = 512	Crop 50%–100% of original, resize to 512 × 512 (applied before CLIP pre-processing; the final input is resized to the model's resolution)
Training	RandomHorizontalFlip	p = 0.5	Uniform sampling
	RandomJPEGCompression	quality = (65, 100)	Mixed-precision; applies only to the factorized-head runs, which re-encode images. The canonical classifier is fitted on cached embeddings in fp32
	Precision	bf16-mixed	
Model architecture	Classification head	Orthogonal activation head	
	Intermediate dimensions	8	Single intermediate projection; no nonlinearity
Regularization	Orthogonal initialization	Yes	Applied to weights
	Orthogonality loss weight	0.33	Auxiliary orthogonality loss, Eq. (17)

An alternative is to enforce weight orthogonality through $\|\mathbf{I} - \mathbf{W}_1^\top \mathbf{W}_1\|_F^2$. However, orthogonal weights do not generally imply orthogonal activations because

$$\mathbf{A}^\top \mathbf{A} = \mathbf{W}_1^\top \mathbf{H}^\top \mathbf{H} \mathbf{W}_1, \quad (18)$$

which also depends on the empirical feature geometry. We therefore use activation orthogonality when the intermediate dimensions themselves are the object of analysis. Appendix B.6.1 compares both regularizers empirically.

A.5. Interpretation stability

We assess the stability of learned linear directions across five independent refits. Because raw parameter-space cosine can be misleading when feature dimensions have unequal variance and covariance, we use the covariance-weighted cosine

$$\cos_{\Sigma}(\mathbf{w}_1, \mathbf{w}_2) = \frac{\mathbf{w}_1^\top \Sigma \mathbf{w}_2}{\sqrt{\mathbf{w}_1^\top \Sigma \mathbf{w}_1} \sqrt{\mathbf{w}_2^\top \Sigma \mathbf{w}_2}}, \quad (19)$$

Table 8

Stability across five independent refits on *SynthCLIC*. Direction similarity is measured with the covariance-weighted cosine $\cos_{\Sigma}$; cue-profile similarity is measured with Spearman correlation ρ . For the factorized $k = 8$ head, the effective direction is $\mathbf{w}_{\text{eff}} = \mathbf{W}_1 \mathbf{w}_2$. Individual axes are compared after Hungarian matching.

Explanation unit	AUROC	$\cos_{\Sigma}$	Cue-profile ρ
Canonical detector D_h	0.921	0.989	0.991
Factorized $k = 8$: effective direction	0.915	0.963	0.979
Factorized $k = 8$: individual axes	–	0.289	0.119

Table 9

Hyperparameters for Concept Bottleneck Model Experiments.

Category	Parameter	Value
Training	Max epochs	4000
	Batch size	256
	Early stopping patience	10 epochs
	Validation frequency	Every 40 epochs
	Random seed	123
	Early stopping metric	val/loss
Optimizer	Type	AdamW
	Learning rate	1×10^{-3}
	Weight decay	1×10^{-4}
Model	Temperature (τ)	0.1
	Sparsity weight (β)	1×10^{-4}
	Target sparsity (α)	1×10^{-4}
Vocabularies	C	$C_{\text{antonyms}}, C_{\text{textspan}}$

where Σ is the empirical feature covariance. This quantity measures agreement between the centered linear scores induced by two directions on the observed feature distribution.

For the canonical detector D_h and the factorized $k = 8$ head, Σ is computed from the *SynthCLIC* pre-projection visual representations. Cue-profile stability is measured by the Spearman correlation ρ between the corresponding 168-dimensional cue-association profiles. Individual $k = 8$ axes are matched between refits with the Hungarian algorithm using absolute cosine similarity to account for permutation and sign ambiguity (see Table 8).

The effective decision boundary is therefore reproducible across refits, whereas the individual axes of the $k = 8$ factorization are not.

A.6. Concept model—implementation details

Table 9 shows the implementation details of the concept model (see Section 3.7).

Appendix B. Additional results

B.1. Characterizing datasets with CLIP-IQA

We use CLIP-IQA [61] to compare perceptual image attribute scores across the datasets and image sources. Fig. 10 shows the resulting score distributions.

