Detecting AI-Generated Content on Social Media with Multi-modal Language Models

Chenyang Yang[1]*, Shen Yan[2], Yibo Yang[2], Litao Hu[2], Yuchen Liu[2], Yuan Zeng[2], Hanchao Yu[2], Yinan Zhu[2], Sumedha Singla[2], Brian Vanover[2], Huijun Qian[2], Zihao Wang[2], Fujun Liu[2], Aashu Singh[2], Jianyu Wang[2], Xuewen Zhang[2]

[1]Carnegie Mellon University, [2]Meta

Abstract

Generative AI has enabled the creation of photorealistic images and videos that are increasingly disseminated on social media, often used for spam, misinformation, manipulation, and fraud. Existing AI-generated content (AIGC) detection methods face challenges including poor generalization to new generation models, reliance on single modalities, and lack of interpretable explanations. We present our pipeline that mitigates these issues by continuously curating diverse multi-modal social media data and training a compact vision-language model for detection and explanation. Our model achieves state-of-the-art detection performance on public benchmarks and demonstrates robust detection and explanation capabilities on internal social media datasets across multiple platforms. We deployed our model for post recommendation on social media platforms and observed positive downstream impacts on user engagement, demonstrating that it is feasible to perform effective AIGC detection in dynamic, real-world social media environments.

1 Introduction

Generative AI is increasingly capable of generating photorealistic images and videos disseminated over social media (Rombach et al., 2022; Brooks et al., 2024). AI-generated contents are being used to generate spammy clickbait (Forbes Technology Council, 2024), to create and spread misinformation (Associated Press, 2024), to manipulate and influence behaviors (Tarsney, 2025), as well as to harass, scam, and fraud users (Wired, 2024).

This has motivated a large body of work on AI-generated content (AIGC) detection: Dedicated models have been developed to detect deepfakes (Frank et al., 2020), AI-generated arts (Rahman et al., 2023), and lately photorealistic images (Yan et al., 2024). However, the existing approaches generally suffered from three problems: (1) They are trained on static datasets and **do not generalize well** to images generated by newer models (Epstein et al., 2023), (2) they only rely on information from one single modality for detection (e.g., Frank et al., 2020; Yan et al., 2024), and can not utilize rich **multi-modal signals** from social media posts, and (3) they mostly do not provide human-readable explanations, that make their judgments **hard to interpret** for humans.

In this work, we share our approach to detect and explain social media posts with AI-generated content at Meta, aiming to mitigate the above three problems (Section 3). First, we develop a dedicated data curation pipeline, that allows us to continuously curate new high-quality AIGC data from various social media platforms. Second, we adopt the latest developments in multi-modal foundation models (Grattafiori et al., 2024), and train a small yet effective vision language model (IFM-AIGCSPOTTER-3B) that detects and explains social media posts with AI-generated content.

We evaluated our approach through two sets of experiments: We first benchmark our approach against public datasets (Section 4.1), showing our model achieve state-of-the-art performance on par with other open-sourced models. We next evaluate our trained model on an internal dataset (Section 4.2), demonstrating our model's ability to accurately detect AI-generated contents and provide useful explanations. We conducted further ablations and deployed our model for post recommendation at social media platforms. Overall, we demonstrate that it is possible to detect AI-generated content effectively from a large quantity of, diverse, and continuously evolving social media posts.

2 Related Work

A substantial body of work studies AI-generated content (AIGC) detection. Early approaches exploit low-level spatial or frequency artifacts (Mc-

*Work done at Meta.

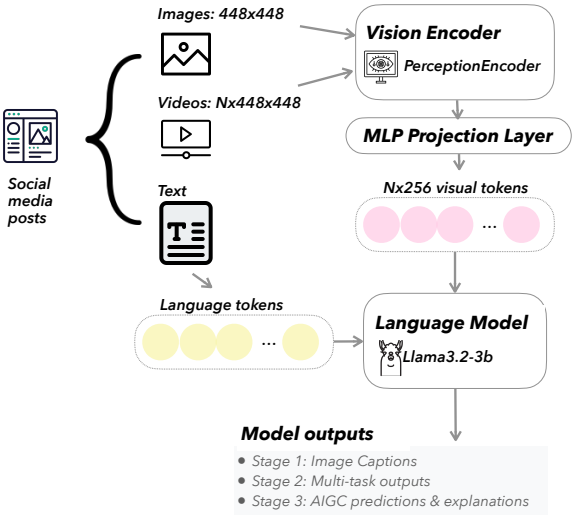

Figure 1: Overview of our LLaVA-based model architecture. Our model processes images and videos with PerceptionEncoder into visual tokens, which are fed into a language model together with language tokens.

Closkey and Albright, 2018; Frank et al., 2020), while lightweight CNNs trained on GAN datasets capture generator-specific cues (Wang et al., 2020; Rossler et al., 2019). The rise of diffusion models has spurred new detectors, including CLIP-based methods that show improved cross-generator generalization (Cozzolino et al., 2024). To cope with rapidly evolving generators, recent work frames detection as a continual-learning problem, studying online or continuous adaptation (Epstein et al., 2023; Tassone et al., 2024).

