AIGUARD: A Benchmark and Lightweight Detection for E-commerce AIGC Risks

Wenhua Zhang[1]*, Weicheng Li[1]*, Xuanrong Rao[1]*, Lixin Zou[1]†, Xiangyang Luo[2], Chubin Zhuang[3], Yongjie Hong[3], Zhen Qin[3], Hengyu Chang[3], Chenliang Li[1], Bo Zheng[3]†

[1]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education,
School of Cyber Science and Engineering, Wuhan University
[2]State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China
[3]Taobao & Tmall Group of Alibaba, China
{wenhuazhang_ncc, liweicheng, raoxuanrong, zoulixin}@whu.edu.cn, xiangyangluo@126.com,
{bryant.zcb, hongyongjie.hyj, qinzhen.qz, hongbo}@taobao.com,
cllee@whu.edu.cn, bozheng@alibaba-inc.com

Abstract

Recent advancements in AI-generated content (AIGC) have heightened concerns about harmful outputs, such as misinformation and malicious misuse. Existing detection methods face two key limitations: (1) lacking real-world AIGC scenarios and corresponding risk datasets, and (2) both traditional and multimodal large language models (MLLMs) struggle to detect risks in AIGC. Towards this end, we introduce AIGUARD, the first benchmark for AIGC risk detection in real-world e-commerce. It includes 253,420 image-text pairs (*i.e.,* the risk content and risk description) across four critical categories: *abnormal body*, *violating physical laws*, *misleading or illogical context*, and *harmful or problematic message*. To effectively detect these risks, we propose distilling text annotations into dense soft prompts and identifying risk content through image soft prompt matching during inference. Experiments on the benchmark show that this method achieves a 9.68% higher recall than leading multimodal models while using only 25% of the training resources and improving inference speed by 37.8 times. For further research, our benchmark and code are available at https://github.com/wenh-zhang/aiguard-dataset.

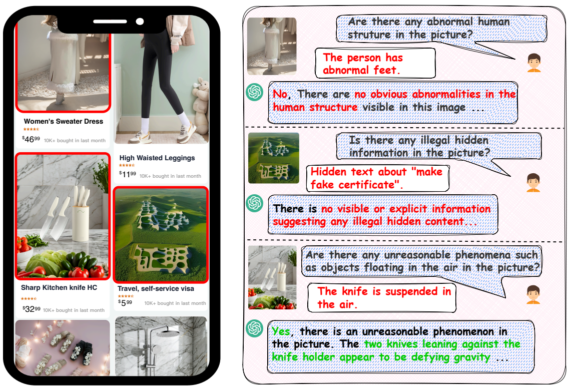
Women's Sweater Dress
$46^99 10K+ bought in last month
High Waisted Leggings
$11^99 10K+ bought in last month
Sharp Kitchen knife HC
$32^99 10K+ bought in last month
Travel, self-service visa
$5^99 10K+ bought in last month
Are there any abnormal human struture in the picture?
The person has abnormal feet.
No, There are no obvious abnormalities in the human structure visible in this image ...
Is there any illegal hidden information in the picture?
Hidden text about "make fake certificate".
There is no visible or explicit information suggesting any illegal hidden content...
Are there any unreasonable phenomena such as objects floating in the air in the picture?
The knife is suspended in the air.
Yes, there is an unreasonable phenomenon in the picture. The two knives leaning against the knife holder appear to be defying gravity ...

Figure 1: Online E-commerce platforms with the risky AIGC content, *e.g.,* "Abnormal foot" or "Make fake certificate", that evade the system detection. However, current MLLMs lack knowledge and have limited capability to detect the risk of AIGC content.

1 Introduction

Recent advancements in AIGC have significantly improved creative workflows in text (Zhang et al., 2024; Achiam et al., 2023), image (Saharia et al., 2022; Koh et al., 2024), and video generation (Blattmann et al., 2023; Liu et al., 2024c), demonstrating substantial commercial potential. According to the AI index report (Clark and Perrault, 2024), in 2024, 42% of surveyed companies reduced their operating costs due to AI technology and 59% increased their revenue. For example, platforms like Google Ads[1] and Alimama[2] now use AI-power tools to automate creative processes that previously required weeks of human effort.

While these systems offer significant advantages, the associated risks require careful attention. On one hand, the inherent randomness of generative models can lead to outputs such as hallucinations (Ji et al., 2023; Li et al., 2023b) or toxic content (Wen et al., 2023; Smith et al., 2022), which can undermine reliability and erode user trust. On the other hand, these systems also pose the risk of being misused in illegal domains. For example, malicious users might exploit these systems to

*First three authors contributed equally.
†Corresponding author.

[1] https://ads.google.com/home/
[2] https://www.alimama.com/index.htm

Findings of the Association for Computational Linguistics: ACL 2025, pages 12437–12450
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

generate risky content that evades detection, such as biased materials or illegal items for criminal activities like selling counterfeit or regulated products (Nadeem et al., 2021; Smith et al., 2022; Qu et al., 2023). As shown in Figure 1, real-world online e-commerce systems might contain AIGC risky content, such as “abnormal foot” and “make fake certificate”, which is too sophisticated to be detected by the system. These issues pose significant threats to user safety, create legal risks for platforms, thus demanding immediate attention.

The solutions to this challenge can be divided into two directions. The first focus is on **detecting real-world risky content**, such as pornography, hate speech, or sensitive political material (Pavlopoulos et al., 2020; Ratkiewicz et al., 2011; Clarke et al., 2023). These approaches for this detection have evolved from rule-based methods (Warner and Hirschberg, 2012; Gitari et al., 2015), to deep learning (Gambäck and Sikdar, 2017; Markov et al., 2023), and now leverage pre-trained models and large language models (Cohen et al., 2023; Pan et al., 2023). However, AIGC risks, such as disproportion or object suspension, are more complex than risks like pornography or hate speech. Detecting these subtle issues requires more world knowledge of MLLMs.

With the development of large language models (LLMs), the second focus is on **controllable generation**, which aims to align models with human preferences and ethical guidelines. Due to the difficulty of annotating training data, these methods primarily rely on reinforcement learning methods, such as reinforcement learning from human feedback (Ouyang et al., 2022), process reward modeling (Lightman et al.), and group relative policy optimization (Shao et al., 2024; Mu et al., 2024). Consequently, the heavy computational demands of reinforcement learning limit their adaptability to intentional misuse by malicious users.

