中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-21

论文引用格式:Yang Jinxu, Wang Zhiyuan, Zhao Pengcheng, Chen Yanxiang. Semantic alignment dataset and benchmark for AI-generated image detection[J/OL]. Journal of Image and Graphics, XXXX:1-21. DOI: 10.11834/jig.260297.(杨锦旭, 王志远, 赵鹏铖, 陈雁翔. 面向人工智能生成图像检测的语义对齐数据集与基准研究[J/OL]. 中国图象图形学报,XXXX:1-21. DOI: 10.11834/jig.260297.)[DOI:10.11834/jig.260297]

面向人工智能生成图像检测的语义对齐数据集与基准研究

杨锦旭[1],王志远[2],赵鹏铖[3],陈雁翔[2*]

1. 合肥工业大学卓越工程师学院,合肥　230009; 2. 合肥工业大学计算机与信息学院,合肥　230601; 3. 南京审计大学计算机学院,南京　211815

摘　要:**目的**　当前人工智能生成图像技术在推动高质量视觉内容创作的同时,也带来了信息安全挑战。在图像伪造检测任务中,现有训练与评测数据中的真实图像与伪造图像存在细粒度语义内容差异,使语义内容与真伪标签之间容易形成虚假关联,从而形成语义偏见——这是一种捷径学习现象,即检测器借助真伪图像间易于区分的语义内容差异进行判别,而非真正学习生成过程遗留的伪造痕迹;这意味着检测器在现有基准上借助语义差异取得的高性能,并不能代表其在语义对齐条件下的真实泛化能力。针对以上问题,提出了一种基于语义对齐约束的图像伪造检测数据集与检测基准。**方法**　为解决上述语义偏见问题,本文以实例级语义对齐为约束构建AIGI-Align数据集:将每张真实图像作为语义参照,生成与之在细粒度语义维度上对齐的伪造图像,从数据层面弱化语义偏见的形成条件。具体而言,从ImageNet中筛选真实图像,利用多模态大模型进行语义类别验证,确保真实图像内容与类别标签一致;在此基础上,引入视觉语言模型为真实图像生成细粒度结构化图像描述,并利用这些描述驱动12种生成模型生成与真实图像语义高度对齐的伪造图像。**结果**　实验表明,现有检测器可能利用真伪图像之间的语义差异进行捷径判别,而AIGI-Align能够提高真实图像与伪造图像之间的语义对齐程度,缓解语义偏见对检测器判别过程的干扰,从而提升其跨生成模型泛化能力。在3个公开基准和AIGI-Align基准上对8种代表性检测器进行了泛化性评估,结果显示:在传统训练设置下,部分检测器在语义对齐测试场景中的性能明显下降,说明其原有性能可能部分依赖语义捷径而非对伪造痕迹的鲁棒识别;在引入语义对齐训练数据后,多数检测器在AIGI-Align和公开基准上的跨生成模型检测性能均得到提升。训练数据控制消融进一步表明,实例级语义对齐约束对AIGI检测任务具有实际意义。**结论**　本文构建了一个面向AIGI检测任务的大规模实例级语义对齐数据集AIGI-Align。在训练方面,语义对齐数据能够引导模型关注底层伪造痕迹,降低真伪判别中对语义内容差异的依赖,从而提升对未见生成模型的泛化能力;在评测方面,AIGI-Align测试集提供了控制语义变量的评估场景,能够在削弱语义捷径的条件下评估模型的真实泛化性能,有助于更准确地反映检测器的跨生成模型泛化水平。论文相关数据集地址:https://huggingface.co/datasets/likeYousif617/AIGI-Align。

关键词:人工智能生成图像;图像伪造检测;语义偏见;语义对齐;数据集;基准

Semantic alignment dataset and benchmark for AI-generated image detection

Yang Jinxu[1], Wang Zhiyuan[2], Zhao Pengcheng[3], Chen Yanxiang[2*]

1. Graduate College for Engineers, Hefei University of Technology, Hefei 230009, China; 2. School of Computer Science and Information