B.2. Content-controlled paired analysis

The paired analysis of Section 4.5 exploits the structure of *SynthCLIC*, where each real photograph has synthetic counterparts generated from its caption. For cue direction $\mathbf{v}_q$ and shared-space representation $\mathbf{e}$, we define the cue shift for source photograph i and generator g as

$$\Delta c_{igq} = \langle \mathbf{e}_{\text{syn}}^{(i,g)}, \mathbf{v}_q \rangle - \langle \mathbf{e}_{\text{real}}^{(i)}, \mathbf{v}_q \rangle, \quad (20)$$

where positive values indicate movement toward the positive pole of the attribute. We summarize each cue by the standardized paired effect

$$d_q = \frac{\text{mean}_{ig}(\Delta c_{igq})}{\text{sd}_{ig}(\Delta c_{igq})}. \quad (21)$$

The analysis uses the *SynthCLIC* test split, comprising 428 source photographs and four synthetic counterparts per photograph (1712 pairs). Fig. 6 pools all pairs when estimating d_q . Uncertainty in the mean cue shifts is estimated with a cluster bootstrap over source photographs (2000 draws), keeping each photograph and its synthetic counterparts in the same resampled cluster. The pairing controls image content only approximately: synthetic counterparts are generated from captions of the real photographs and therefore remain semantically similar but are not identical in content.

B.3. Score-space cue alignment

Additional diagnostics support the score-association analysis of Section 4.4. On the *SynthCLIC* test split, the median $|r_q^{\text{within}}|$ across the 168 cues is 0.119, compared with 0.459 for the strongest cue, indicating a distributed association profile rather than dominance by a single attribute. The shared-space analysis head D_e produces a broadly similar cue ranking to the canonical detector D_h (Spearman $\rho = 0.864$). Uncertainty estimates use a cluster bootstrap over source photographs so that paired variants are resampled together.

B.4. Decision boundary differences across training datasets

We fit three independent D_e heads per training dataset and report cross-dataset decision boundary similarities averaged over all nine cross-seed pairs. We use the covariance-weighted cosine of Eq. (19). Under the *SynthCLIC* shared-space covariance,

$$\cos_{\Sigma}(D_{e,\text{SynthCLIC}}, D_{e,\text{CNNSpot}}) = -0.2068. \quad (22)$$

compared with a raw cosine of -0.061 . *SynthBuster+* is positively aligned with both *SynthCLIC* ($\cos_{\Sigma} = 0.311$) and *CNNSpot* (0.353), indicating that the strong disagreement is specific to the *SynthCLIC*–*CNNSpot* pair rather than a general opposition between generator families.

Fig. 7 decomposes the signed difference between the *CNNSpot*- and *SynthCLIC*-trained directions onto the 168 cue axes. The dominant terms separate processing and artifact cues on the *CNNSpot* side from photographic, optical, and grain-related cues on the *SynthCLIC* side, supporting a distribution-specific interpretation of the learned boundary. Because *CNNSpot* is nearly perfectly separable in the shared representation, its exact fitted normal is weakly identified; we therefore interpret the boundary similarities and cue decomposition as diagnostic summaries rather than as a unique or universal explanation of synthetic imagery.

B.5. Concept modeling

B.5.1. Sparsity sensitivity (β)

Table 10 shows the effect of varying the sparsity weight β on concept-model detection and on the number of active concepts (mean $\pm$ std over five seeds). Detection peaks at $\beta = 10^{-4}$ (0.88 ± 0.01 mAP with 13.7 active concepts) and declines mildly toward both ends of the range, to 0.84 at $\beta = 10^{-5}$ and at $\beta = 10^{-2}$, while the mean number of active concepts falls from about 73 to 0.2. A useful operating range is therefore roughly 10^{-4} to 3×10^{-4} , combining strong detection with moderate gate activity.

At the strongest regularization, the classifier compensates for fewer active gates with larger remaining weights, so active-concept count should not be interpreted directly as explanation complexity.