The weakness of existing methods is primarily due to limitations in available datasets. Existing datasets often concentrate on specific model safety issues, such as evaluating hallucinations (Hartvigsen et al., 2022; Wang et al., 2023a; Li et al., 2023a) or detecting toxic content (Shen et al., 2025; Podolak et al., 2024; Tang et al., 2025), while overlooking sophisticated risks, such as hidden illegal messages or disharmonious background. This narrow focus weakens detection methods. Additionally, MLLMs excel at understanding real-world content but struggle to recognize risky AIGC outputs, as illustrated in Figure 1, where models like Qwen2-VL-7B fail to detect such risks. This limitation arises because MLLMs are primarily trained on standard real-world data and lack exposure to risky or adversarial AIGC examples (Schuhmann et al., 2022).

To address these challenges, we introduce AIGUARD, the first comprehensive benchmark for detecting risks in AIGC within real-world e-commerce scenarios. Our dataset comprises real-world adversarial examples and industrial risks (*e.g.,* product flaws), accompanied by expert annotations and detailed risk descriptions. It includes 253,420 image-text pairs, with text descriptions categorizing risks into four critical types: *abnormal body*, *violating physical laws*, *misleading or illogical context*, and *harmful or problematic message*. We also propose a lightweight detection method based on the pre-trained BLIP model (Li et al., 2022). Risk detection is optimized by distilling human annotations into soft prompts through image soft prompt matching and causal risk decoding tasks. During inference, risks are identified by matching images with the soft prompts, achieving high accuracy at minimal computational cost. This approach enables efficient detection of AIGC risks, conserving computational resources in real-world e-commerce applications.

The contributions are summarized as follows:

- Introduce AIGUARD, the first comprehensive AIGC risk detection benchmark, compiling a dataset of 253,420 image-text pairs covering four critical risk categories (*abnormal body, violating physical laws, misleading or illogical context, harmful or problematic message*).
- Propose a lightweight detection method using a pre-trained BLIP model with the soft prompts, achieving high accuracy via image soft prompt matching while minimizing computational overhead for real-world applications.
- Conduct extensive experiments on the benchmark, identifying key challenges and highlighting critical research problems that merit further systematic investigation.

2 Related Work

2.1 Risk Detection Benchmarks

Prior research on risk datasets has primarily focused on text-based risks, such as pornogra-

Benchmark	Task	Risk Types			Size
		Toxic	Hallulation	Illegal	
FELM (Zhao et al., 2023)	Factuality Evaluation		✓		847
ToxiGen (Hartvigsen et al., 2022)	Hate Speech Detection	✓			274,186
HaluEval (Li et al., 2023a)	Hallucination Recognition		✓		30,000
CHIFRAUD (Tang et al., 2025)	Fraud Text Detection	✓		✓	411,934
MHaluBench (Chen et al., 2024)	Multimodal Hallucination Detection		✓		1,860
M-HalDetect (Gunjal et al., 2024)	Multimodal Hallucination Detection		✓		4,000
MM-safetybench (Liu et al., 2024b)	Safety-critical Evaluation	✓			5,040
AIGUARD (Ours)	Multimodel Risk Detection	✓	✓	✓	253,420

Table 1: Comparison of AIGC risk detection benchmarks.

phy detection (Pavlopoulos et al., 2020), fraud identification (Tang et al., 2025), and politically sensitive content (Ratkiewicz et al., 2011). With the rise of AIGC, datasets evaluating risks in AI-generated text, such as factual hallucinations (Zhao et al., 2023; Li et al., 2023a) and toxic outputs like hate speech (Hartvigsen et al., 2022), have gained prominence. For example, FELM (Zhao et al., 2023) assesses factual accuracy across domains (*e.g.,* math, reasoning), while ToxiGen (Hartvigsen et al., 2022) catalogs toxic/benign statements targeting 13 minority groups. Recent work has extended to multimodal tasks, exploring hallucination and toxicity in image-text contexts. New benchmarks aim to evaluate hallucination/toxicity severity (Liu et al., 2024b; Ying et al., 2024; Li et al., 2023b; Wang et al., 2023b) or detector performance (Chen et al., 2024). Examples include MM-safetybench (Liu et al., 2024b), which classifies multimodal toxicity risks, and MHaluBench (Chen et al., 2024), a multi-task hallucination detector benchmark spanning three modalities.

However, current research often concentrates on analyzing text in isolation or addressing a single type of risk (*e.g.,* toxic outputs, hallucinations). There is a notable lack of exploration into composite risk data derived from real-world scenarios. We compare the recent risk detection benchmarks with AIGUARD in Table 1.

2.2 Risk Detection and Model Alignment

The approaches to risk detection have primarily evolved alongside the development of deep learning. Early solutions rely on rule-based methods, such as template-based strategies (Warner and Hirschberg, 2012) or syntactic features (Gitari et al., 2015), which often lack generalization ability. Subsequently, deep learning-based methods, such as CNN-based detectors (Gambäck and Sikdar, 2017) and domain adversarial training (Markov et al., 2023), are introduced to enhance performance. More recently, detectors leveraging pre-trained models and large language models have gained traction (Cohen et al., 2023; Pan et al., 2023). For instance, Pan et al. introduce program-guided fact-checking, which decomposes complex claims into simpler sub-tasks using reasoning programs generated by large language models. Nevertheless, these methods are either too outdated or lack generalizability for detecting diverse multimodal risks.

For large language models (LLMs), alignment is a hot topic aimed at reducing risky outputs by aligning models with human preferences. Existing alignment techniques primarily follow the reinforcement learning from human feedback (RLHF) paradigm (Ouyang et al., 2022; Yu et al., 2024; Sun et al., 2023; Xu et al., 2023), evolving into variants including group relative policy optimization (Shao et al., 2024), and rule-based reward modeling (Mu et al., 2024), among others. For example, Wu et al. use dense reward signals for fine-grained control. Recently, Lightman et al. propose a process reward model that provides feedback on each step of the model's reasoning process, rather than focusing solely on the final result.

However, these methods primarily focus on alignment with real-world content rather than addressing risks in AIGC. Furthermore, these methods require meticulous parameter tuning and necessitate further research to develop fast and adaptive approaches for quickly responding to adversarial risks in real-world scenarios.