收稿日期:2026-05-26;**修回日期**:2026-08-14

* **通信作者**:陈雁翔 chenyx@hfut.edu.cn

基金项目:国家自然科学基金项目(61972127)

Supported by:National Natural Science Foundation of China(61972127)

Engineering, Hefei University of Technology, Hefei 230601, China; 3. School of Computer Science, Nanjing Audit University, Nanjing 211815, China

Abstract: Objective Recent advances in artificial intelligence generated image (AIGI) synthesis have improved the efficiency and quality of visual content creation, but they have also introduced information security risks. Realistic generated images can be misused for identity forgery, fake news fabrication, and evidence manipulation, making real-fake image discrimination an important task in digital media forensics. In image forgery detection, fine-grained semantic differences often exist between real and generated images in training and evaluation datasets. These differences can create spurious correlations between semantic content and authenticity labels, resulting in semantic bias. Such semantic bias can be regarded as a shortcut learning phenomenon: instead of learning forgery traces left by the generation process, detectors may rely on distinguishable semantic content differences between real and fake images for classification. This means that the high performance achieved by detectors on existing benchmarks with the help of semantic shortcuts may not fully reflect their generalization ability when semantic shortcuts are weakened. To address this problem, this paper proposes AIGI-Align, an instance-level semantic alignment dataset and benchmark for AIGI detection. **Method** To mitigate semantic bias from the data perspective, AIGI-Align is constructed under an instance-level semantic alignment constraint. The core idea is to use each real image as a semantic reference and generate corresponding fake images that are aligned with it in fine-grained semantic dimensions. Real images are selected from ImageNet and filtered according to image resolution to ensure basic visual quality. A multimodal large model is used to verify whether the image content matches the target category, ensuring semantic category correctness and reducing category mismatch. Unlike conventional datasets that rely on coarse category labels such as "a photo of [class]" as generation prompts, this work introduces a vision-language model to generate fine-grained structured descriptions for each real image. These descriptions cover four visual dimensions: subject description, photographic style, image details, and background elements. These image-level descriptions are then used as text prompts to drive generative models, so that the generated images are constrained by the semantic content of the corresponding real images rather than by a coarse class label alone. Based on these prompts, 12 generative models are used to synthesize fake images, covering U-Net-based latent diffusion models, diffusion Transformers, rectified-flow Transformers, cascaded diffusion models, and a text-to-video model used in a single-frame setting. After generation, the fake images are checked by a multimodal large model for semantic category consistency, and samples with obvious visual artifacts are removed through manual inspection. The dataset contains 532 000 high-resolution images, including 70 000 real images and 462 000 generated images. It covers 14 semantic categories and 12 generative models, providing both instance-level semantic alignment training data and a controlled benchmark for evaluating cross-generator generalization. **Result** Experiments show that AIGI-Align improves the degree of semantic alignment between real and fake images, alleviates the interference of semantic bias in the decision process of detectors, and improves their cross-generator generalization ability. Semantic bias analysis further indicates that existing detectors may use semantic differences between real and fake images for shortcut classification. We conduct a systematic evaluation on eight representative AIGI detectors, including CNN-based methods, frequency- or transformation-based methods, and methods based on pretrained visual representations. Evaluation is performed on ForenSynths, UniversalFakeDetect, GenImage, and the proposed AIGI-Align benchmark. Experimental results show that, under the conventional training setting, some detectors exhibit a performance drop in the semantic alignment evaluation scenario. This suggests that their original performance may be partially supported by semantic shortcuts rather than by robust recognition of forgery traces. When real and fake images are more consistent in semantic content, the semantic differences available to detectors are weakened, making the evaluation more focused on whether detectors can capture generation-related traces. After introducing instance-level semantic alignment training data, most detectors achieve improved cross-generator detection performance on both AIGI-Align and public benchmarks. These results indicate that instance-level semantic alignment data can reduce detectors′ dependence on semantic content differences and encourage them to learn more general forgery-related cues. **Conclusion** This paper presents AIGI-Align, a large-scale instance-level semantic alignment dataset for AIGI detection. In terms of training, instance-level semantic alignment data can guide detectors to focus more on generation-related forgery traces and reduce their reliance on semantic content differences during real-fake dis-

crimination, thereby improving their generalization ability to unseen generative models. In terms of evaluation, the AIGI-Align test set provides a controlled scenario in which semantic variables are constrained, making it possible to evaluate the actual generalization performance of detectors when semantic shortcuts are weakened. Our dataset is available athttps://huggingface. co/datasets/likeYousif617/AIGI-Align.