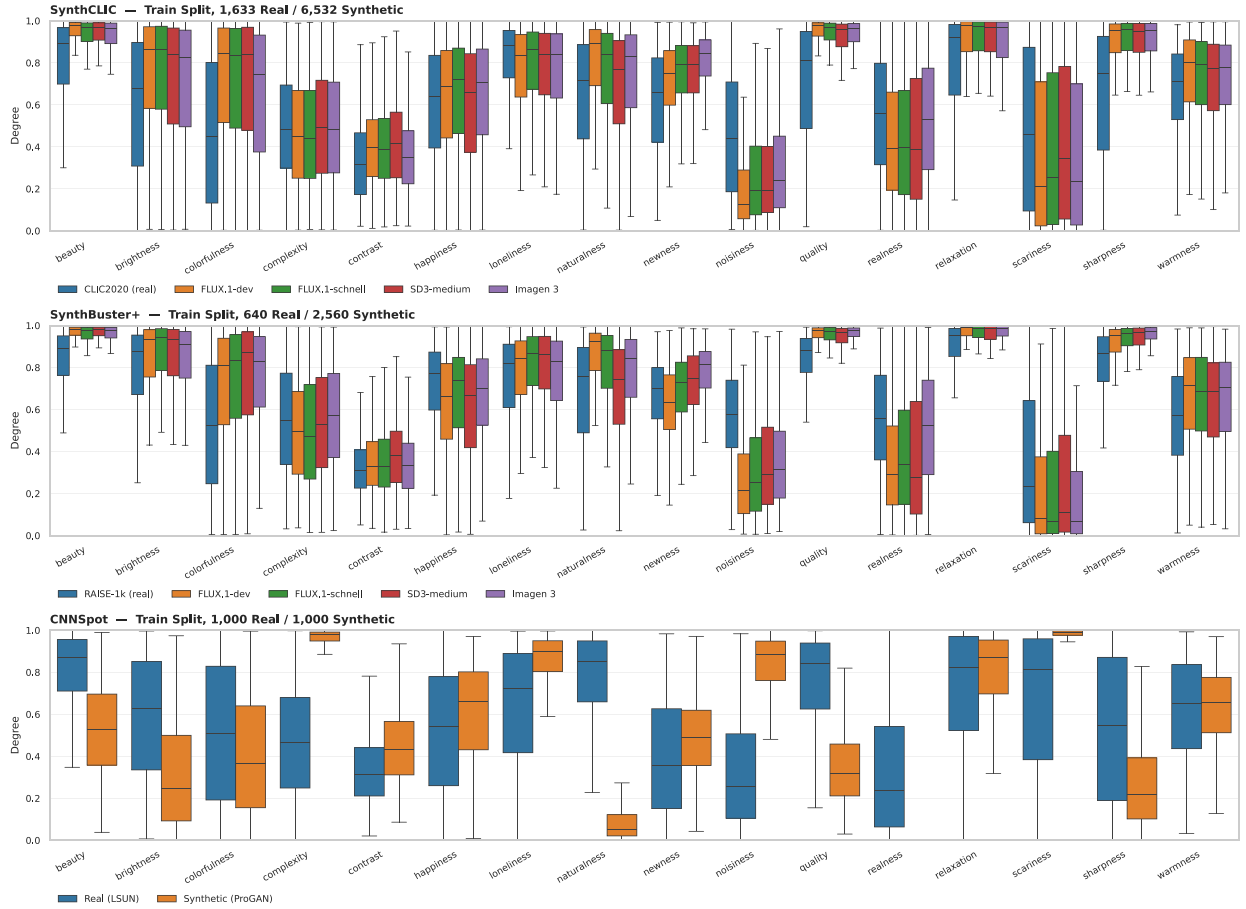


Fig. 10. Distribution of 16 CLIP-IQA [61] attribute scores for *CNNSpot*, *SynthBuster+*, and *SynthCLIC*, grouped by image source. The scores represent CLIP-based perceptual attributes rather than direct pixel measurements.

Table 10

Concept-sparsity weight β vs. detection on SynthCLIC (per-generator mAP), mean number of active concepts, total gate mass, and max classifier weight (mean $\pm$ std over 5 seeds). As β grows the gates collapse but the classifier rescales ($\max|W_{\text{clf}}|$ grows), so mAP is preserved at near-zero active concepts.

β	mAP	#active	gate mass	$\max W_{\text{clf}} $
10^{-5}	0.84 ± 0.01	72.6 ± 2.1	73.00 ± 1.98	6 ± 0
10^{-4}	0.88 ± 0.01	13.7 ± 2.7	14.78 ± 2.75	13 ± 2
3×10^{-4}	0.87 ± 0.02	6.7 ± 1.8	6.98 ± 1.84	21 ± 6
10^{-3}	0.84 ± 0.02	1.8 ± 0.5	1.94 ± 0.49	32 ± 7
10^{-2}	0.84 ± 0.01	0.2 ± 0.1	0.29 ± 0.06	51 ± 11

B.5.2. Concept-selection stability

Across seeds, detection performance and mean sparsity remain stable, but the selected concepts do not. At $\beta = 10^{-4}$, the top-30 concept sets have mean Jaccard overlap ≈ 0.24 . We therefore interpret recurring cue families rather than individual concepts and treat Fig. 8 as a representative profile rather than a unique explanation.

B.6. CLIP classifier

Table 11 shows detailed results for each model.

Table 12 shows ablations with respect to the number k of orthogonal dimensions.

B.6.1. Effect of orthogonality constraint

Table 13 compares activation- and weight-level orthogonality regularization for the factorized $k = 8$ head. We report

$$O_W = \|\mathbf{I} - \mathbf{W}_1^T \mathbf{W}_1\|_F, \quad (23)$$

$$O_A = \|\mathbf{I} - \bar{\mathbf{A}}^T \bar{\mathbf{A}}\|_F, \quad (24)$$

where $\bar{\mathbf{A}}$ is the column-normalized activation matrix defined in Appendix A.4. The activation regularizer used during training is λO_A^2 (Eq. (17)). Table 13 reports the corresponding unsquared diagnostics.