3 Risk Detection Problem Formulation

Detecting risky content in AI-generated e-commerce images involves predicting a probability $\hat{y} \in [0, 1]$ for a given image $\boldsymbol{I}$, representing the likelihood that the image contains risky content.

However, purely predicting the label may overfit to specific patterns and lack interpretability, which undermines generalization and contravenes the principles of developing robust detectors. Therefore, we formalize the task as interpretable risk detection, where models generate textual explanations $\boldsymbol{S} = (w_0, w_1, ..., w_T)$ that explicitly identify and describe harmful content while aligning with expert annotations. Here, T denotes the number of decoding steps. The goal is to learn a function $f_\theta : \boldsymbol{I} \rightarrow \{\hat{y}, \boldsymbol{S}\}$ that jointly optimizes classification accuracy (*e.g.,* F1, AUC-ROC) and explanation verifiability, ensuring that predictions are grounded in causal expert rationales rather than spurious correlations.

4 Benchmark Description

This section outlines the construction of the dataset, detailing the workflow for collecting AIGC images, the expert annotation procedure, and the construction of the benchmarks.

Online AIGC Workflow Our dataset comprises images sourced from a real-world e-commerce application. The risky images primarily originate from our advertising creative platforms powered by AI-driven generative tools, such as text-to-image, image-to-image, doodle-style art, virtual model synthesis, and personalized portrait generation. These tools enable creative and cost-effective advertising. The image generation workflow is depicted in Figure 2. As illustrated, product images produced by the AIGC platform's cutout tool, paired with descriptive prompts, are processed by a Flux-based model (Labs, 2024). This model dynamically selects LoRA fine-tuning parameters (Hu et al., 2021), such as visual model, background, and style, to align with the input prompt. This process may inadvertently generate risky content, including hallucinations (*e.g.,* unrealistic product attributes) or toxic information. Additionally, malicious actors could exploit advanced AI techniques to embed inconspicuous text or illegal content, evading standard OCR systems and enabling deceptive material to proliferate undetected. Further technical details of the image generation pipeline are provided in Appendix B.

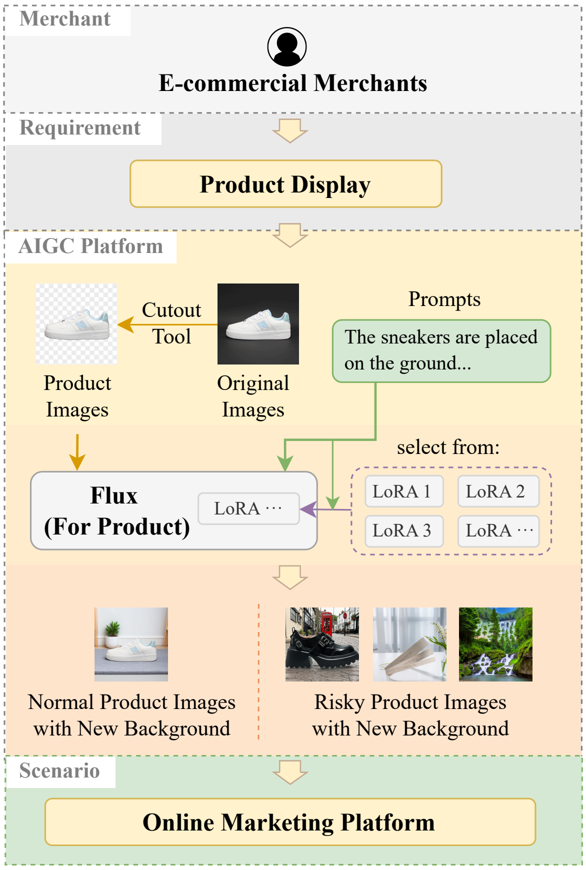

Figure 2: The workflow for generating AI-produced advertisement images. The process begins with the extraction of product images from the original inputs provided by merchants. These extracted images are then combined with descriptive prompts, and subsequently fed into a Flux-based image generation model. Throughout this process, the platform dynamically integrates pre-tuned LoRA modules to tailor features according to user preferences.

Data Collection Procedure Our dataset comprises a subset of samples collected from the e-commerce platform between January 1 and December 31, 2024. During this period, the platforms generated a significantly large volume of images. To filter out normal images, we employ a multi-stage process combining user complaints, manual inspection, and model-based checks. Specifically, all user-submitted complaints flagged with risky tags are subjected to expert annotation procedures. For the large unlabeled dataset, we first recruit professional annotators to manually label the images. Subsequently, we train a ViT-S/16 model (Dosovitskiy et al., 2021), a small version of the Vision Transformer, using professionally annotated data. The trained model is employed to filter out images identified as deemed certainly non-risky, ensuring the resulting dataset contains only high-quality samples with potential risks relevant to real-world e-commerce platforms. The final dataset comprises 253,420 samples, including 43,885 risky and 209,535 normal instances.

Expert Annotation To enhance the generalization and usability of the dataset, we include detailed annotations for each sample. Specifically, we recruit three domain experts to label the data following a standardized procedure.

- **Risk Classification**: Experts are required to categorize the images into four distinct classes based on their risk type: *Abnormal Body* refers to the unrealistic human features (*e.g.,* "a man with three arms"). *Violating Physical Laws* involves images that defy the laws of physics (*e.g.,* "a smartphone floating in mid-air"). *Misleading or Illogical Context* describes images where the background is inconsistent with the main subject (*e.g.,* "a giant toothbrush in a forest landscape"). *Harmful or Problematic Message* includes images with hidden illegal message in the background (*e.g.,* "make fake certificate") in Figure 1.

- **Content Annotation**: Experts then describe whether the image contains risky content. Both risky and normal samples are annotated strictly in a certain format to ensure clarity and accuracy. For *Abnormal Body*, normal images are labeled as "Characters do not have any abnormal features, such as missing bodies, flying heads, twisting limbs, etc.", and risky images are described using a **"Abnormal part + Identification"** format (*e.g.,* "The woman's left hand is deformed and the right hand is missing, and there is an abnormal structure in her body"). For *Misleading or Illogical Context*, normal images are annotated as "The product has no reasonableness issues, the product size is reasonable, and the background is coordinated (not floating in the water or standing on the table, etc.).", and risky images are described using a **"Observation + Assessment"** format (*e.g.,* "Shoes appear on the ground, obviously too large"). More specific annotation rules are detailed in Appendix C.