Key words: artificial intelligence generated image; image forgery detection; semantic bias; semantic alignment; dataset; benchmark

论文引用格式：Yang J X, Wang Z Y, Zhao P C and Chen Y X. 2026. Semantic alignment dataset and benchmark for AI-generated image detection. Journal of Image and Graphics, x(x): x-x (杨锦旭, 王志远, 赵鹏铖, 陈雁翔. 2026. 面向人工智能生成图像检测的语义对齐数据集与基准研究. 中国图象图形学报, x(x): x-x) [DOI:10. 11834/jig. 260297]

0 引　言

图像生成技术的进步丰富了数字内容的创作形式，推动了数字艺术和工业设计等相关产业的发展。但是，当前AI(Artificial Intelligence)生成的图像在视觉上具有较强的欺骗性，人眼难以分辨真伪。这类技术若被用于伪造人像、捏造虚假新闻或伪造司法证据，将直接威胁公共信息安全，引发严重的社会信任危机，造成不可估量的经济损失。

为了应对AI生成图像带来的真伪鉴别挑战，研究者们提出了多种图像伪造检测方法。例如，CNN-Spot(Wang 等，2020)通过卷积网络捕捉ProGAN(Karras 等，2017)生成的高频指纹；LaDeDa(Cavia 等，2024)将感受野严格限制在局部图像块以专注鉴别底层噪声；UnivFD(Ojha 等，2023)直接利用预训练视觉大模型CLIP(Radford 等，2021)的丰富特征空间进行真伪分类；FatFormer(Liu 等，2024)通过在Transformer中插入伪造感知适配器来学习通用特征；LTD(Yang 等，2026)利用真实图像与伪造图像在深层网络中的层间特征过渡差异进行伪造鉴别。然而，随着生成模型的持续更新和生成图像质量的不断提升，现有检测器的跨生成模型泛化能力仍然有限(王亚斌等，2025)。

检测器在面对未见生成模型时的泛化能力不足，其中一个可能原因是语义偏见问题。本文讨论的语义偏见，是指检测器在真伪判别过程中受到图像语义内容差异的干扰，将类别、主体姿态、光照条件、背景环境等语义内容误作为区分真实图像与伪造图像的依据，而不是根据生成图像中的伪造痕迹进行判断。语义偏见是捷径学习(shortcut learning)在AIGI检测领域的具体体现，其产生原因在于：随着生成模型不断进步，生成图像中的伪造痕迹逐渐减弱，而真实图像与伪造图像在语义内容上的差异可能相对更加突出，因此更容易被检测器捕捉。在这种情况下，检测器可能利用语义内容与真伪标签之间的虚假相关来区分真实图像和生成图像，从而形成语义偏见。由于这种语义判别依据并不真正对应生成图像中的伪造痕迹，当测试图像来自未见生成模型或者真伪图像在语义内容上保持对齐时，检测器原本依赖的语义差异被削弱，其检测性能可能下降。因此，在真伪图像语义对齐的评估场景下，语义偏见可能会限制检测器的跨生成模型泛化能力。