Detection performance is unchanged across variants (0.92 mAP), but the regularizers affect different objects. Weight regularization enforces $O_W = 0.00$ while leaving substantial activation non-orthogonality ($O_A = 2.64$). Activation regularization directly reduces O_A to 0.43 and also produces more orthogonal weights ($O_W = 1.05$) than using no auxiliary loss ($O_W = 2.25$). This supports activation orthogonality when the resulting activation dimensions are the objects of analysis.

B.6.2. CLIP backbone ablation

Table 14 gives the evidence for the backbone claim of Section 4.1: in-domain mAP on SynthCLIC is similar across the three OpenAI CLIP ViT backbones (0.89–0.92), with a slight edge for the largest (ViT-L/14-336), and the gap between ViT-B/16 and ViT-B/32 is small and dataset-dependent.

Data availability

We will release *SynthCLIC*, *SynthBuster+*, the images used in *CNNSpot*, our antonym vocabulary C_{antonyms} , the prompts for *SynthCLIC* and *SynthBuster+*, the model checkpoints, and all inference code upon publication <https://github.com/marco-willi/clip-cues>, subject to the licenses of the underlying photographs.

Table 11

Per-generator detection performance of the CLIP linear head (D_h , 1024-d frozen pooler_output), for each training set (columns) on each test set (rows). AP is the per-generator average precision of the paper's Convention A: on *CNNSpot* each generator is scored against *its own* real images (matched pairing), on *SynthCLIC* and *SynthBuster+* against the single shared real pool (shared pairing). ACC uses the head's own decision rule, $\hat{p} > 0.5$ (equivalently $\logit z > 0$), applied uniformly to every cell and computed on the same generator group as AP; no threshold is tuned on any test set. Per-column means reproduce the cross-dataset matrix of Table 3. All heads are trained under one matched recipe on frozen cached features without augmentation.

Test set	Generator	SynthCLIC		SynthBuster+		CNNSpot		Combined	
		ACC	AP	ACC	AP	ACC	AP	ACC	AP
SynthCLIC	FLUXDev	0.71	0.93	0.50	0.58	0.49	0.47	0.65	0.87
	FLUXSchnell	0.71	0.95	0.50	0.61	0.50	0.53	0.65	0.91
	Imagen3	0.70	0.88	0.50	0.58	0.49	0.48	0.64	0.80
	SD3M	0.71	0.93	0.50	0.67	0.52	0.61	0.66	0.89
	Mean	0.71	0.92	0.50	0.61	0.50	0.52	0.65	0.87
SynthBuster+	DALLE2	0.39	0.43	0.88	0.97	0.65	0.73	0.76	0.88
	DALLE3	0.77	0.98	0.90	1.00	0.33	0.31	0.81	1.00
	Firefly	0.57	0.63	0.90	0.99	0.53	0.57	0.80	0.97
	FluxDev	0.77	0.97	0.90	1.00	0.33	0.32	0.81	0.99
	FluxSchnell	0.77	0.98	0.90	1.00	0.34	0.33	0.81	0.99
	Glide	0.53	0.57	0.90	1.00	0.61	0.70	0.81	0.98
	Imagen3	0.76	0.95	0.90	0.99	0.34	0.33	0.81	0.97
	MJv5	0.75	0.93	0.90	0.99	0.36	0.35	0.81	0.98
	SD1.3	0.66	0.77	0.90	1.00	0.48	0.52	0.81	0.99
	SD1.4	0.69	0.78	0.90	1.00	0.48	0.50	0.81	0.98
	SD2	0.50	0.58	0.90	1.00	0.65	0.76	0.81	0.98
CNNSpot	SD3M	0.77	0.96	0.90	1.00	0.40	0.40	0.81	0.99
	SDXL	0.64	0.78	0.90	1.00	0.55	0.63	0.81	0.99
	Mean	0.66	0.79	0.90	1.00	0.47	0.50	0.81	0.98
	BigGAN	0.46	0.47	0.51	0.66	0.98	1.00	0.79	0.96
	CRN	0.72	0.74	0.58	0.78	0.74	1.00	0.83	0.90
	CycleGAN	0.31	0.35	0.52	0.76	0.93	1.00	0.63	0.91
	DALLE	0.34	0.36	0.51	0.69	0.97	1.00	0.75	0.89
	DeepFake	0.55	0.60	0.51	0.50	0.69	0.92	0.65	0.76
	GauGAN	0.42	0.41	0.52	0.88	0.97	1.00	0.82	0.98
	GLIDE-100-10	0.27	0.34	0.52	0.74	0.86	0.98	0.76	0.89
	GLIDE-100-27	0.26	0.34	0.52	0.75	0.85	0.98	0.75	0.88
	GLIDE-50-27	0.28	0.34	0.52	0.74	0.88	0.99	0.76	0.90
	Guided	0.34	0.36	0.53	0.67	0.73	0.90	0.66	0.72
	IMLE	0.54	0.55	0.56	0.75	0.74	1.00	0.85	0.92
	LDM-100	0.24	0.33	0.52	0.75	0.98	1.00	0.77	0.91
	LDM-200	0.23	0.33	0.52	0.76	0.97	1.00	0.76	0.91
	LDM-200-CFG	0.41	0.40	0.52	0.66	0.86	0.98	0.74	0.84
	ProGAN	0.25	0.33	0.50	0.68	1.00	1.00	0.98	1.00
	SAN	0.43	0.42	0.52	0.63	0.57	0.71	0.55	0.64
	SeeingDark	0.46	0.45	0.57	0.87	0.67	0.83	0.70	0.78
	StarGAN	0.45	0.44	0.44	0.33	0.99	1.00	0.87	0.92
	StyleGAN	0.39	0.41	0.53	0.50	0.88	0.99	0.92	0.98
	StyleGAN2	0.47	0.46	0.50	0.46	0.77	0.98	0.87	0.94
	WhichFacesReal	0.46	0.45	0.53	0.48	0.94	0.99	0.73	0.86
	Mean	0.39	0.42	0.52	0.67	0.86	0.96	0.77	0.88