Vision-language models. Large pretrained VLMs provide a much stronger multimodal prior for AIGC detection. Prior work explores linear probes or prompt tuning on frozen CLIP backbones (Ojha et al., 2023; Keita et al., 2024), in-context prompting of multimodal LLMs (Jia et al., 2024; Ye et al., 2024), and fine-tuning MLLMs to reason about visual semantics (Liu et al., 2024; Wen et al., 2025b).

In-the-wild benchmarks. Recent datasets better reflect real-world distribution shifts. MiRAGe-News (Huang et al., 2024) pairs Midjourney images with synthetic captions, revealing weaknesses on unseen publishers. Deepfake-Eval-2024 (Chandra et al., 2025) and BusterX++ (Wen et al., 2025a) collect recent social-media fakes, causing large performance drops for state-of-the-art detectors. These benchmarks highlight the need for multimodal reasoning and continual adaptation.

Discussion. Most existing detectors are specialized to particular generators or rely solely on visual artifacts, and current benchmarks largely omit social context. Our work addresses these limitations by continuously collecting image-text pairs with social signals and training a unified VLM that leverages both visual and contextual cues for robust, real-time AIGC detection.

3 Method

In this section, we provide a detailed description of our method. We first outline our model architecture, and then present our data curation pipelines and their corresponding model training recipes.

3.1 Model Architecture

We follow the general architecture of LLaVA (Liu et al., 2023), a widely adopted architecture for vision language modeling such as in Qwen2.5-VL (Bai et al., 2025) and InternVL-2.5 (Chen et al., 2024a). The LLaVA architecture has three major components: a Vision Encoder, a Language Model, and an MLP projection layer that connects the Vision Encoder and Language Model, as demonstrated in Figure 1.

Implementation. To ensure high inference efficiency, we choose a model of small size, `Llama3.2-3B-Instruct` (Grattafiori et al., 2024) as our Language Model. For Vision Encoder, we choose Perception Encoder (Bolya et al., 2025) with 300M parameters, as it demonstrates strong zero-shot classification results compared to other vision encoders (Bolya et al., 2025). To improve inference efficiency, we crop the image to a resolution of 448×448, and adopt an adaptive pooling strategy to produce a total of 256 output visual tokens per image. For the projection layer, we use a single linear layer with zero-weight initialization.

3.2 Data Curation

We initialize our model with the pretrained weights and train our model on a wide range of curated image-text data. The data can be categorized into three parts as illustrated in Figure 2:

- **Image Captioning** data to align the image-text modality.
- **Multi-task Fine-tuning** data to improve multimodal content understanding capabilities.
- **AIGC Fine-tuning** data to adapt the model to the downstream AIGC detection and explanation tasks.

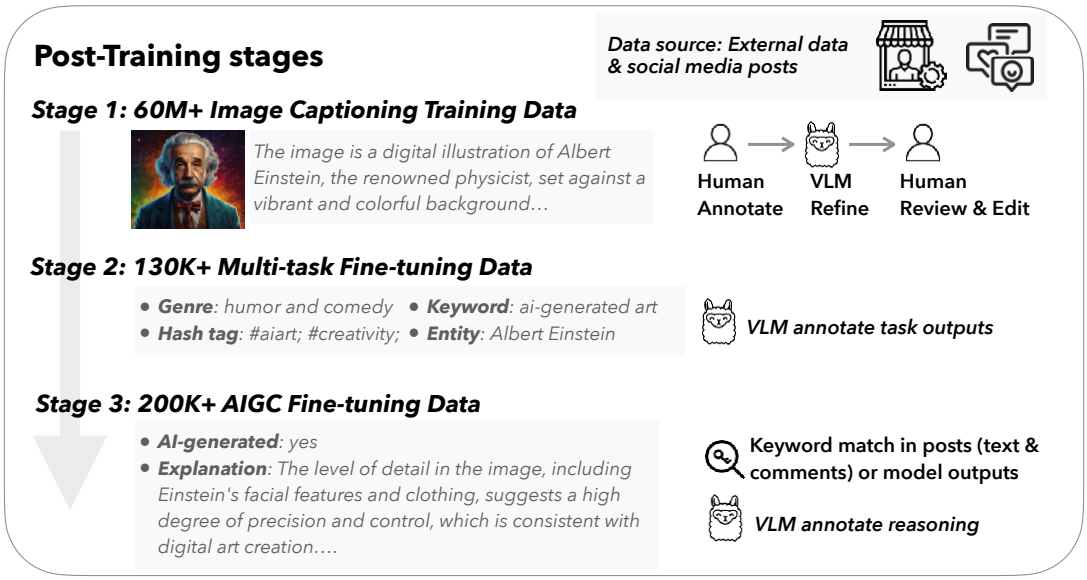

Figure 2: Overview of the data curation and post-training pipeline. The model is initialized with pretrained weights of `Llama3.2-3B-Instruct` (Grattafiori et al., 2024) and Perception Encoder (Bolya et al., 2025). It is further post-trained on three curated data categories: (1) Image Captioning data from Shutterstock and social media, including human-annotated high-quality captions; (2) Multi-task Fine-tuning data for enhanced multimodal understanding with tasks like named entity recognition and hashtag generation; (3) AIGC Fine-tuning data for AI-generated content detection and explanation, curated via heuristic labeling and continuous automated collection.