语义偏见的产生与数据集构建方式密切相关。现有训练与评测数据构建方式主要有以下三类：第一类数据集依赖粗粒度类别级提示词生成伪造图像，例如GenImage(Chen 等，2023)、DiffusionForensics(Wang 等，2023)和AIGCDetect(Zhong 等，2023)使用“a photo of [class]”等类别模板作为生成条件，生成图像只能在粗粒度类别层面与真实图像保持一致；第二类数据集从在线平台收集真实图像和AI生成图像，例如Chameleon(Yan 等，2025)从AI艺术社区和真实图像平台中收集高质量图像；第三类数据集引入大语言模型批量生成多样化文本描述，例如AIGIBench(Li 等，2026)利用大语言模型批量生成面向人物、动物、物体和风景四类内容的多样化句子描述，并基于这些描述生成图像。以上数据集在构建过程中没有以具体真实图像为语义参照，导致真实图像与伪造图像之间存在语义内容差异。这种语义内容差异会使语义因素与真伪标签之间形成虚假关联，从而为检测器学习语义捷径提供条件，使图像真伪判别过程受到语义偏见的影响。

为了缓解上述问题，本文提出实例级语义对齐

约束思想:以每张真实图像作为语义参照,在保证类别一致性的基础上,进一步控制真实图像与对应伪造图像之间的细粒度语义内容差异。具体而言,本文从主体描述、摄影风格、图像细节和背景元素等维度建立真伪图像间的语义对齐关系,从数据层面削弱语义内容与真伪标签之间的虚假关联。基于该思想,本文设计了语义对齐数据构建方法:引入视觉语言模型为每张真实图像生成细粒度图像描述提示词,利用这些图像描述驱动多种生成模型生成语义受约束的伪造图像,利用多模态大模型进行语义类别验证。最终,本文构建了面向AIGI检测任务的实例级语义对齐训练与评估数据集AIGI-Align,为分析和缓解图像伪造检测任务中的语义偏见问题提供数据基础。

本文的主要贡献包括以下三点:1)通过实例级语义对齐约束,构建面向人工智能生成图像检测任务的语义对齐训练与评估数据集AIGI-Align。数据集包含532 000张高分辨率图像,涵盖14个语义类别及12种生成模型。2)在评估阶段,通过AIGI-Align测试集提供控制语义变量的评估场景,在语义对齐条件下评测现有检测器的真实泛化能力,揭示语义偏见对模型判别的影响。3)在训练阶段,引入实例级语义对齐真伪图像数据,验证语义对齐约束是否有助于引导检测器关注伪造痕迹,减少对语义内容差异的依赖,提升模型在语义捷径受限条件下对未见生成模型的泛化能力。