Table 12

Mean AP (mAP) by dataset and k -value for models trained on *SynthCLIC*. These runs use the augmented factorized-head protocol (Table 7), not the canonical classifier.

k	2	4	8	16
Test dataset				
CNNSpot	0.39	0.38	0.37	0.38
SynthBuster+	0.81	0.79	0.79	0.78
SynthCLIC	0.91	0.94	0.92	0.92

Table 13

Orthogonality ablation on *SynthCLIC*. Detection performance is unchanged across variants. Weight regularization enforces orthogonality of W_1 , whereas activation regularization most strongly reduces the activation-space orthogonality error O_A .

Auxiliary loss	mAP	$O_W \downarrow$	$O_A \downarrow$
None	0.92	2.25	6.81
Activation	0.92	1.05	0.43
Weight	0.92	0.00	2.64

Table 14

In-domain mAP across CLIP backbones.

Test set (in-domain)	ViT-B/16	ViT-B/32	ViT-L/14
CNNSpot	0.93	0.87	0.96
combined	–	–	0.49
SynthBuster+	0.99	0.98	1.00
SynthCLIC	0.89	0.90	0.92

References

- [1] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution image synthesis with latent diffusion models. 2022, URL: <http://arxiv.org/abs/2112.10752>. arXiv:2112.10752 [cs].
- [2] Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, et al. SDXL: Improving latent diffusion models for high-resolution image synthesis. 2023, URL: <http://arxiv.org/abs/2307.01952>. arXiv:2307.01952 [cs].
- [3] Esser P, Kulal S, Blattmann A, Entezari R, Müller J, Saini H, et al. Scaling rectified flow transformers for high-resolution image synthesis. 2024, <http://dx.doi.org/10.48550/arXiv.2403.03206>, URL: <http://arxiv.org/abs/2403.03206> arXiv:2403.03206 [cs].
- [4] Imagen-Team-Google, Baldridge J, Bauer J, Bhutani M, Brichtova N, Bunner A, et al. Imagen 3. 2024, <http://dx.doi.org/10.48550/arXiv.2408.07009>, URL: <http://arxiv.org/abs/2408.07009> arXiv:2408.07009 [cs].
- [5] Lipman Y, Chen RTQ, Ben-Hamu H, Nickel M, Le M. Flow matching for generative modeling. 2023, <http://dx.doi.org/10.48550/arXiv.2210.02747>, URL: <http://arxiv.org/abs/2210.02747> [cs].
- [6] Lu Z, Huang D, Bai L, Qu J, Wu C, Liu X, et al. Seeing is not always believing: Benchmarking human and model perception of AI-generated images. 2023, URL: <http://arxiv.org/abs/2304.13023> arXiv:2304.13023 [cs].
- [7] Casu M, Guarnera L, Caponnetto P, Battiato S. AI mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions. 2023, URL: <http://arxiv.org/abs/2312.16220> arXiv:2312.16220 [cs].
- [8] Nightingale SJ, Farid H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proc Natl Acad Sci 2022;119:e2120481119. <http://dx.doi.org/10.1073/pnas.2120481119>, URL: <https://pnas.org/doi/full/10.1073/pnas.2120481119>.
- [9] Chesney B, Citron D. Deep fakes: A looming challenge for privacy, democracy, and national security. Calif L Rev 2019;107:1753. <http://dx.doi.org/10.2139/ssrn.3213954>.
- [10] Cordey S. Cyber influence operations: an overview and comparative analysis. Technical report, ETH Zurich; 2019, <http://dx.doi.org/10.3929/ETHZ-B-000382358>, URL: <http://hdl.handle.net/20.500.11850/382358>, artwork Size: 38 p. Medium: application/pdf.
- [11] Farid H. Creating, using, misusing, and detecting deep fakes. J Online Trust Saf 2022;1. <http://dx.doi.org/10.54501/jots.v1i4.56>, URL: <https://tsjournal.org/index.php/jots/article/view/56>.
- [12] Marra F, Gragnaniello D, Verdoliva L, Poggi G. Do GANs leave artificial fingerprints? In: 2019 IEEE conference on multimedia information processing and retrieval (MIPR). San Jose, CA, USA: IEEE; 2019, p. 506–11. <http://dx.doi.org/10.1109/MIPR.2019.00103>, URL: <https://ieeexplore.ieee.org/document/8695364/>.
- [13] Wang SY, Wang O, Zhang R, Owens A, Efros AA. CNN-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 8695–704, URL: <http://arxiv.org/abs/1912.11035>.
- [14] Cozzolino D, Poggi G, Niefesner M, Verdoliva L. Zero-shot detection of AI-generated images. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer vision – ECCV 2024. vol. 15076, Cham: Springer Nature Switzerland; 2025, p. 54–72. http://dx.doi.org/10.1007/978-3-031-72649-1_4, URL: https://link.springer.com/10.1007/978-3-031-72649-1_4, series Title: Lecture Notes in Computer Science.
- [15] Tariang D, Corvi R, Cozzolino D, Poggi G, Nagano K, Verdoliva L. Synthetic image verification in the era of generative AI: What works and what isn't there yet. 2024, <http://dx.doi.org/10.48550/arXiv.2405.00196>, URL: <http://arxiv.org/abs/2405.00196> arXiv:2405.00196 [cs].