Image Captioning Data. We collected a total of 60M+ image captioning data from three sources:

- **Shutterstock Split**: We leveraged a substantial subset of the Shutterstock dataset, comprising 60 million images, after rigorous cleaning. The initial round of data cleaning removed low-quality data, including non-ASCII-coded characteristics, repeating words, incomplete sentences, etc. We then used `Llama-3.2-90B` to generate elaborate captions for the filtered data, while providing original captions for grounding.
- **Social Media Split**: We sampled 10M images and 2.2M videos from public posts on social media platforms based on post engagement. We then generated captions for these images with `Llama4-Scout`.
- **High-quality Social Media Split**: For a subset of 164K social media images, we have humans first annotate captions, including visual details, OCR, and a summary. These captions are further refined by `Llama4-Scout` and then reviewed and edited by human annotators.

Multi-task Fine-tuning Data. In this stage, we collected data for various downstream tasks that improve model's multi-modal understanding capabilities. We included tasks like named entity recognition, keyword generation, hashtag generation, and genre classification. We sampled 130K images from the **Social Media Split** and generated task-specific labels with `Llama4-Maverik`. Note that in this stage, the training data often contains signals related to AI-generated content (e.g., in *generated* keywords or hashtags), which bootstraps the model to obtain some initial knowledge for later AIGC detection and explanation.

AIGC Fine-tuning Data. In the final stage, we collected 200K samples for AIGC fine-tuning. We sampled images from social media posts and labeled them using simple heuristics: Posts are labeled as AIGC if their text or comments explicitly mention AIGC-related keywords based on a curated vocabulary (e.g., *"ai-generated," "Midjourney," "AI art"*), or if the multi-task model from the previous stage predicts an AIGC-related keyword or hashtag; all others are labeled as non-AIGC. We further cleaned the data by clustering them with embeddings and keeping the top clusters with human confirmation. This results in 30K AIGC and 170K negative samples, with explanations annotated with `Llama4-Maverik`. This stage runs continuously, enabling data collection and model re-training on

emerging AI-generated content.

3.3 Post-training Recipe

Our post-training pipelines consist of three stages:

- **Stage 1 (Image Captioning)**: This initial stage equips our model with foundational image understanding capabilities.
- **Stage 2 (Multi-task Fine-tuning):** This stage improves our model's multimodal content understanding capabilities and helps prepare data for the final stage.
- **Stage 3 (AIGC Detection Fine-tuning)**: This final stage exclusively focuses on the ability to detect and explain AIGC.

Stage 1 (Image Captioning). In the first stage, our goal is to equip our model with foundational image understanding capabilities by aligning the text-image modality. We first freeze both Language Model and Vision Encoder, and exclusively train the MLP projection layer with Shutterstock Split of image captioning data. We use an effective batch size of 2048 and a learning rate of 0.002 using a Cosine Scheduler with 200 warm-up steps.

We then unfreeze the Vision Encoder and train on the **Social Media Split**. Finally, we train the model end-to-end with the small **High-quality Social Media Split**. We use an effective batch size of 128 and 64 respectively for the last two parts of Stage 1 training, and a learning rate of 4e-5.

Stage 2 (Multi-task Fine-tuning). The goal of the second stage is to improve model's multi-modal content understanding capabilities. We fine-tuned the model from stage 1 on the curated **Multi-task Fine-tuning Data**. We use an effective batch size of 256, a learning rate of 4e-5, and Adam 8-bit for optimization.

Stage 3 (AIGC Detection Fine-tuning). The goal of the final stage is to improve our model's capability to detect and explain posts with AI-generated content. We performed standard supervised fine-tuning on the curated **AIGC Fine-tuning Data**. During the fine-tuning, we randomly drop 90% of the text information, such that the trained model does not solely rely on the textual clues but also learns from related features in the images.

Discussion. Our training pipeline produces a designated model, IFM-AIGCSPOTTER-3B, that understands multi-modal social media posts well. Different from existing vision language models (Grattafiori et al., 2024; Bai et al., 2025), IFM-AIGCSPOTTER-3B is exposed to a wide range of images posted on social media, which provides strong signals of the latest trend of AIGC, laying a foundation for detection and explanation.

4 Experiments

4.1 Detecting and Explaining AI-generated Images in Benchmarks

We first evaluate our approach by benchmarking on public datasets and comparing it against a wide range of closed-sourced and open-sourced models.