- **Peer Review**: To ensure label accuracy throughout the annotation process, annotators perform a peer review of each other's annotations and resolve disagreements through majority voting. Corrections are made as needed to adhere to established guidelines. This step is crucial to maintain consistency and reliability in the annotations, ensuring that the dataset is robust and usable for various applications.

Category	Total	Risky	Normal	Ratio
Abnormal Body	76,800	12,768	64,032	≈1:5
Violating Physical Laws	90,880	15,154	75,726	≈1:5
Misleading or Illogical Context	65,280	10,847	54,433	≈1:5
Harmful or Problematic Message	20,460	5,116	15,344	≈1:3

Table 2: The statistic of the dataset.

AIGUARD Benchmark After annotation, the benchmark dataset comprises a total of 253,420 samples. The distribution across categories is as follows: *Abnormal Body* (76,800 samples), *Violating Physical Laws* (90,880 samples), *Misleading or Illogical Context* (65,280 samples), and *Harmful or Problematic Message* (20,460 samples). To balance the dataset, we remove many normal samples, resulting in a risky-to-normal ratio close to 1:5. The statistic of the dataset is shown in Table 2.

5 Lightweight Detection Method

To balance efficiency and effectiveness, we develop a lightweight detection model based on the BLIP framework (Li et al., 2022), which unifies image-text contrastive learning (ITC), image-text matching (ITM), and language modeling (LM) to achieve strong performance across multimodal tasks. Specifically, we distill expert-annotated risk information into soft prompts using image soft prompt matching and language modeling tasks. During inference, we rely solely on image soft prompt matching to reduce detection time. The framework overview is illustrated in Figure 3.

Cross-Attentive Image Soft Prompt Matching To address the absence of text during inference, the module employs learnable soft prompts to encode risk information into general dense vector representation. Specifically, the input image $\boldsymbol{I}$ is first encoded into a feature sequence $\boldsymbol{H}_{\texttt{IMG}} \in \mathbb{R}^{N\times d}$, where N is the number of image patches and d is the embedding dimension. Simultaneously, the soft prompts are represented as $\boldsymbol{H}_{\texttt{S}} \in \mathbb{R}^{L\times d}$, with L denoting the prompt length. These features are concatenated with a `[CLS]` token $\boldsymbol{h}_{\texttt{CLS}}$ and fed into n transformer encoder layers $\texttt{Encoder}_n$, initialized from BLIP's cross-encoder. Then, the final output $\boldsymbol{H}_n$ is calculated as

$$\boldsymbol{H}_n = \texttt{Encoder}_n([\boldsymbol{h}_{\texttt{CLS}}, \boldsymbol{H}_{\texttt{IMG}}, \boldsymbol{H}_{\texttt{S}}]), \qquad (1)$$

where $[\cdot,\cdot,\cdot]$ is the concentrate operation. Finally, the last layer output of the summarized token $\boldsymbol{h}^n_{\texttt{CLS}}$

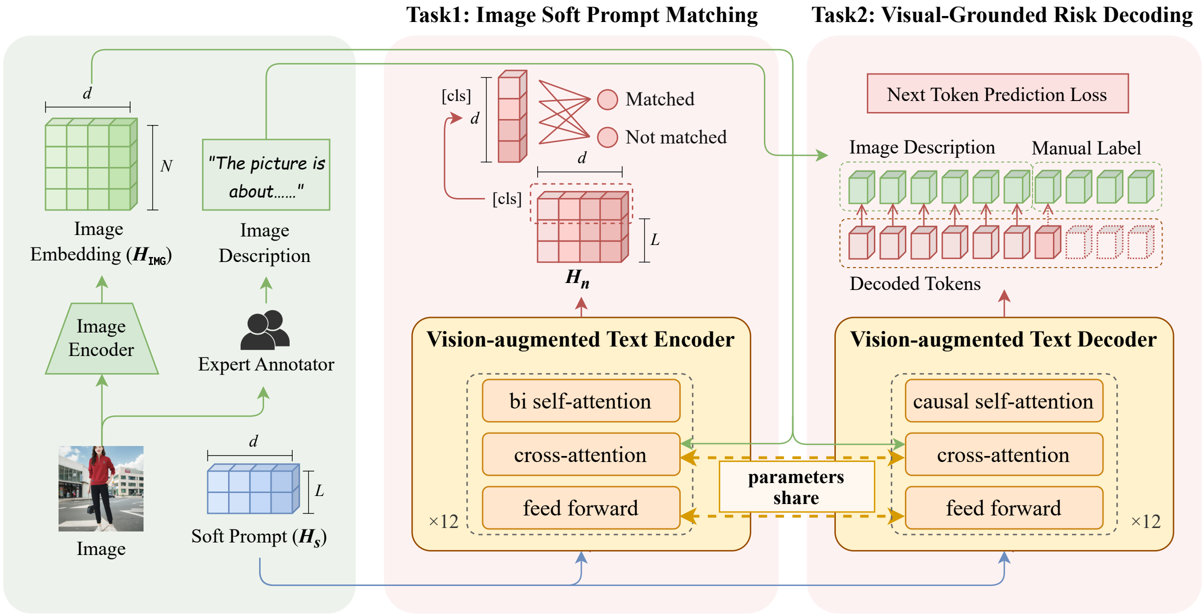

Figure 3: An overview of the lightweight detection method. It contains two components: (1) The image-grounded text encoder leverages cross-attention and soft prompts to identify AIGC risk, guided by an image-text matching objective. (2) The image-grounded text decoder utilizes causal self-attention and is refined through a language modeling objective, embedding semantic information into the soft prompts.

is passed through a MLP to yield the binary classification probabilities as

$$\hat{y} = \mathtt{Sigmoid}\left(\mathtt{MLP}(\boldsymbol{h}^n_{\mathtt{CLS}})\right), \tag{2}$$

where Sigmoid transforms the output of the MLP into a probability score between 0 and 1.