- [16] Zhang X, Karaman S, Chang SF. Detecting and simulating artifacts in GAN fake images. In: 2019 IEEE international workshop on information forensics and security (WIFS). Delft, Netherlands: IEEE; 2019, p. 1–6. <http://dx.doi.org/10.1109/WIFS47025.2019.9035107>, URL: <https://ieeexplore.ieee.org/document/9035107/>.
- [17] Carlini N, Farid H. Evading deepfake-image detectors with white- and black-box attacks. In: 2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW). Seattle, WA, USA: IEEE; 2020, p. 2804–13. <http://dx.doi.org/10.1109/CVPRW50498.2020.00337>, URL: <https://ieeexplore.ieee.org/document/9150604/>.
- [18] Epstein DC, Jain I, Wang O, Zhang R. Online detection of AI-generated images. In: 2023 IEEE/CVF international conference on computer vision workshops. ICCVW, Paris, France: IEEE; 2023, p. 382–92. <http://dx.doi.org/10.1109/ICCVW60793.2023.00045>, URL: <https://ieeexplore.ieee.org/document/10350523/>.
- [19] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Meila M, Zhang T, editors. Proceedings of the 38th international conference on machine learning. PMLR; 2021, p. 8748–63, URL: <https://proceedings.mlr.press/v139/radford21a.html>.
- [20] Ojha U, Li Y, Lee YJ. Towards universal fake image detectors that generalize across generative models. 2023, URL: <https://arxiv.org/abs/2302.10174> [cs].
- [21] Cioni D, Tzelepis C, Seidenari L, Patras I. Are CLIP features all you need for universal synthetic image origin attribution?. 2024, URL: <http://arxiv.org/abs/2408.09153> arXiv:2408.09153 [cs].
- [22] Huang J, Chen C, Mishra A, Kwon BC, Liu Z, Bryan C. ASAP: Interpretable analysis and summarization of AI-generated image patterns at scale. 2024, URL: <http://arxiv.org/abs/2404.02990> arXiv:2404.02990 [cs].
- [23] Khan SA, Dang-Nguyen DT. CLIPping the deception: Adapting vision-language models for universal deepfake detection. 2024, <http://dx.doi.org/10.48550/arXiv.2402.12927>, URL: <http://arxiv.org/abs/2402.12927> arXiv:2402.12927 [cs].
- [24] Gaintseva T, Kushnareva L, Magai G, Piontkovskaya I, Nikolenko S, Benning M, et al. Improving interpretability and robustness for the detection of AI-generated images. 2024, URL: <http://arxiv.org/abs/2406.15035> arXiv:2406.15035 [cs].
- [25] Koutlis C, Papadopoulos S. Leveraging representations from intermediate Encoder-blocks for synthetic image detection. 2024, <http://dx.doi.org/10.48550/arXiv.2402.19091>, URL: <http://arxiv.org/abs/2402.19091> arXiv:2402.19091 [cs].
- [26] Cozzolino D, Poggi G, Corvi R, Nießner M, Verdoliva L. Raising the bar of AI-generated image detection with CLIP. 2024, URL: <http://arxiv.org/abs/2312.00195> arXiv:2312.00195 [cs].
- [27] De Rosa V, Guillard F, Poggi G, Cozzolino D, Verdoliva L. Exploring the adversarial robustness of CLIP for AI-generated image detection. In: 2024 IEEE international workshop on information forensics and security. WIFS, Rome, Italy: IEEE; 2024, p. 1–6. <http://dx.doi.org/10.1109/WIFS61860.2024.10810719>, URL: <https://ieeexplore.ieee.org/document/10810719/>.
- [28] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Adv Neural Inf Process Syst 2014;3:2672–80. <http://dx.doi.org/10.3156/jsoft.29.5.177.2>.
- [29] Kingma DP, Welling M. Auto-Encoding variational Bayes. 2013, CoRR abs/1312.6114 URL: <https://api.semanticscholar.org/CorpusID:216078090>.
- [30] Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using real NVP. 2017, <http://dx.doi.org/10.48550/arXiv.1605.08803>, URL: <http://arxiv.org/abs/1605.08803> arXiv:1605.08803 [cs].