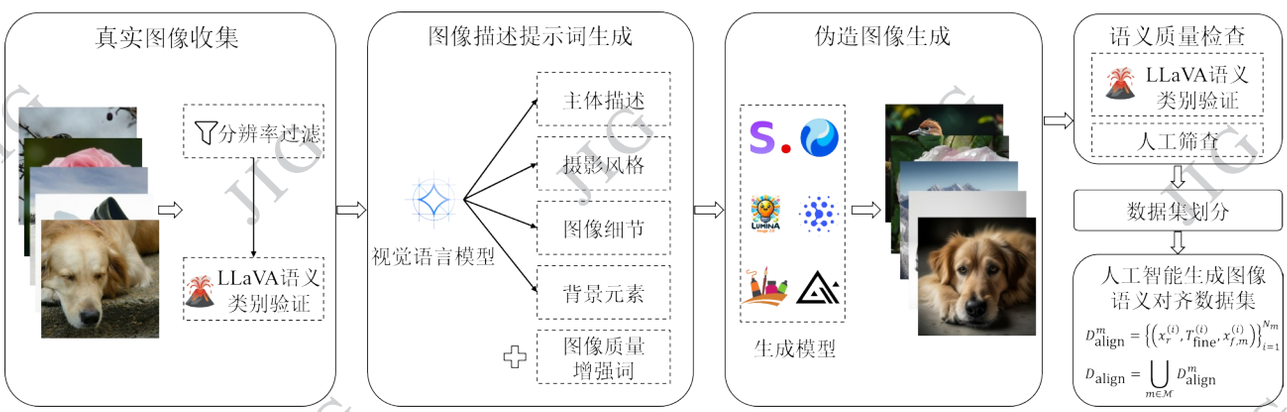

图1 AIGI-Align语义对齐数据集构造流程

Fig. 1 Construction pipeline of the semantic alignment dataset AIGI-Align

1 数据集构建

AIGI-Align数据集的整体构建流程如图1所示,包括真实图像收集、图像描述提示词生成、伪造图像生成、语义质量检查和数据集划分五个步骤。首先,从ImageNet(Deng等,2009)中筛选高分辨率真实图像,并进行语义类别验证;其次,利用视觉语言模型为每张真实图像生成细粒度结构化图像描述;然后,将这些描述作为提示词输入多种生成模型,生成以真实图像为语义参照的伪造图像;最后,通过语义类别检查和人工筛查保证伪造图像与真实图像之间的语义对齐程度和视觉质量,并按照训练、验证和测试的实验设置对数据集进行划分。

1.1 真实图像收集

为了确保数据集的语义多样性并模拟复杂真实场景,本文选用了大规模视觉识别领域的数据集ImageNet作为数据源。本文从ImageNet中筛选出包含14个语义类别的真实图像子集,这些图像包含了生物(如猫、鱼)、物体(如鞋子、瓶子)、景观(如山脉、桥梁),具体类别包括:飞机、猫、花、鱼、汽车、狗、水果、鞋子、鸟、船只、山峰、瓶子、桥梁和昆虫。本文根据图像分辨率对真实图像进行过滤,筛除了宽高低于256 × 256的图像。

为了避免出现类别泄漏和类别不匹配的问题,本文引入多模态大模型LLaVA(Liu等,2023)对真实图像进行严格的语义类别检查,确保图像内容与其类别标签一致,避免一个目标类别中的图像中出现其他类别对象。

1.2 伪造图像构建

1.2.1 基于视觉语言模型的提示词生成

以往图像伪造检测数据集比如GenImage和Dif-

fusionForensics依赖粗粒度的类别提示词模板(例如“a photo of [class]”)生成伪造图像,这种方法会导致生成图像与真实图像在语义内容上不匹配,从而在训练阶段误导检测器关注真伪图像的语义差异,在评估阶段又使这些差异成为检测器判别真伪的语义捷径,难以真实反映跨生成模型的泛化能力。针对这一问题,本文使用视觉语言模型Gemma-3(Gemma Team 等,2025)对每一张经过语义类别验证的真实图像生成细粒度的描述提示词,精确捕捉物体姿态、光照条件等细节,从而提高生成图像与对应真实图像之间的实例级语义对齐程度。

本文设计了结构化的图像描述模板,要求视觉语言模型输出四个维度的结构化信息:1)主体描述,记录核心对象的物理状态和动作。2)摄影风格,包括摄像机视角、镜头景深和光照条件。3)图像细节,描述目标物体的材质纹理或关键局部特征。4)背景元素,描述主体所处的环境和背景信息。

1.2.2 伪造图像生成

本文使用了12种具有代表性的图像生成模型,包括U-Net潜在扩散模型、基于Transformer的扩散或流匹配模型、级联扩散模型和跨模态生成模型。

1)基于U-Net的潜在扩散模型,包括Stable Diffusion v1.4(Rombach 等,2022)、Stable Diffusion v1.5、Stable Diffusion XL(Podell 等,2024)、Kandinsky 3.0(Arkhipkin 等,2023)和Kolors(Kolors Team,2024)。其中,Stable Diffusion v1.4采用CLIP ViT-L/14作为文本编码器,并利用变分自编码器VAE(variational autoencoder)将图像从像素空间压缩至潜在空间进行去噪生成。Stable Diffusion v1.5在SD-v1.4基础上,基于LAION数据集进行了更多步数的微调,提升了生成图像的美学质量。Stable Diffusion XL引入更大规模的U-Net主干网络,采用双文本编码器(CLIP-ViT-L/14和OpenCLIP ViT-bigG/14)来增强语义理解能力。Kandinsky 3.0采用Flan-UL2作为文本编码器,通过MoVQ编解码器进行特征的量化与潜空间映射。Kolors采用ChatGLM3-6B-Base作为文本编码器来增强中英文复杂提示词理解能力,利用CogVLM对训练图像进行细粒度重描述,通过两阶段训练、高美学图像数据和高分辨率图像噪声调度优化提升生成图像的真实感和视觉质量。