This module is trained using an image soft prompt matching objective, denoted as $\mathcal{L}_{\text{VTM}}$, to determine whether an image contains risk content. The objective function is defined as

$$\mathcal{L}_{\text{VTM}} = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}), \tag{3}$$

where $y = 1$ indicates that the image contains risk content, and $y = 0$ is not.

Visual-Grounded Risk Decoding This module employs a shared architecture that combines causal self-attention, optimized via a language modeling objective. This design enables the integration of semantic information into the soft prompts, allowing the creation of more expressive representations conditioned on visual semantics.

Specifically, given an input image $\boldsymbol{I}$ and the soft prompts, the model autoregressively predicts the t-th token of the risk description $\boldsymbol{S} = (w_0, w_1, ..., w_T)$ at each decoding step t. This is achieved using an m-layer decoder, initialized from BLIP's pre-trained decoder, as $\mathtt{Decoder}_m(w_t|\boldsymbol{I}, \boldsymbol{H}_{\mathtt{S}}, \boldsymbol{H}_{w_{<t}})$ where $\boldsymbol{H}_{w_{<t}}$ denotes the token embeddings of the first $t-1$ tokens. The soft prompts $\boldsymbol{H}_{\mathtt{S}}$ are jointly optimized with the decoder by minimizing the next-token prediction loss $\mathcal{L}_{\text{RD}}$, which encodes expert annotations into the soft prompts as

$$\mathcal{L}_{\text{RD}} = -\sum_{t=1}^{T}\log\mathtt{Decoder}_m(w_t|\boldsymbol{I}, \boldsymbol{H}_{\mathtt{S}}, \boldsymbol{H}_{w_{<t}}). \tag{4}$$

This process ensures that the soft prompts adapt to visual semantics while aligning with annotated risk descriptions.

The final training loss $\mathcal{L}$ is formulated by combining two components:

$$\mathcal{L} = \mathcal{L}_{\text{VTM}} + \lambda\mathcal{L}_{\text{RD}}, \tag{5}$$

where λ is a hyper-parameter that balances the influence of $\mathcal{L}_{\text{RD}}$.

6 Experiments

This section examines intuitive risk detection methods on AIGUARD and compares their effectiveness against our proposed lightweight approach. The experimental results provide valuable insights and suggest promising directions for refining risk detection models in future work.

Model	Params	Abnormal Body			Violating Physical Laws			Misleading or Illogical Context			Harmful or Problematic Message			Overall		
		R	P	F1	R	P	F1	R	P	F1	R	P	F1	$\overline{\text{R}}$	$\overline{\text{P}}$	$\overline{\text{F1}}$
ResNet-50	235M	48.48	52.80	50.54	75.59	32.08	45.04	40.47	79.01	53.52	2.78	52.57	5.28	50.36	41.45	45.47
ViT-B/16	86M	77.21	76.17	76.68	84.29	80.55	82.38	**83.68**	77.20	80.31	42.78	53.97	47.73	77.47	75.87	76.66
BLIP-LM	224M	73.00	20.38	31.86	84.40	87.09	85.72	7.26	21.10	10.81	53.15	50.62	51.85	58.60	37.25	45.54
BLIP-ITM	447M	79.09	81.50	80.28	83.49	89.10	86.20	77.55	82.53	79.96	**95.19**	25.69	40.46	82.24	63.80	71.86
Qwen2-VL-7B	7.6B	10.54	63.77	18.09	24.90	68.22	36.49	7.64	37.85	12.72	16.11	24.79	19.53	15.53	51.16	23.83
Qwen2-VL-7B (sft)	7.6B	65.66	56.93	60.98	80.28	86.57	83.30	71.42	79.60	75.29	71.48	**86.35**	**78.21**	72.88	74.96	73.90
LLaVa-1.6-7B	7.6B	5.27	44.29	9.42	6.36	**100**	11.96	6.70	33.81	11.18	0.37	50.00	0.73	5.43	51.01	9.82
GLM-4V-9B	9B	3.99	**90.91**	7.64	5.92	78.61	11.01	40.66	30.16	34.63	0.56	17.65	1.09	13.32	35.45	19.36
GPT-4o	-	46.78	35.86	40.60	49.28	37.44	42.55	82.76	36.01	50.18	13.08	27.23	17.67	52.61	36.08	42.80
Ours	500M	**87.74**	84.92	**86.31**	**84.47**	90.27	**87.27**	80.75	**86.64**	**83.59**	67.41	35.24	46.28	**82.40**	**76.06**	**79.10**

Table 3: Performance comparison of different methods on AIGUARD. "Params" denotes the number of parameters in the model. "Overall" is calculated from the entire dataset. The best results are shown in **bold**.

6.1 Experimental Setting

Baseline Models To comprehensively evaluate our proposed method, we compare it against five baseline approaches from distinct categories: **(1) ResNet-50** (He et al., 2016): A foundational convolutional neural network pre-trained on ImageNet (Deng et al., 2009). We adapt this model for risk detection via full fine-tuning. **(2) ViT-B/16** (Dosovitskiy et al., 2021): A base version of the transformer-based vision model using 16×16 patches, pre-trained on ImageNet. We adapt this model for risk detection via full fine-tuning. **(3) BLIP-LM** (Li et al., 2022): The decoder part of BLIP is designed to generate descriptive text from visual data and identify the risk from the descriptive text information. **(4) BLIP-ITM** (Li et al., 2022): We utilize image-text matching part of BLIP for risk classification by simply setting the query text as "The image does not contain any risk information" and fine-tune the model with the benchmark dataset. **(5) Qwen2-VL-7B** (Bai et al., 2023): A state-of-the-art large multimodal model with 7.6 billion parameters. We evaluate its performance before and after LoRA (Hu et al., 2021) fine-tuning. **(6) LLaVa-1.6-7B** (Liu et al., 2024a): A state-of-the-art large multimodal model with 7.6 billion parameters, which showcases remarkable zero-shot capabilities in Chinese. **(7) GLM-4V-9B** (GLM et al., 2024): A state-of-the-art large multimodal model with 9 billion parameters, which achieves impressive performance in both Chinese and English tasks and can effectively utilize tools to complete complex tasks. **(8) GPT-4o** (Achiam et al., 2023): A state-of-the-art multimodal language model developed by OpenAI[3]. It is capable of processing and generating high - quality text based on the input prompts, and has been widely used in various natural language processing tasks. In this paper, it serves as one of the baselines for performance comparison.