- [31] Yu J, Xu Y, Koh JY, Luong T, Baid G, Wang Z, et al. Scaling autoregressive models for content-rich text-to-image generation. 2022, <http://dx.doi.org/10.48550/arXiv.2206.10789>, URL: <http://arxiv.org/abs/2206.10789> arXiv:2206.10789 [cs].
- [32] Oord Avd, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks. 2016, <http://dx.doi.org/10.48550/arXiv.1601.06759>, URL: <http://arxiv.org/abs/1601.06759> arXiv:1601.06759 [cs].
- [33] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. 2020, <http://dx.doi.org/10.48550/arXiv.2006.11239>, URL: <http://arxiv.org/abs/2006.11239> arXiv:2006.11239 [cs].
- [34] Midjourney, URL: <https://www.midjourney.com/home/>.
- [35] Announcing black forest labs. 2024, URL: <https://bfl.ai/announcements/24-08-01-bfl>.
- [36] Adobe firefly, URL: <https://www.adobe.com/products/firefly.html>.
- [37] Introducing 4o image generation. 2025, URL: <https://openai.com/index/introducing-4o-image-generation/>.
- [38] Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, et al. A survey on multimodal large language models. Natl Sci Rev 2024;11:nwae403. <http://dx.doi.org/10.1093/nsr/nwae403>, URL: <http://arxiv.org/abs/2306.13549> arXiv:2306.13549 [cs].
- [39] Zhou C, Yu L, Babu A, Tirumala K, Yasunaga M, Shams L, et al. Transfusion: Predict the next token and diffuse images with one multi-modal model. 2024, <http://dx.doi.org/10.48550/arXiv.2408.11039>, URL: <http://arxiv.org/abs/2408.11039> arXiv:2408.11039 [cs].
- [40] Corvi R, Cozzolino D, Poggi G, Nagano K, Verdoliva L. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In: 2023 IEEE/CVF conference on computer vision and pattern recognition workshops. CVPRW, Vancouver, BC, Canada: IEEE; 2023a, p. 973–82. <http://dx.doi.org/10.1109/CVPRW59228.2023.00104>, URL: <https://ieeexplore.ieee.org/document/10209003/>.
- [41] Corvi R, Cozzolino D, Zingarini G, Poggi G, Nagano K, Verdoliva L. On the detection of synthetic images generated by diffusion models. In: ICASSP 2023 - 2023 IEEE international conference on acoustics, speech and signal processing. ICASSP, Rhodes Island, Greece: IEEE; 2023b, p. 1–5. <http://dx.doi.org/10.1109/ICASSP49357.2023.10095167>, URL: <https://ieeexplore.ieee.org/document/10095167/>.
- [42] Ricker J, Damm S, Holz T, Fischer A. Towards the detection of diffusion model deepfakes. 2024, URL: <http://arxiv.org/abs/2210.14571> arXiv:2210.14571 [cs].
- [43] Sha Z, Li Z, Yu N, Zhang Y. DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. 2023, URL: <http://arxiv.org/abs/2210.06998> arXiv:2210.06998 [cs].
- [44] Wang Z, Bao J, Zhou W, Wang W, Hu H, Chen H, et al. DIRE for diffusion-generated image detection. In: 2023 IEEE/CVF international conference on computer vision. ICCV, Paris, France: IEEE; 2023, p. 22388–98. <http://dx.doi.org/10.1109/ICCV51070.2023.02051>, URL: <https://ieeexplore.ieee.org/document/10377654/>.
- [45] Wong YJ, Ng TK. Local statistics for generative image detection. 2023, URL: <http://arxiv.org/abs/2310.16684> arXiv:2310.16684 [cs].
- [46] Wu H, Zhou J, Zhang S. Generalizable synthetic image detection via language-guided contrastive learning. 2023, URL: <http://arxiv.org/abs/2305.13800> arXiv:2305.13800 [cs].
- [47] Cozzolino D, Thies J, Rössler A, Riess C, Nießner M, Verdoliva L. Forensic-Transfer: Weakly-supervised domain adaptation for forgery detection. 2019, URL: <http://arxiv.org/abs/1812.02510> arXiv:1812.02510 [cs].
