A Novel Approach for Detecting AI-Generated Images in Zero-Shot Setting

Nithin Skantha Murugan
University of Maryland
College Park, MD 20740
nithin10@umd.edub

Krishna P Taduri
University of Maryland
College Park, MD 20740
kptaduri@umd.edu

Rishie Raj
University of Maryland
College Park, MD 20740
rraj27@umd.edu

Abstract

As AI-generated content becomes increasingly pervasive, distinguishing between AI-generated and human-captured images is a critical challenge. This paper introduces a novel approach to detecting AI-generated images based on cross-perplexity and perplexity computations. Leveraging LlamaGen, a state-of-the-art autoregressive image generation model, the proposed framework utilizes tokenization, vision encoding, and class-conditioned starting tokens to compute detection metrics that reliably identify synthetic visual content. Preliminary results indicate the effectiveness of this approach in a zero-shot setting, highlighting its potential for broad applicability without dependence on training data from generative models.

1 Introduction

AI-generated content has proliferated into every sphere of media consumption and it is becoming increasingly difficult to distinguish between human-captured and synthetic visual content. Advances in autoregressive and diffusion-based generative models such as LlamaGen, GANs, Stable Diffusion, DALL·E and MidJourney have enabled the creation of highly realistic images. While these models have a lot of creative and practical applications, they are also being used for nefarious purposes such as spreading misinformation, deepfakes and harmful synthetic media content. This has resulted in a growing demand for reliable and generalizable detection mechanisms that can identify synthetic images from a variety of model families, including models having unseen or novel architectures.

Most of the existing detection methods are designed on a supervised paradigm and their classifiers are trained on labeled datasets containing real and synthetically generated images. While these approaches work for images generated by specific models, they do not generalize well to detecting images generated by unseen models. These supervised learning approaches learn features and artifacts of images generated by the respective generation models and hence they fail when they see images from architectures that were not part of their training dataset. Also, the issue with supervised training is that it requires a lot of labeled traning data which is not often practical.

In order to overcome these challenges, we have come up with an approach that uses a zero-shot detection strategy already implemented in the language domain Hans et al. [3]. The *Binoculars* approach uses *perplexity* and *cross-perplexity* metrics to measure the level of certainty of token sequences. The underlying assumption for this approach is that human and machine-generated content will have different measures of these metrics. Perplexity is a concept of language modeling which measures how ”surprising” a token sequence is to a given language model. Meanwhile, cross-perplexity gives the same measure for different model outputs against each other.

In this paper, we have integrated the Binoculars approach with LlamaGen, an autoregressive image generation model. LLamaGen is used to tokenize images into a sequence using a vector-quantized

autoencoder (VQ-VAE). This allows us to calculate perplixity and cross-perplexity in the same way as was done for text token sequences in Hans et al. [3]. We follow the Binoculars appraoch, wherein we have chosen an *observer* from the LlamaGen family of models. The token sequences of the images generated by all our test models, called *performers*, are contrasted with the observer to calculate the perplexity and cross-perplexity. By comparing the outputs of performer models against the expectations of the observer, we can identify the deviations in natural and synthetic data distributions. This helps us detect synthetic content without actually traning on any model-specific data, in a zero-shot detection setting.

In summary, this report details the adaptation of the Binoculars approach to the image domain and its integration with LlamaGen to calculate perplexity-related metrics for a wide-range of image generative models for synthetic image detection. The key contributions of this work are as follows:

1. **Adaptation of Perplexity-Based Detection:** We adapted a two novel metrics, namely perplexity and cross-perplexity, developed for language domain, into the image domain by tweaking the tokenizing mechanism.
2. **Integration with LlamaGen:** We have used the LlamaGen tokenization approach to calculate the decision metrics from the synthetically generated images.
3. **Evaluation on Diverse Generative Models:** We have validated our appraoch on images generated by a wide range of models such as LlamaGen, Stable Diffusion, DALL·E and Emu3, getting promising results. This highlights the ability of our approach to generalize well to different models.
4. **Zero-Shot Detection:** Our approach does not require any training procedure, hence making it fast, efficient and generalizable.

2 Related Work

All existing approaches in the AI-image detection space primarily rely on supervised learning. While this is effective within specific domains, it struggles to generalize to images from unseen generative models. These limitations come from the fact that supervised detectors tend to overfit to model-specific artifacts, while their reliance on large amounts of labeled training data make them highly impractical to obtain and train.

In order to address these challenges, researchers are exploring zero-shot detection methods that do not require labeled synthetic data for training. Cozzolino et al. [2] introduced a Zero-Shot Entropy-based Detector (ZED), where pixel-level probability distributions are predicted using a lossless image encoder. By measuring the uncertainty or entropy of a synthetic image and comparing it with the expected coding costs of a real image, ZED achieves robust detection across various generative models such as DALL·E, MidJourney, and Stable Diffusion, without training.