Implementation Details We employ the ViT-B/16 model as our baseline, initializing it with Googles official checkpoint (Dosovitskiy et al., 2021). For BLIP-based caption generation and image-text retrieval, we utilize COCO-fine-tuned checkpoints provided by the BLIP authors (Li et al., 2022). As for the trainable soft prompts, they comprise 25 embeddings, which are initialized by averaging the predefined negative label token embeddings. Consistent with the BLIP model configuration, the Transformer architecture comprises both encoder and decoder layers, each with a layer size of 12, and the hyper-parameter λ is set to 1. Prior to encoding, all input images are resized to 384×384 resolution. The training and test datasets are split in a 9:1 ratio. For the four risk categories in AIGUARD, baseline models (except Qwen2-VL-7B) are trained for 25 epochs, while Qwen2-VL-7B is trained for one epoch. The training process uses an initial learning rate of 1×10^{-5} and a weight decay of 0.05. For all experiments, we report precision (P), recall (R), and F1-score (F1) as performance metrics.

6.2 Performance Comparison

Table 3 presents the recall, precision, and F1 scores of the evaluated baseline models on our AIGUARD benchmark. All experiments are repeated four times to ensure reliability, with results averaged across runs to reflect consistent performance metrics. From the table, we have followed observations: **(1) Our experiments reveal significant room for improvement, underscoring the need for further research.** Current baseline methods, including MLLMs (*e.g.,* Qwen2-VL-7B), demonstrate limited effectiveness on our dataset. For instance, Qwen2-VL-7B after supervised fine-tuning (sft) only achieves a recall of 65.66% and a precision of 56.93% on the abnormal body detection task. **(2) Our lightweight framework es-**

[3] https://openai.com/

tablishes state-of-the-art results. Our method outperforms Qwen2-VL-7B by 9.68% in recall and 1.10% in precision on the overall dataset. It also reduces GPU memory consumption by 4.45× during training and increases inference QPS by 37.8×, as shown in Table 6. **(3) MLLMs exhibit limited proficiency in image risk detection.** We evaluate MLLMs (e.g., Qwen2-VL-7B, LLaVa-1.6-7B, GLM-4V-9B) on our AIGUARD dataset. Despite their extensive world knowledge, the models' performance remain suboptimal (*e.g.,* recall of 80.28% in *Violating Physical Laws* category for Qwen2-VL). Fine-tuning improves results slightly, but our approach still outperforms it across most risk types. This suggests MLLMs currently lack specialized knowledge for AIGC content analysis. **(4) General MLLMs hold promise for future risk detection.** The fine-tuned Qwen2-VL-7B model achieves strong performance in the *Harmful or Problematic Message* category, demonstrating MLLMs' potential for complex harmful message detection tasks (*e.g.,* detecting hidden text). As a comparison, the best-performing method ViT-B/16 (Dosovitskiy et al., 2021) among the remaining baselines, including our method, achieves only an F1 score of 51.85%. This highlights the difficulties these methods encounter in performing the task, attributable to their deficiencies in context comprehension and world knowledge.

6.3 Analysis Experiments

Benefit of Textual Description We evaluate the impact of text descriptions by comparing model performance with/without text inputs under identical settings, where text is used only during training (not inference). As shown in Table 4, **integrating descriptive text description allows the model to improve F1 scores on specific tasks while balancing precision and recall.** For image-intensive tasks (*e.g., Violating Physical Laws* detection), the improvement is particularly obvious, with precision improving by 2.74% and recall improving by 0.33%. For the *Harmful or Problematic Message* detection task, which requires the model to identify and detect risky hidden text accurately, the incorporation of descriptive image labels is counterproductive and adversely affects the model's training, leading to lower identification precision.

Category	Text-Description	Recall	Precision	F1-score
Abnormal	✗	87.66	84.54	86.07
Body	✓	**87.74**	**84.92**	**86.31**
Violating Physical	✗	84.14	87.53	85.80
Laws	✓	**84.47**	**90.27**	**87.27**
Misleading or	✗	**81.60**	85.81	**83.66**
Illogical Context	✓	80.75	**86.64**	83.59
Harmful or	✗	67.04	**43.77**	**52.96**
Problematic Message	✓	**67.41**	35.24	46.28

Table 4: Comparison of model performance influenced by text description labels generated by large multimodal models. The symbol "✗" indicates the absence of a text-description, while "✓" indicates its presence. The best results are shown in **bold**.

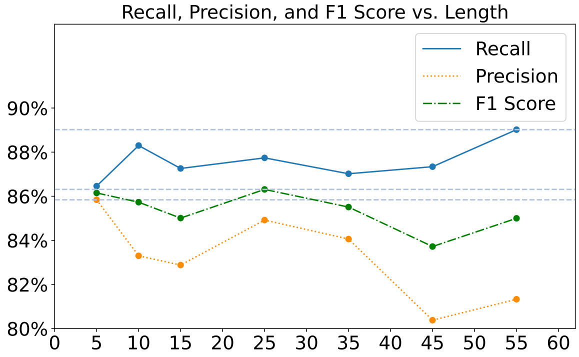

Figure 4: Performance comparison of different prompt lengths on the *abnormal body* detection task. The horizontal axis represents the length of soft prompts, and the vertical axis represents the corresponding recall, precision, and F1-score.

Influence of Prompt Length We compare the model performance of the learnable soft prompts under different length settings. As shown in Figure 4, **our approach achieves the highest F1 score when the prompt length is set to 25.** When the prompt length is 5, we observe a suboptimal F1 score, but recall and precision remain well balanced. As the prompt length moves away from 25, both recall and precision decline, which confirms that our experimental setup is sound.

Further ablation study and analysis experiments are presented in Appendix D.