4.1.1 Setups

To ensure a fair comparison, we use the model checkpoint from Stage 2 before dedicated AIGC fine-tuning. We then follow the setup of prior work (Wen et al., 2025b), fine-tuning models on the same training dataset, FakeClue (Wen et al., 2025b), and evaluating them on FakeClue's test split (in-distribution), as well as an independent test set LOKI (Ye et al., 2024) (out-of-distribution). The goal here is to understand how well different models can learn to detect AIGC, and how well they generalize to different AIGC datasets.

Baselines. We compare against three baselines: (1) seven VLMs with zero-shot prompting (Wu et al., 2024; Chen et al., 2024b; Wang et al., 2024) with reported results in prior work (Wen et al., 2025b), (2) three open-source VLMs (Qwen2.5-VL-3b (Bai et al., 2025), Gemma3-4b, `Llama3.2-11b` (Grattafiori et al., 2024)) fine-tuned with and without explanations, and (3) prior SOTA with explanation, FakeVLM (Wen et al., 2025b).

Datasets. We evaluate all approaches on two datasets: The test split of FakeClue (Wen et al., 2025b) and an independent test set LOKI (Ye et al., 2024), to test the approaches' abilities to detect AIGC both in in-distribution and out-of-distribution settings. The datasets cover a wide range of AIGC, from deepfakes, synthesized images of animals, humans, objects, and scenes, to AI-generated documents and satellite images.

We report accuracy and F1 scores on both datasets. Additional ablation results can be found at Appendix B.

4.1.2 Results

Overall, we found IFM-AIGCSPOTTER-3B among the best-performing models to detect AI-generated content (Table 1), with an accuracy of

Method	FakeClue (in-distribution)		LOKI (out-of-distribution)	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑
Prompting only				
Deepseek-VL2-small	0.404	0.542	0.253	0.387
Deepseek-VL2	0.475	0.541	0.431	0.392
InternVL2-8B	0.506	0.490	0.526	0.340
InternVL2-40B	0.507	0.463	0.507	0.376
Qwen2-VL-7B	0.457	0.592	0.571	0.350
Qwen2-VL-72B	0.578	0.565	0.554	0.409
GPT-4o	0.474	0.420	0.634	0.572
Detection-only training				
Qwen2.5-VL-3b	0.950	0.962	0.767	0.809
Gemma3-4b-it	0.946	0.958	0.748	0.810
Llama3.2-11b	0.956	0.966	0.802	0.848
IFM-AIGCSPOTTER-3B	**0.986**	**0.989**	**0.839**	**0.870**
Detection + Explanation training				
Qwen2.5-VL-3b	0.932	0.948	0.739	0.792
Gemma3-4b	0.937	0.951	0.784	0.824
Llama3.2-11b	0.951	0.962	0.789	**0.846**
FakeVLM (11b)	**0.986**	**0.981**	**0.843**	0.837
IFM-AIGCSPOTTER-3B	0.971	**0.977**	0.816	**0.844**

Table 1: Performance comparison of various models on FakeClue and LOKI datasets. IFM-AIGCSPOTTER-3B achieves top detection accuracy (0.986 on FakeClue and 0.839 on LOKI) among fine-tuned models, demonstrating strong generalization especially in out-of-distribution settings. We observe that incorporating explanation generation during training slightly reduces detection accuracy, suggesting a trade-off between explainability and performance.

0.986 on FakeClue and an accuracy of 0.839 on LOKI (out-of-distribution).

Breaking down the results, we first found that **existing VLMs struggle to detect AIGC correctly without specialized fine-tuning**, with the best performing model achieving < 0.6 accuracy on the FakeClue dataset.

Second, comparing different fine-tuned models, we found **fine-tuned models with a small size are able to detect AIGC reliably**. Specifically, we found IFM-AIGCSPOTTER-3B demonstrates better generalizations to LOKI under both training schemes – it is particularly effective for images with persons (+5.4%), animals (+6.2%), and scenes (+3.5%), but less so for documents (-0.3%). We believe this largely reflects our model's training data, where documents appear much less frequently compared to persons, animals, or scenery.

Third, we found adding explanation to the training process slightly compromises detection quality, which contradicts with findings in prior work (Wen et al., 2025b). We hypothesize that, adding explanation generation during training introduces a trade-off in model capacity and optimization, which lead the model to allocate resources to generating explanations at the expense of optimizing detection accuracy. However, we consider this is a trade-off that is potentially worth making, to introduce better explainability and user trust.

4.2 Detecting AI-generated Content on Social Media

Next, we evaluate IFM-AIGCSPOTTER-3B's ability to detect AIGC on real-world social media posts. We use the model checkpoint from stage 3 for this evaluation. Additional experiments on exploring the model's reasoning capabilities are reported in Appendix C. We also share more detailed error analysis in Appendix D.

4.2.1 Internal Evaluation

We curate test data using the same pipeline as in Section 3 and report results by platform. We compare against two baselines: (1) an embedding-based MLP trained on visual embeddings, and (2) a lightweight adapter variant with the backbone frozen. In addition, we run ablations to isolate key design choices: (1) removing random text drop, (2) swapping the Perception Encoder for a CLIP-based vision backbone, and (3) using 4- and 8-frame videos instead of the default 1-frame input.