On the other hand, pre-trained Vision-Language Models (VLMs) like CLIP have shown great generalization capabilities in detecting synthetic images. Cozzolino et al. [1] utilized CLIP features in a lightweight framework to differentiate synthetic images with high accuracy across unseen architectures. Ojha et al. [4] expanded on this idea by showing that linear probing and nearest-neighbor methods in VLM feature spaces can outperform traditional classifiers, particularly for diffusion-based generators and GAN-based models. These methods prove that there is potential for zero-shot methods to replace supervised paradigms by solving their limitations.

There is a huge diversity in contemporary image generative models, such as GANs, diffusion models, and autoregressive models. This poses a complex challenge for detection. Supervised classifiers often fail to capture the minute differences in statistical patterns between different generative families. For example, detectors trained on GAN-based images may misclassify diffusion-generated images due to differences in noise distribution and artifact characteristics. Ojha et al. mentioned these limitations and emphasized the need for detectors to treat all real and synthetic images with equal importance, rather than treating real and synthetic classification as a binary problem.

An interesting innovation happened with the adaptation of perplexity, which is a language modeling concept, for detecting AI-generated language content. This concept was first shown in Binoculars, which uses perplexity and cross-perplexity metrics to measure the predictability of token sequences. This approach helped them classify real and synthetic text content in a zero-shot fashion. By com-

paring the differences between two closely related models (observer and performer), Binoculars was able to identify the subtle patterns that indicate synthetic content without requiring model-specific training.

Inspired by this approach, our work extends the use of perplexity-based metrics to the image domain. By integrating the Binoculars framework with the autoregressive capabilities of LlamaGen, we tokenize images into sequences and calculate these metrics over image tokens. This adaptation enables robust detection of synthetic content across diverse generative models, including unseen architectures such as Stable Diffusion, DALL·E, and Emu3.

3 Methodology

Similar to how large language models (LLMs) tend to generate text that appears predictable to an LLM, autoregressive image generation models often create images that are predictable to similar models. Conversely, human-created images can exhibit less predictable features, resulting in higher perplexity when assessed by an AI observer. This makes raw perplexity an attractive metric for distinguishing human-created images, as higher perplexity can suggest human authorship.

However, this intuition falters when context-specific cues, akin to hand-crafted prompts in text, are involved. For instance, a model conditioned on a specific context, such as "blue skies with a single bird," might generate an image that aligns perfectly with that prompt, resulting in low perplexity. But without knowledge of the prompt, the generated image may seem highly surprising, leading to falsely categorizing it as human-created. Similarly, an image containing unexpected combinations, such as "a panda riding a bicycle in the desert," may exhibit high perplexity in the absence of its conditioning prompt, causing naive perplexity-based detection to fail.

Relying solely on raw perplexity for AI-generated image detection can lead to significant inaccuracies without considering the conditioning context. This is where cross-perplexity comes into play—by contrasting the predictability of the same image across models with shared contexts, we can mitigate the biases introduced by missing prompts, enabling a more reliable detection framework.

3.1 Framework and Model Setup

We employ two models, viz. performer and observer model, from the LlamaGen family of models based on an autoregressive framework for image generation, as the backbone of our detection metric. The dual-model setup allows us to compute cross-perplexity, capturing discrepancies in token predictability between models. The methodology incorporates two distinct models from the LlamaGen family [1]:

- **Performer Model**: This model is used to generate the logits which are in-turn used in perplexity calculation.
  - **Model Used:** LlamaGen-L
- **Observer Model**: This model is used to generate logits which are then used to calculate the cross-perplexity. This calculation also involves the logits generated by the performer model.
  - **Model Used:** LlamaGen-B

This dual-model setup allows us to compute cross-perplexity, capturing discrepancies in token predictability between models.

3.2 Tokenization and Input Processing

Our framework employs the LlamaGen vision encoder to tokenize each input image into a sequence of discrete tokens, suitable for processing by an autoregressive generation model. For this generation process, we condition the model on predefined class labels. These labels are transformed into token

[1] https://github.com/FoundationVision/LlamaGen

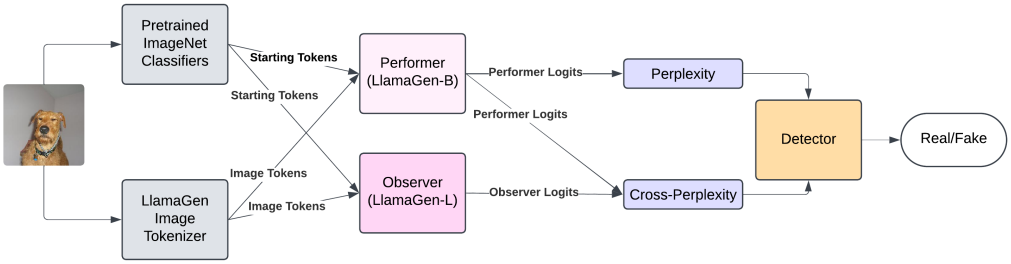


Figure 1: Model Setup

sequences, which serve as starting points for the autoregressive sampling process. The model generates a sequence of tokens based on these initial inputs, capturing the distributional characteristics of the image data.