7 Discussion

Based on the dataset and experimental results, this section highlights key challenges and emerging research opportunities for advancing AIGC safety. **(1) Advanced Risk Detection**: Though lightweight detection methods show promise, significant improvements are still needed. Existing approaches struggle particularly with *Harmful or Problematic Message* risks, where threats are well-hidden and demand more generalizable solutions. Furthermore, as detection methods improve, malicious users may adapt their tactics to hide illegal content, highlighting the need for robust and adaptable detection strategies. **(2) MLLMs and AIGC Risks**: Our experiments reveal that current MLLMs often fail to detect subtle risks in AIGC due to their training on standard real-world data. This underscores the need to expose MLLMs to AIGC outputs, particularly adversarial or risky content, to enhance their understanding. With the increasing prevalence of AIGC, it is crucial to introduce AIGC content to MLLMs and develop specialized algorithms (*e.g.,* contrastive learning, adversarial training, pre-training), to improve the models' ability to recognize and process AIGC content effectively. **(3) AIGC Alignment Datasets**: Current alignment efforts mainly focus on real-world images. As these data sources become limited, using composite data generated by models can enhance model understanding and improve generalization from weak to strong. Future work could expand the dataset to AIGC data, which can be generated at a low cost and with controlled parameters. Our dataset can also serve as a foundation for building alignment datasets to improve models' comprehension of e-commerce AIGC content. **(4) Safer AIGC Generation**: This work provides a real-world scenario for studying the safety of AIGC generation. As the source of this problem, we also highlight the need to develop safer and controllable generation methods, which can address risks in AIGC content at root.

8 Conclusion

This work introduces AIGUARD, the first benchmark designed to detect AIGC risks within e-commerce contexts. The benchmark comprises 253,420 image-text pairs, each annotated with corresponding risk information. Then, we propose an effective and lightweight detection method that distills risk annotations into learnable soft prompts via image-text matching and next-token prediction tasks. Experimental results demonstrate the superior performance of our approach and provide insights into future directions for developing robust detection methods in real-world systems.

9 Limitation

This study faces two primary limitations. First, the dataset has inconsistent annotation standards. Specifically, while risky data are labeled in detail, normal data are labeled uniformly across different categories. This inconsistency requires further exploration of dataset caption methods to fully utilize the dataset. Second, the risk content in the system is dynamic. Therefore, models over-fitted on this dataset may not perform well in real-world systems, where risk patterns change rapidly. As a result, this work can only provide guidance for method development and encourages the development of general and powerful models that can generalize across diverse and different risk types.

10 Ethical Considerations

Privacy While constructing AIGUARD from content generated by the AIGC platform, we observe that the content is influenced by the pre-training data of the underlying model and may include elements that resemble human features. We affirm that our dataset does not include any personal information, ensuring that it can be safely released and utilized.

Legitimacy Certain images in AIGUARD contain content associated with illegal black and gray market transactions. We wish to clarify that our intention is not to promote illegal transactions. Instead, their focus is on analyzing the detection effectiveness of *harmful or problematic message* risk. All the risky images we collect have been prohibited on e-commerce platforms.

Acknowledgment

We express our sincere gratitude for the financial support provided by the CCF-ALIMAMA TECH Kangaroo Fund (NO. CCF-ALIMAMA OF 2024009), the National Natural Science Foundation of China (No. U23A20305, NO. 62302345), and the Natural Science Foundation of Wuhan (NO. 2024050702030136), Innovation Scientists and Technicians Troop Construction Projects of Henan Province, China (No. 254000510007), National Key Research and Development Program of China (No. 2022YFB3102900), the Xiaomi Young Scholar Program.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575.

Xiang Chen, Chenxi Wang, Ningyu Zhang, Yida Xue, YUE SHEN, GU Jinjie, Huajun Chen, et al. 2024. Unified hallucination detection for multimodal large language models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Jack Clark and Ray Perrault. 2024. Ai index report. https://hai.stanford.edu/research/ai-index-report.

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: harnessing logical rules for explainable hate speech detection. *arXiv preprint arXiv:2307.12935*.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12621–12640.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024b. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403. Springer.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. 2024c. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. 2024. Rule based rewards for language model safety. *arXiv preprint arXiv:2411.01111*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744*.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305.

Jakub Podolak, Szymon Łukasik, Paweł Balawender, Jan Ossowski, Jan Piotrowski, Katarzyna Bakowicz, and Piotr Sankowski. 2024. Llm generated responses to mitigate the impact of hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15860–15876.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. 2023. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3403–3417.

Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and tracking political abuse in social media. In *Proceedings of the International AAAI Conference on Web and social media*, volume 5, pages 297–304.

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.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Xinyue Shen, Yixin Wu, Yiting Qu, Michael Backes, Savvas Zannettou, and Yang Zhang. 2025. Hatebench: Benchmarking hate speech detectors on llm-generated content and hate campaigns. *arXiv preprint arXiv:2501.16750*.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. im sorry to hear that: Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Min Tang, Lixin Zou, Zhe Jin, ShuJie Cui, Shiuan Ni Liang, and Weiqing Wang. 2025. Chifraud: A long-term web text dataset for chinese fraud detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5962–5974.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023a. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Xinpeng Wang, Xiaoyuan Yi, Han Jiang, Shanlin Zhou, Zhihua Wei, and Xing Xie. 2023b. Tovilag: Your visual-language generative model is also an evildoer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3508–3533.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

Zonghao Ying, Aishan Liu, Siyuan Liang, Lei Huang, Jinyang Guo, Wenbo Zhou, Xianglong Liu, and Dacheng Tao. 2024. Safebench: A safety evaluation framework for multimodal large language models. *arXiv preprint arXiv:2410.18927*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2023. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36:44502–44523.

A AIGUARD License

The AIGUARD dataset is available for free download at https://huggingface.co/datasets/yinyueguilai/AIGUARD_dataset and can be used for non-commercial purposes under a custom license, CC BY-NC 4.01. In addition to the existing tasks in the dataset directory, users are permitted to define their own tasks under this license.

B Details of Image Generation in the Dataset

In this section, we will give a detailed description of image generation in the online AIGC platform.

The dataset proposed in this work consists of images from two sources: product advertisement images and images that may contain illegal information. For images from the first source, the sample generation process is illustrated in Figure 2, which demonstrates the process by which an e-commerce merchant can obtain AI-generated advertisement images for product display. The process begins with the merchant providing requirements for product display, then the requirements are addressed by the AIGC platform. The workflow involves the following steps:

- **Original Images**: The original images of the products to be displayed. The merchants need to start by taking product photos as original images and uploading them to the platform.

- **Cutout Tool**: A cutout tool developed by the AIGC platform, which is used to isolate the product from the original images, creating product images. The merchant can obtain product images with the assistance of the tool and seamlessly input them into the model.