2)基于Transformer的扩散或流匹配模型,包括PixArt-α(Chen 等,2024)、Hunyuan-DiT(Li 等,2024)、Lumina-Image-2.0(Qin 等,2025)和FLUX.1-schnell(Black Forest Labs,2024)。PixArt-α采用Diffusion Transformer作为主干网络,使用Flan-T5-XXL作为文本编码器,并通过交叉注意力注入文本条件;该模型引入了adaLN-single模块以共享减少冗余参数,采用分阶段训练策略和高信息密度图像描述来提升图文对齐能力和高分辨率图像生成质量。Hunyuan-DiT采用潜空间扩散Transformer架构,结合双语CLIP和多语言T5作为文本编码器,并通过交叉注意力将文本条件注入去噪网络,从而增强中英文提示词理解能力和多分辨率图像生成能力。Lumina-Image-2.0采用Unified Next-DiT架构和Gemma2-2B文本编码器,将文本嵌入和带噪图像潜变量作为统一序列进行联合自注意力建模,引入面向文本到图像任务的UniCap为训练图像生成多粒度、多视角和多语言描述。FLUX.1-schnell基于修正流Transformer架构,参数规模达到120亿,引入了潜在对抗扩散蒸馏技术,能够在极少的推理步数内生成高质量图像。

3)级联式生成模型,包括Stable Cascade(Pernias 等,2024)和DeepFloyd IF(DeepFloyd Lab,2023)。Stable Cascade基于Würstchen架构,由Stage A、Stage B和Stage C三个阶段组成,Stage A和Stage B实现图像的压缩与重建,Stage C在高度压缩的潜在空间中进行文本条件生成。DeepFloyd IF是基于U-Net的级联像素级扩散模型,由T5-XXL文本编码器和三个级联扩散模块组成;该模型不依赖VAE压缩,直接在像素空间中进行扩散建模。

4)跨模态生成模型。CogVideoX(Yang 等,2025)是基于扩散Transformer的文本到视频生成模型,采用T5编码文本提示词,通过3D因果VAE对视频进行时空压缩;其去噪网络采用专家Transformer结构,并引入专家自适应层归一化和3D全注意力机制。

本文在NVIDIA RTX A40 GPU上部署以上12种生成模型,完成伪造图像生成流程。所有生成模型都以第1.2.1节中由视觉语言模型生成的图像描述作为基础提示词,并在此基础上加入质量增强词(如“photorealistic,8k UHD”),以提升生成图像的真实感和视觉质量。在生成过程中,本文对SD-v1.4、SD-v1.5、SDXL使用负向提示词,主要用于进行图像基础质量控制,以减少严重卡通化、主体结构异常以

表1　不同生成模型的参数配置

Table 1　Parameter configurations of different generative models

模型	文本编码器	采样算法	采样步数	引导尺度	分辨率
SD-v1.4	CLIP ViT-L/14	PNDM	50	7.5	512 × 512
SD-v1.5	CLIP ViT-L/14	PNDM	50	7.5	512 × 512
SDXL	CLIP ViT-L/14 & OpenCLIP ViT-bigG/14	Euler a	50	7.5	1024 × 1024
Kolors	ChatGLM3-6B-Base	DPM++ 2M Karras	30	5.0	1024 × 1024
Kandinsky 3.0	Flan-UL2	DDIM	25	4.0	1024 × 1024
Stable Cascade	OpenCLIP ViT-bigG/14	DDPM	30	4.0 (Prior)	1024 × 1024
PixArt-α	Flan-T5-XXL	DPM-Solver++	20	4.5	1024 × 1024
Hunyuan-DiT	Bilingual CLIP & Multilingual T5	DPM++ 2M Karras	30	6.0	1024 × 1024
Lumina-Image-2.0	Gemma2-2B	FlowMatch Euler	50	4.0	1024 × 1024
DeepFloyd IF	T5-XXL	DPM-Solver++ / DDPM	50 (S1) + 50 (S2)	7.0 (S1)	1024 × 1024
FLUX.1-schnell	CLIP ViT-L/14 & T5-XXL	FlowMatch Euler	4	0.0	1024 × 1024
CogVideoX	T5-XXL	DDIM	50	6.0	720 × 480