- [48] Lai Y, Wang H, Yang J, Kang X, Li B, Shen L, et al. GM-DF: Generalized multi-scenario deepfake detection. In: Proceedings of the 33rd ACM international conference on multimedia. Dublin Ireland: ACM; 2025, p. 4300–9. <http://dx.doi.org/10.1145/3746027.3755386>, URL: <https://dl.acm.org/doi/10.1145/3746027.3755386>.
- [49] Lai Y, Yu Z, Wang J, Shen L, Xu Y, Cao X. Agent4faceforgery: Multi-agent llm framework for realistic face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2026, p. 14073–83.
- [50] Park J, Owens A. Community forensics: Using thousands of generators to train fake image detectors. In: 2025 IEEE/CVF conference on computer vision and pattern recognition. CVPR, Nashville, TN, USA: IEEE; 2025, p. 8245–57. <http://dx.doi.org/10.1109/CVPR52734.2025.00772>, URL: <https://ieeexplore.ieee.org/document/11093773/>.
- [51] Zhou Z, Sun K, Chen Z, Kuang H, Sun X, Ji R. StealthDiffusion: Towards evading diffusion forensic detection through diffusion model. In: ACM multimedia 2024. 2024, URL: <https://openreview.net/forum?id=jnQFcU9Bw>.
- [52] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020, URL: <http://arxiv.org/abs/2010.11929> arXiv:2010.11929 [cs].
- [53] Cherti M, Beaumont R, Wightman R, Wortsman M, Ilharco G, Gordon C, et al. Reproducible scaling laws for contrastive language-image learning. In: 2023 IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2023, p. 2818–29. <http://dx.doi.org/10.1109/CVPR52729.2023.00276>, URL: <http://arxiv.org/abs/2212.07143> arXiv:2212.07143 [cs].
- [54] Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. 2022, <http://dx.doi.org/10.48550/arXiv.2210.08402>, URL: <http://arxiv.org/abs/2210.08402> arXiv:2210.08402 [cs].
- [55] Gandelsman Y, Efros AA, Steinhart J. Interpreting CLIP's image representation via text-based decomposition. 2024, <http://dx.doi.org/10.48550/arXiv.2310.05916>, URL: <http://arxiv.org/abs/2310.05916> arXiv:2310.05916 [cs].
- [56] Panousis KP, Ienco D, Marcos D. Sparse linear concept discovery models. 2023, <http://dx.doi.org/10.48550/arXiv.2308.10782>, URL: <http://arxiv.org/abs/2308.10782> arXiv:2308.10782 [cs].
- [57] Yu F, Seff A, Zhang Y, Song S, Funkhouser T, Xiao J. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. 2016, <http://dx.doi.org/10.48550/arXiv.1506.03365>, URL: <http://arxiv.org/abs/1506.03365> arXiv:1506.03365 [cs].
- [58] Bammey G. Synthbuster: Towards detection of diffusion model generated images. IEEE Open J Signal Process 2023;1–9. <http://dx.doi.org/10.1109/OJSP.2023.3337714>, URL: <https://ieeexplore.ieee.org/document/10334046/>.
- [59] Dang-Nguyen DT, Pasquini C, Conotter V, Boato G. RAISE: a raw images dataset for digital image forensics. In: Proceedings of the 6th ACM multimedia systems conference. Portland Oregon: ACM; 2015, p. 219–24. <http://dx.doi.org/10.1145/2713168.2713194>, URL: <https://dl.acm.org/doi/10.1145/2713168.2713194>.
- [60] Comanici G, Bieber E, Schaeckermann M, Pasupat I, Sachdeva N, Dhillon I, et al. Gemini 2.5: Pushin the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. 2025, <http://dx.doi.org/10.48550/arXiv.2507.06261>, URL: <http://arxiv.org/abs/2507.06261> arXiv:2507.06261 [cs].
- [61] Wang J, Chan KCK, Loy CC. Exploring CLIP for assessing the look and feel of images. 2022, <http://dx.doi.org/10.48550/arXiv.2207.12396>, URL: <http://arxiv.org/abs/2207.12396> arXiv:2207.12396 [cs].