- **Prompts**: Textual prompts created to guide the image generation process, *e.g.,* "The sneakers are placed on the ground...". The merchants control the background generation by inputting these descriptive prompts into the model.

- **Flux Model**: The product images and the prompts are fed into the Flux-based model (Labs, 2024), which is designed for product image generation.

- **LoRA Selection**: The AIGC platform selects appropriate LoRA (Low-Rank Adaptation) (Hu et al., 2021) modules to fine-tune the Flux-based model based on the given prompts.

- **Output**: The Flux-based model may generate two types of images: (1) *Normal Product Images with New Background*: These are standard images of the product with a new background. (2) *Risky Product Images with New Background*: These are abnormal images that may not meet the desired quality or could be inappropriate.

During this process, different types of risky content may be generated. In this work, we divide the risk information from this source into the following three categories based on their specific forms: *Abnormal Body*, *Violating Physical Laws*, and *Misleading or Illogical Context*. The specific connotations of them are elaborated in Section 4.

As for the images that may contain illegal information, we primarily focus on the issue of embedding illicit text into images using AI technology in this work, for they are relatively easy to generate but difficult to detect. For instance, they can be generated by some open-source text-to-image web applications for image generation developed based on Stable Diffusion (Rombach et al., 2022) along with ControlNet (Zhang et al., 2023) plugin. These applications can accept a text prompt and a control signal image for ControlNet, allowing it to generate an image that highlights the white areas in the control signal while adhering to the prompt in the background. Following this manner, malicious merchants can convert an illegal text to a black-and-white binary image, whose white areas are the text patterns, and then feed this image along with a prompt (describing the background of the image) into the platform to generate an image that subtly incorporates the illegal text. We describe this category of risks as "*Harmful or Problematic Message*".

Figure 5 illustrates the four categories of risky images mentioned in this section.

C Additional Annotation Rules

For *Harmful or Problematic Message*, all images contain hidden text information that is difficult to detect. The images are described using a **"Whether it is a violation + Hidden text"** format, *e.g.,* "There is violation information in the hidden text ’pinhole camera’ in the figure", or "There is no violation information in the hidden text ’fashion shoes’ in the figure".

For *Violating Physical Laws*, we primarily focus on whether the images display any phenomena that defy **the law of gravity**. In particular,

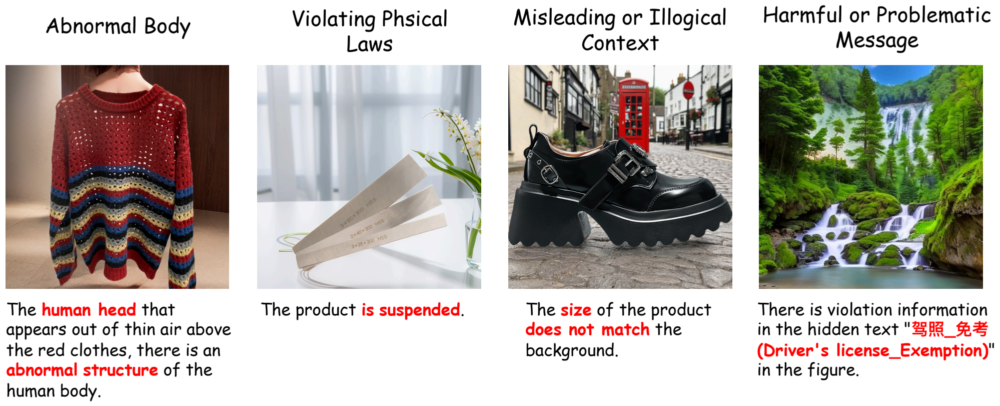
Abnormal Body
Violating Phsical Laws
Misleading or Illogical Context
Harmful or Problematic Message
The human head that appears out of thin air above the red clothes, there is an abnormal structure of the human body.
The product is suspended.
The size of the product does not match the background.
There is violation information in the hidden text "驾照_免考 (Driver's license_Exemption)" in the figure.

Figure 5: Our dataset covers four categories of AI-generated risk images: *Abnormal Body*, *Violating Physical Laws*, *Misleading or Illogical Context*, and *Harmful or Problematic Message*. This figure provides specific examples for each category, including images and manually annotated labels.

Category	Soft Prompts	LM	Recall	Precision	F1
Abnormal Body			79.09	81.50	80.28
	✓		**88.46**	82.20	85.22
	✓	✓	87.74	**84.92**	**86.31**
Violating Physical Laws			83.49	89.10	86.20
	✓		84.40	89.07	86.68
	✓	✓	**84.47**	**90.27**	**87.27**
Misleading or Illogical Context			77.55	82.53	79.96
	✓		**82.36**	84.59	83.46
	✓	✓	80.75	**86.64**	**83.59**
Harmful or Problematic Message			**95.15**	25.69	40.46
	✓		65.93	32.28	43.34
	✓	✓	67.41	**35.24**	**46.28**

Table 5: Performance comparison of our detection method under different component configurations."✓" represents the corresponding components are equipped. The best results are shown in **bold**.

we describe the images using overall statements. Normal images are labeled as "The product is not suspended.", and risky images are labeled as "The product is suspended.".

D More Analysis Experiments

Contribution of Each Component We evaluate the contributions of each component to the model's overall performance. The results, presented in Table 5, confirm that both **the soft prompts and language model components positively impact precision and F1 score**.

Computational Resource Consumption Table 6 shows the computation resource consumption of our method compared to Qwen2-VL-7B. Under the same experimental settings, our method reduces GPU memory usage by **4.45**× during training and **16**× during inference, decreases training time by **4.89**×, and increases the QPS of inference by **37.8**×.

Resource Consumption	**Training**		**Inference**	
	GPU Memory(GB)↓	Time(min)↓	GPU Memory(GB)↓	QPS↑
Qwen2-VL(sft)	60.1	72.8	17.6	1.5
ours	**13.5**	**14.9**	**1.1**	**56.7**

Table 6: Comparison of resource consumption between our method and fine-tuned Qwen2-VL-7B during training and inference under consistent settings on a single NVIDIA H20 GPU. Training involves 10,000 samples with a batch size of 4 over one epoch, while inference uses a batch size of 16. "↓" means lower values are better, and "↑" means the opposite. The best results are shown in **bold**.