及明显偏离真实图像语义参照的低质量生成结果。

考虑到各生成模型在底层架构上的差异，本文为不同模型设定了相应的参数配置（见表1）。在生成阶段，各模型以文本提示词为条件输入，在特定的采样器、引导尺度和推理步数下进行生图过程，最终输出PNG格式的高分辨率伪造图像。特别地，对于文本生成视频模型CogVideoX，本文将生成帧数限制为单帧，为数据集引入跨模态伪造图像，以评估检测器对不同类型生成模型的检测能力。

为保证生成图像的语义准确性，在图像生成完成后同样使用多模态大模型LLaVA对伪造图像进行语义类别检查，确保每张图像的内容与目标类别匹配，避免类别泄漏。此外，通过人工检查筛除了明显偏离真实图像语义参照的低质量样本。

2　数据集属性

2.1　数据集统计

本文构建的图像伪造检测数据集共包含532 000张高分辨率图像。其中，真实图像从ImageNet数据集中收集，共计70 000张；伪造图像由12种生成模型生成，共计462 000张。这些生成模型包含潜在扩散模型、扩散Transformer、级联扩散模型、修正流模型和文本生成视频模型等多种生成范式，能够为图像伪造检测任务提供多来源、多架构的伪造样本。

数据集共包含14个语义类别，覆盖了生物、物体、场景等内容。每种生成模型生成的伪造图像都包含14个语义类别，保证数据集在语义内容上的多样性。

2.2　数据集组织与划分

本文构建的图像数据集由真实图像、图像描述提示词以及对应的伪造图像组成。对于第i张真实图像，记为$x_r^{(i)}$，由视觉语言模型生成的细粒度图像描述提示词为$T_{fine}^{(i)}$。对于任意生成模型m，基于$T_{fine}^{(i)}$生成的伪造图像记为$x_{f,m}^{(i)}$。因此，生成模型m对应的语义对齐子集可表示为：

图2　实例级语义对齐数据集样例

Fig. 2　Samples from the instance-level semantic alignment dataset

$$D_{\text{align}}^{m}=\left\{\left(x_r^{(i)},T_{\text{fine}}^{(i)},x_{f,m}^{(i)}\right)\right\}_{i=1}^{N_m}\#(1)$$

式中，N_m表示生成模型m生成的伪造图像数量。

整体语义对齐数据集可以表示为：

$$D_{\text{align}}=\bigcup_{m\in\mathcal{M}}D_{\text{align}}^{m}\#(2)$$

式中，$\mathcal{M}$表示本文采用的12种生成模型集合。

数据集中真伪图像总配对数为：

$$N=\sum_{m\in\mathcal{M}}N_m\#(3)$$

数据集中真实图像保存为JPEG格式，伪造图像保存为PNG格式。图像描述提示词与对应的图像路径统一存储于CSV文件中，便于伪造图像生成、数据读取和后续实验调用。

为支持检测器训练和跨生成模型泛化评估，本文将数据集划分为训练集、验证集和测试集，分别记为D_{train}、D_{val}和D_{test}。

本文对真实图像进行固定划分，其中56 000张用于训练，7 000张用于验证，7 000张用于测试。针对SD-v1.4、SD-v1.5、SDXL、Kandinsky 3.0、Stable Cascade和PixArt-α这6种模型，其伪造图像都按照对应真实图像的划分结果进行归类，使得训练、验证和测试子集中的真伪图像均保持语义对齐关系。每种生成模型都包含56 000张训练伪造图像、7 000张验证伪造图像和7 000张测试伪造图像。因此，训练集共包含392 000张图像，其中包括56 000张真实图像和336 000张伪造图像；验证集共包含49 000张图像，其中包括7 000张真实图像和42 000张伪造图像。