Results. Overall, we found our model can reliably detect AIGC across all three platforms we look at (Table 3), and significantly improve over baselines (Table 6).

For ablations, we found that random text drop

Dataset	CNNSpot	FreDect	Fusing	GramNet	LNP	UnivFD	DIRE	PatchCraft	NPR	AIDE	Ours
Chameleon	60.11	56.86	57.07	60.95	55.63	55.62	59.71	56.32	58.13	62.60	**86.00**

Table 2: Performance comparison across different detection methods on the Chameleon dataset. Our method significantly outperforms existing baselines without adaptation.

Platform	Precision	Recall
Facebook	0.888	0.818
Instagram	0.883	0.835
Threads	0.926	0.919
Overall	0.886	0.853

Table 3: Performance of our model in detecting AI-generated content (AIGC) across three platforms. These results demonstrate the model's reliable detection capability across diverse social media environments.

is crucial for recall, as models will over-rely on the textual clues otherwise. We also found the Perception Encoder offers a slight precision advantage over CLIP vision encoder, and observed significant improvement by extending the inputs to multi-frame (Table 6).

4.2.2 Generalizations to Benchmarks

We further evaluate our trained model on public benchmarks *without* further adaptation. We choose Chameleon (Yan et al., 2024) as it contains highly realistic AIGC images commonly observed on social media. We compare our method to the reported results of 10 existing methods (Yan et al., 2024).

Results. We found our model achieves 86.0% accuracy (Table 2), outperforming all baselines by a large margin (>20%). This result demonstrates that our data curation pipeline is high-quality and is able to train detectors that reliably detect realistic AIGC images without dedicated adaptation.

4.3 Deployment in Production

We deployed the model with 4 Nvidia H100 nodes. It will trigger all of the posts with views > 500 (millions per day) for global users. We aggregated the results per content creator and identified the ones with excessively AIGC posts. We filtered out the creators with high engagement and demoted the remaining ones for new users and marginal users. We conducted standard online A/B testing, with the model deployed for the treatment group.

Results. After deploying the model for 6 weeks, in the 7-day backtest we noticed there is a statistically significant views gain for marginal users (+0.21%). For global, the launch achieved +0.22% Young-Adult session gains. This demonstrates that, our deployment strategy successfully improves users' experience on social media, resulting in higher user engagements.

5 Discussion

Our results demonstrate that our approach achieves state-of-the-art performance in detecting AI-generated content. We next discuss practical implications of our work and future challenges.

Mitigating harmful content and preserving platform integrity. AI-generated media increases the risk of deceptive and harmful content, such as deepfakes and misinformation that are difficult for users to identify (e.g. Associated Press, 2024). These materials can enable harassment, fraud, and manipulation, with real-world consequences for public safety and democratic processes (e.g. Wired, 2024). At the same time, widespread generation risks saturating social feeds with synthetic content, crowding out authentic voices. Automated AIGC detection supports *early intervention*, enabling platforms to flag or demote harmful media while preserving content diversity and platform integrity.

Detecting realistic AIGC. A challenge in AIGC detection lies in identifying highly realistic AI-generated content that closely mimics authentic content. As generative models improve, synthetic images and videos increasingly exhibit surface coherence that evade existing detectors (Google DeepMind, 2024). This underscores the need for evolving detection systems that go beyond shallow cues, instead leveraging semantic and contextual information for detection.

6 Conclusion

We propose a unified framework for detecting and explaining AI-generated content (AIGC) on social media. Our approach continuously curates in-the-wild multimodal data, leverages an efficient vision-language model for robust detection. Extensive experiments on public benchmarks and internal datasets show state-of-the-art detection perfor-

mance, and large-scale deployment demonstrates tangible gains in user engagement, highlighting the practical impact of our system.

Limitations

Despite strong empirical performance, our work has several limitations. First, while our continuous data collection pipeline mitigates distribution shift, it remains reactive to emerging generators and may lag behind new models or adversarial attacks specifically designed to evade detection. Second, although we train a small vision-language model for efficient inference, large-scale deployment still incurs nontrivial computational costs. Finally, the generated textual explanations aim to improve interpretability but do not guarantee faithful causal attribution, and may occasionally reflect model biases or overconfidence.

References

Associated Press. 2024. Trump arrested? putin jailed? fake ai images spread online.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, and 1 others. 2025. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators.

Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, and 1 others. 2025. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. 2024. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366.

David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. 2023. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 382–392.

Forbes Technology Council. 2024. The rise of advanced clickbait: How ai is tricking unsuspecting users.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR.

Google DeepMind. 2024. Veo.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Runsheng Huang, Liam Dugan, Yue Yang, and Chris Callison-Burch. 2024. Miragenews: Multimodal realistic ai-generated news detection. *arXiv preprint arXiv:2410.09045*.

Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. 2024. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4324–4333.

Mamadou Keita, Wassim Hamidouche, Hessen Bougueffa Eutamene, Abdelmalik Taleb-Ahmed, and Abdenour Hadid. 2024. Fidavl: Fake image detection and attribution using vision-language model. In *International Conference on Pattern Recognition*, pages 160–176. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Jiawei Liu, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. 2024. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*.

Scott McCloskey and Michael Albright. 2018. Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489.

Md Awsafur Rahman, Bishmoy Paul, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, and Shaikh Anowarul Fattah. 2023. Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2200–2204. IEEE.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Christian Tarsney. 2025. Deception and manipulation in generative ai. *Philosophical Studies*, pages 1–23.

Francesco Tassone, Luca Maiano, and Irene Amerini. 2024. Continuous fake media detection: adapting deepfake detectors to new generative techniques. *Computer Vision and Image Understanding*, 249:104143.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704.

Haiquan Wen, Tianxiao Li, Zhenglin Huang, Yiwei He, and Guangliang Cheng. 2025a. Busterx++: Towards unified cross-modal ai-generated content detection and explanation with mllm. *arXiv preprint arXiv:2507.14632*.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. 2025b. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *Preprint*, arXiv:2503.14905.

Wired. 2024. Deepfake scams are distorting reality itself. https://www.wired.com/story/youre-not-ready-for-ai-powered-scams/.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2024. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, and 1 others. 2024. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. *arXiv preprint arXiv:2410.09732*.

A Prompt for AIGC Detection

Image-level detection Prompt:
<image>
Does the image look real/fake?

Post-level detection Prompt:
You are an Assistant to answer questions from users. Your task is to evaluate the likelihood of a post being AI-generated. Specially, pay attention to the image included in the post.

<image>

BODY TEXT: {body text}
OCR TEXT: {ocr text}
TITLE TEXT: {title text}
VIDEO TRANSCRIPT: {truncated video transcript}

Given the post information, determine if the post is AI-generated or not. Answer in "yes" or "no".

Detection + explanation prompt:
You are an assistant designed to analyze images for potential AI-generated content by examining semantic clues, stylistic clues, and local visual artifacts.

<image>
Analysis Steps

1. **Semantic Clues Analysis**

- **Objective**: Identify dramatic or unlikely content that suggests the use of an AI generator.
- **Output**: Provide a caption and reasoning for the scene, highlighting any elements that appear improbable or exaggerated.

2. **Stylistic Clues Analysis**

- **Objective**: Detect waxy styles or overly high-quality features typical of generated images. Consider color, texture, and lighting.
- **Output**: Offer a description of the image's style, noting any characteristics that deviate from natural or expected artistic styles.

3. **Local Visual Artifacts Analysis**

- **Objective**: Identify imperfections and artifacts that are uncommon in real photos or traditional art. Consider physical artifacts (e.g., optical display issues, violations of physical laws, and spatial/perspective errors), structural artifacts (e.g., deformed objects, asymmetry, or distorted text), and distortion artifacts (e.g., color/texture distortion, noise/blur, artistic style errors, and material misrepresentation).
- **Output**: Detail any local visual artifacts, explaining their nature and why they suggest synthetic generation.

4. **Conclusion**

- After conducting the three-step analysis, you will provide a final assessment of the image's authenticity, considering the findings from each step.

Answer in the following format:

<semantic>The image shows... </semantic>
<stylistic>The image is... </stylistic>
<local>The image has... </local>
<conclusion>AI-generated: Yes.</conclusion>

Method	FakeClue		LOKI	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑
IFM-AIGCSPOTTER-3B	0.929	0.945	0.793	0.833
+ CLIP vision encoder	0.902	0.925	0.803	0.850
+ Larger input size	0.967	0.974	0.833	0.872
+ More visual tokens	**0.973**	**0.979**	**0.859**	**0.884**

Table 4: Impact of Perception Encoder and vision budget allocation on IFM-AIGCSPOTTER-3B's performance *when training for 1 epoch*. The Perception Encoder improves in-distribution accuracy compared to the smaller CLIP vision encoder but generalizes slightly worse on the LOKI dataset. Increasing vision input size or visual tokens significantly boosts performance.

Method	FakeClue		LOKI	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑
IFM-AIGCSPOTTER-3B, LM loss	0.986	0.989	0.839	0.870
IFM-AIGCSPOTTER-3B, BCE loss	0.991	0.993	0.829	0.863
IFM-AIGCSPOTTER-3B, focal loss	**0.991**	**0.993**	**0.845**	**0.875**

Table 5: Performance of IFM-AIGCSPOTTER-3B variants on FakeClue and LOKI datasets. BCE and focal loss variants show slightly improved performance compared to the standard SFT loss.

B Ablation study

B.1 Ablations on Vision Encoders

To understand the effectiveness of our vision encoder, we conducted ablation analysis comparing different variants:

- **CLIP vision encoder**: We replaced the vision encoder with a CLIP vision encoder trained on internal data.
- **Larger input size**: We changed the input size from 448×448 to 896×896.
- **More visual tokens**: We changed the number of visual tokens from 256 to 1024.