During the sampling process, we employ standard autoregressive techniques to generate predictions for each token in the sequence. The generated token sequences are subsequently decoded into image representations using a vector quantized (VQ) decoder, ensuring the outputs are consistent with the tokenized structure. Class-conditioned initialization and controlled sampling ensure reliable logits which are key to perplexity and cross-perplexity computation for comparing observer and performer models.

3.3 Perplexity and Cross-Perplexity Computation

The proposed methodology relies on two key metrics to detect synthetic images: **perplexity** and **cross-perplexity**. These metrics evaluate the predictability of token sequences and discrepancies in token predictability, respectively.

Perplexity. Perplexity measures how well the performer model predicts the token sequences generated by an input image. Given a sequence of tokens $\mathbf{t} = \{t_1, t_2, \ldots, t_N\}$ with N tokens, and the probability $P(t_i|t_{1:i-1})$ assigned by the performer model to each token t_i, the perplexity PPL is computed to captures how "expected" the token sequence is according to the performer model, with lower perplexity indicating higher predictability. On a logarithmic scale, we have:

$$\log \text{PPL}_M(s) = -\frac{1}{L}\sum_{i=1}^{L} \log(Y_{ix_i}), \tag{1}$$

where

- M is a language model that predicts the next token in a sequence by outputting a probability distribution over the vocabulary $V = \{1, 2, \ldots, n\}$,
- $Y_{ij} = P(v_j \mid x_{0:i-1})$ is the probability assigned by M to the token $v_j \in V$, conditioned on the preceding tokens $x_{0:i-1}$,
- x_i is the i-th token ID in the tokenized input sequence $\vec{x} = T(s)$, where T is the tokenizer,
- L is the total number of tokens in the sequence s,
- Y_{ix_i} is the probability assigned by M to the actual i-th token x_i.

Cross-Perplexity. Cross-perplexity evaluates the alignment between the token predictions of the performer model and their surprisal when evaluated by the observer model. The cross-perplexity $CPPL$ is computed using the conditional entropy on a logarithmic scale:

$$\log \text{X-PPL}_{M_1,M_2}(s) = -\frac{1}{L}\sum_{i=1}^{L} M_1(s)_i \cdot \log(M_2(s)_i), \tag{2}$$

where

- M_1 and M_2 are two language models being compared,
- s is the input string, and $\vec{x} = T(s)$ is the tokenized form of s,
- L is the total number of tokens in the sequence s,
- $M_1(s)_i$ represents the i-th token probability predicted by M_1,
- $M_2(s)_i$ represents the i-th token probability predicted by M_2,
- $\log \text{X-PPL}_{M_1,M_2}(s)$ measures the average per-token cross-entropy between the outputs of M_1 and M_2.

Ratio. We compute the ratio of perplexity to cross-perplexity, defined as **Binocular Score** by Hans et al. [3]:

$$B_{M_1,M_2}(s) = \frac{\log \text{PPL}_{M_1}(s)}{\log \text{X-PPL}_{M_1,M_2}(s)}, \tag{3}$$

Lower values of $R(\mathbf{t})$ suggest high alignment between the performer and observer models, indicative of AI-generated images. Higher values indicate greater unpredictability as observed by the observer model, characteristic of human-authored content.

Implementation. The computation pipeline involves the following steps:

1. **Logit Processing:** Both the observer and performer models process token sequences to generate logits, which are used for entropy and probability calculations.
2. **Perplexity Computation:** The observer model computes perplexity by aggregating probabilities of the token sequence.
3. **Cross-Perplexity Computation:** The conditional entropy of the observer's logits, as evaluated by the observer, yields the cross-perplexity.
4. **Metric Aggregation:** The perplexity and cross-perplexity scores are combined to compute $R(\mathbf{t})$ as the detection metric.

By leveraging both perplexity and cross-perplexity, the framework effectively distinguishes between human-authored and AI-generated content without reliance on labeled training data.

4 Results and Analysis

Based on our model setup as earlier, we have calculated two metrics for classification, *Perplexity* and *Cross-Perplexity*. Our goal is to get a clear segregation between the data-points for real and synthetic images in a perplexity vs. cross-perplexity plot.