We followed the same training and evaluation setup as before and trained each variant for only 1 epoch.

Results. We first found that Perception Encoder does improve IFM-AIGCSPOTTER-3B's performance under in-distribution setting, compared to the smaller CLIP vision encoder, though it generalizes slightly worse to the LOKI dataset (Table 4).

Next, we found that allocating more budgets for the vision components – either supporting larger input size or using more visual tokens – significantly improves model performance when training for 1 epoch, indicating a trade-off between higher compute cost and fast training convergence.

B.2 Ablations on Supervision Loss

To evaluate the impact of different loss functions on model performance beyond standard SFT Language Modeling loss, we conducted ablations using the following loss variants:

- **BCE loss**: The standard binary cross-entropy loss for binary classification. We swap the language modeling head with a classification head.
- **Focal loss**: A modified version of BCE loss that down-weights easy examples and focuses training on hard examples. This is intended to improve model robustness and performance on imbalanced data.

We followed the same setups as before and trained each variant for two epochs.

Results. We found that switching to the classification head can slightly improve IFM-AIGCSPOTTER-3B's AIGC detection performance (Table 5), with focal loss further improving model's performance on out-of-distribution settings where label imbalances differ. This indicates that, for pure detection task without explanation or reasoning, using a classification head with focal loss will achieve the best result.

B.3 Ablations on Visual Backbones and Data

To better understand the contributions of different design choices, we evaluate several ablation setups:

- A variant without random text drop, testing whether having full access to text inputs affects performance.
- A variant where the Perception Encoder is replaced by a CLIP-based vision encoder, testing sensitivity to the choice of vision backbone.
- Variants that process 4-frame and 8-frame video inputs instead of the default setting (1-frame), testing extensions to multi-frame data.

Method	Precision	Recall
IFM-AIGCSPOTTER-3B	0.886	0.853
Embedding-based model	0.663	0.773
Adapter-based model	0.883	0.801
Ablations of IFM-AIGCSPOTTER-3B		
No text drop	0.956	0.500
CLIP vision encoder	0.876	0.854
4-frame video inputs	0.898	0.861
8-frame video inputs	0.904	0.849

Table 6: Comparison of precision and recall for different variants of IFM-AIGCSPOTTER-3B in detecting AI-generated content. Removing the random text drop drastically reduces recall to 0.500, indicating that models tend to infer based only on text content without text drop. Replacing the Perception Encoder with the CLIP vision encoder results in slightly lower precision (0.876) but comparable recall (0.854).

C Reasoning about AI-generated Content on Social Media

We then explore IFM-AIGCSPOTTER-3B's ability to reason about AI-generated content on social media. This is to understand how well IFM-AIGCSPOTTER-3B can generate plausible reasoning traces before making a final judgment.

Setups. We first curated 20k reasoning traces from `Llama4-Maverick` with a prompt (Appendix A) that requires models to provide analysis on image semantics, style, and local visual artifacts (Figure 3). We then fine-tuned IFM-AIGCSPOTTER-3B with the 20k reasoning-enhanced data.

To increase input coverage, we performed rejection sampling with this fine-tuned checkpoint on our curated **AIGC Fine-tuning** dataset, with 8 samples per input. We curated 61k pairwise preference data and 61K SFT data on hard examples this way, by verifying each completion's final conclusion against the label ground truth. We further curated 191k SFT data by collecting all examples with at least one correct response. We explored various post-training methods (rejection sampling finetuning, direct preference optimization, group relative policy optimization) on this dataset.

Results. Overall, we found IFM-AIGCSPOTTER-3B achieves strong performance in detecting AI-generated content while generating plausible reasoning traces (Table 7). Breaking down the results, we observed the following:

First, we found that **reasoning-enhanced fine-tuning alone yields decent performance**, with a precision of 0.639 and a recall of 0.756. This can be further improved by SFT with rejection sampling data, which expose the model to more diverse samples. However, gaps remain between reasoning and non-reasoning methods, indicating a trade-off for generating human-interpretable traces.

Second, we found that **preference alignment methods such as DPO and IPO show mixed results** – models often exhibit serious reward hacking issues with, leading to worse format following and increased repetition that produce malformed outputs.

Third, we found that larger models like `Llama3.2-11b` do not yield better results, similar to what we have observed in Section 4.1, likely due to their limited exposure to social media data.

C.1 Reasoning faithfulness

We conducted an exploratory analysis to estimate the reasoning error rate. We mark the model-provided reasoning as correct if *at least one* of the multiple reasons generated for a sample is valid, since the model typically outputs several supporting reasons per example.

Results. For images, the reasoning error rate is below 5%; the dominant failure mode is not missing genuine artifacts, but hallucinating additional ones (e.g., attributing "unnatural textures" to artifact-free regions). For videos, the reasoning error rate is higher (above 10%), largely due to frame sampling limitations that amplify hallucinations on fast-paced content.

Examples. We share a few anonymized examples (images omitted due to privacy concerns) to demonstrate when we observe the reasoning to be (not) faithful:

- **Detecting unrealistic synthetic scene**: *"The entire scene appears generated from scratch. The dog is anthropomorphized (standing at a bank counter, talking), and the human character has the characteristic smooth, slightly plastic look of AI-generated humans."* (Faithful)
- **Detecting garbled text**: *"The images exhibit definitive markers of AI generation. Most notably, whenever text is supposed to be visible within the images themselves (such as the pages of the Bible or the scroll), it is completely garbled, nonsensical pseudo-script. There are also minor anatomical inconsistencies in the hands across various images."* (Faithful)

Method	Valid Rate	Precision	Recall	Accuracy
SFT on binary classification	-	0.889	0.785	0.949
SFT on 20k llama4-annotated data	0.995	0.639	0.756	0.890
Preference Alignment				
llama4-annotated data + 61K preference data (DPO)	0.568	-	-	-
llama4-annotated data + 61K preference data (IPO)	0.895	0.539	0.732	0.861
SFT with rejection sampling				
llama4-annotated data + 61K rejection sampling data	0.997	0.750	0.698	0.912
191K rejection sampling data	0.997	0.745	0.687	0.910
191K rejection sampling data, llama3.2-11b	0.995	0.695	0.558	0.888

Table 7: Evaluation results of various training and alignment methods on 200k examples for AI-generated content detection. Reasoning-enhanced fine-tuning alone yields decent performance and can be further improved by supervised fine-tuning (SFT) with rejection sampling data. Despite improvements, gaps remain between reasoning and non-reasoning methods, indicating a trade-off between detection accuracy and generating human-interpretable reasoning traces.

Post Image

AIGC Reasoning + Detection

<semantic>The image shows a serene and well-maintained courtyard with a large mosque in the background, featuring a tall minaret and a large dome. The scene is peaceful, with a small pond in t foreground and lush greenery surrounding the building. The architecture is ornate, with intricate designs and patterns on the walls and domes. The overall atmosphere is one of tranquility and religious significance.</semantic>
<stylistic>The image is rendered in a highly realistic style, with detailed textures and lighting that gi it a lifelike quality. The colors are vibrant and well-balanced, with a focus on greens and blues that evoke a sense of calmness. The overall aesthetic is consistent with high-quality architectural photography, suggesting ***a level of precision and attention to detail that is characteristic of AI-generated content****.</stylistic>*
<local>The image has a few subtle inconsistencies in the ***reflection on the water's surface and th lighting on the building's facade****, which could be indicative of digital manipulation. However, thes are not overtly jarring and could also be attributed to the skillful use of lighting and reflection in the image. The image is generally free of obvious artifacts, such as distorted text or deformed objects, b the perfection of the scene and the lack of any visible signs of wear or aging on the building sugges a level of digital enhancement.</local>*
<conclusion>AI-generated: Yes.</conclusion>

Figure 3: Example of AI-generated reasoning trace for AIGC detection. The model provides semantic, stylistic, and local analyses. This structured approach demonstrates the model's ability to generate plausible, human-interpretable reasoning when identifying AI-generated content.

- **Hallucinating about "unnatural motions"**: *However, there are subtle signs of AI-generated motion: the woman's hand movements show slight morphing and instability — her fingers appear to blend or stretch unnaturally as she gestures.* (Unfaithful)

D Error analysis

We summarize common error patterns observed during development and evaluation, along with their likely causes and mitigation.

Over-reliance on textual cues. A frequent false positive occurs when the text contains surface-level "AI" indicators (e.g., explicit mentions of AI tools, prompts, or model names) while the content itself is human-written. This suggests the model can shortcut by keying on lexical cues rather than grounded artifacts. To reduce this behavior, we apply aggressive text dropout (90%) during training, forcing the model to rely more on visual and multimodal signals.

Out-of-distribution categories. The model degrades on categories that differ substantially from the training distribution, such as screenshots of PDFs. Errors here are primarily false negatives: subtle AIGC artifacts are missed when the input departs from common social-media-style images. A practical takeaway is that deploying the model beyond the intended domain should be paired with domain-specific data curation (e.g., document-centric AIGC examples) and/or targeted adaptation.

Large images with partially AI-generated regions. Another recurring failure mode involves high-resolution images that are mostly real but contain a small AI-generated region (e.g., a document image with an inserted AIGC illustration). When the manipulated region is spatially small, a global representation can underweight it, leading to false negatives. We partially mitigate this by allocating more compute budget to the vision components (e.g., higher-resolution features / more visual tokens), which improves sensitivity to localized artifacts.

Videos with partially AI-generated frames. We also observe failures on video inputs where only a subset of frames are AI-generated. When evidence is temporally sparse, frame sampling and temporal aggregation can wash out the signal, yielding false negatives. This suggests that robust video detection likely requires adaptive sampling strategies that increase coverage around suspicious segments.