For classification of real and fake, we used a evaluation metric called *Binoculars Score* from Hans et al. [3], which is basically the ratio between the perplexity and the cross-perplexity. As can see from the plot, we have managed to derive fairly reasonable zones for real and synthetic images. The blues indicate *real images* and oranges indicate *synthetic images*. The plot is also inline with our intuitive reasoning, which is that, real images are supposed to have higher perplexity scores w.r.t. autoregressive models as their probability distribution is quite different than synthetic images, as evidenced by [4, 2, 1, 5, 6] and hence are more "surprising" for generative models.

In Table 1 we have shown the true-positive rates (TPR) at a particular false-positive rate (FPR) threshold. This highlights the effectiveness of our framework, especially at a stricter threshold of 0.1FPR. Our model demonstrates a strong ability to detect synthetic content while minimizing false positives, outperforming both the baseline models from Wang et al. [6] and Ricker et al. [5]. In Table 2 our framework shows a consistent performance across multiple models.

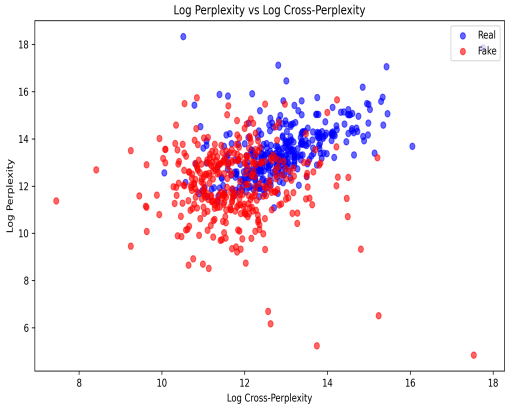
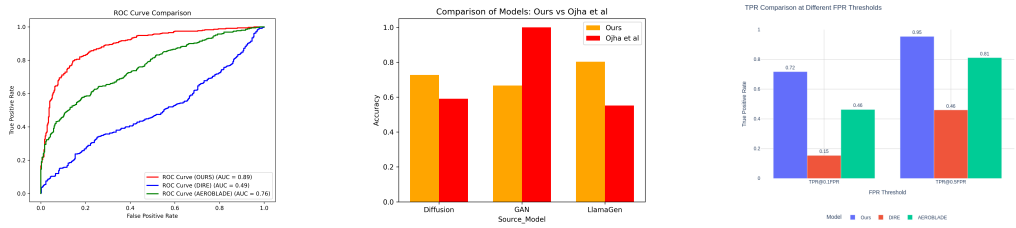


Figure 2: Perplexity vs. Cross-perplexity plot for AI-generated and real images



(a) True Positive Rates @FPR (b) Accuracy Comparison (c) TPR @Different FPRs

Figure 3: Comparison of metrics with state-of-the-art.

Table 1: Comparison of TPR Metrics Across Methods

Name	TPR@0.1FPR	TPR@0.5FPR
Ours	**0.7171**	**0.9543**
DIRE	0.1543	0.4600
AEROBLADE	0.4629	0.8114

Table 2: Comparison of Accuracy Metrics Across Methods

Name	Diffusion	GAN	LlamaGen
Ours	**0.7272**	**0.6660**	**0.8027**
Ojha	0.5909	**1.0000**	0.5513

5 Conclusion

In our work, we have introduced a novel framework for detecting synthetically-generated images by adapting perplexity and cross-perplexity metrics, which was originally developed for language-based detection, into the image domain. Using the capabilities of the LlamaGen family of autoregressive models, we have shown how tokenization and autoregressive generation could be used to compute robust detection metrics without following a supervised training approach. The results also show that our approach performs competitively in a zero-shot setting, achieving better detection accuracy compared to baseline methods at both strict (0.1FPR) and relaxed (0.5FPR) thresholds. By combining the strengths of perplexity with cross-perplexity computation, our framework generalizes effectively across diverse generative architectures, creating a promising future for scalable and efficient synthetic content detection.

6 Future Work

While our preliminary results have been quite promising, we have to extend our experiments on a much larger dataset. As of now, we have tested our model on 5000 images (1000 real and 4000 synthetic). It remains to be seen if we can achieve a good separation with higher number of images. Also, we are planning to add another metric along with perplexity and cross-perplexity, to create a better separation between the data-points. This is help us in coming up with a more robust decision boundary even with a larger testset. Finally, as of now, we are using an SVM to classify the images based on the decision score. As this SVM has to be trained on the testset, it takes away from our purpose to make our framework truly zero-shot. We want to come up with an approach where this classification method does not need to be learnt.

References

[1] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip, 2023.

[2] Davide Cozzolino, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva. Zero-shot detection of ai-generated images. In *European Conference on Computer Vision (ECCV)*, 2024.

[3] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text, 2024.

[4] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023.

[5] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9130–9140, June 2024.

[6] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*, 2023.