为评估检测器对未见生成模型的泛化能力，Kolors、Hunyuan-DiT、Lumina-Image-2.0、DeepFloyd IF、FLUX.1-schnell和CogVideoX这6种前沿生成模型不参与训练，仅用于测试阶段，每种模型包含7

表2　不同生成模型的数据集划分

Table 2　Dataset partitioning by generative model

模型	训练集	验证集	测试集
真实图像	56 000	7 000	7 000
SD-v1.4	56 000	7 000	7 000
SD-v1.5	56 000	7 000	7 000
SDXL	56 000	7 000	7 000
Kandinsky 3.0	56 000	7 000	7 000
Stable Cascade	56 000	7 000	7 000
PixArt-α	56 000	7 000	7 000
Kolors	无	无	7 000
Hunyuan-DiT	无	无	7 000
Lumina-Image-2.0	无	无	7 000
DeepFloyd IF	无	无	7 000
FLUX.1-schnell	无	无	7 000
CogVideoX	无	无	7 000

000张伪造图像。测试阶段使用全部12种生成模型，因此测试集共包含84 000张伪造图像和7 000张真实图像，总计91 000张测试图像。表2给出了各生成模型的训练集、验证集和测试集的具体划分情况。

2.3　数据集样例

本文选取了5种生成模型（SD-v1.4、Stable Cascade、Kolors、Hunyuan-DiT、Lumina-Image-2.0）来构建数据集样例，直观展示数据集的组成与语义对齐特性。图2展示了部分真实图像、由视觉语言模型针对真实图像生成的图像描述提示词、生成模型生成的伪造图像。

3　实　验

3.1　实验设置

3.1.1　基线模型

本文选取了8种有代表性的检测器进行评估，包括CNN-Spot、DFFreq、SAFE、UnivFD、RINE、Effort、AIDE和ForgeLens。

1）传统基于CNN（convolutional neural networks）的检测方法。CNN-Spot采用预训练的ResNet-50作为骨干网络，在ProGAN生成图像以及对应真实图像上进行二分类训练，学习生成图像中具有共性的伪造痕迹。该方法在训练阶段引入了高斯模糊和JPEG压缩等数据增强操作，提升检测器对未知生成模型的泛化能力。

2）基于频域或图像变换的方法。DFFreq（Yan等，2026）通过双频域分支提取频域伪造特征，利用重构滑动窗口注意力机制进行局部细粒度特征建模，增强对伪造痕迹的捕捉能力。SAFE（Li等，2025）以裁剪操作替代下采样，避免细粒度伪造痕迹被削弱；通过颜色扰动、随机旋转和随机掩码进行数据增强，缓解检测器对无关特征的过拟合，提升其局部感知能力；利用DWT（discrete wavelet transform）高频特征作为输入，增强跨生成模型检测的泛化性能。

3）基于预训练视觉模型的方法。UnivFD利用在大规模图文数据上预训练的视觉语言模型（如CLIP ViT-L/14），提取冻结的特征空间，并直接在此特征空间上使用最近邻分类或线性探测来进行图像真伪鉴别。RINE（Koutlis和Papadopoulos，2024）利用CLIP图像编码器的中间层特征，通过轻量线性映射构建伪造感知向量空间，并结合可训练的重要性权重模块进行加权融合，实现对合成图像的判别。AIDE（Yan等，2025）融合低层分块特征与高层语义特征进行AI生成图像检测，分块特征提取模块捕捉图像局部噪声模式和伪造痕迹，语义特征嵌入模块基于预训练OpenCLIP提取整图语义表示，最终通过特征融合完成真伪判别。Effort（Yan等，2024）基于奇异值分解将预训练视觉模型的权重划分为冻结的主导子空间和可学习的残差子空间，保留预训练语义知识并学习伪造相关模式；通过正交约束减少特征空间退化，缓解对特定伪造模式的过拟合，提升跨生成模型泛化能力。ForgeLens（Chen等，2025）以冻结的CLIP-ViT作为图像编码器，引入权重共享引导模块与伪造感知特征整合机制，在少量训练数据下提升模型的跨域泛化能力。

3.1.2　评估指标

本文采用ACC（accuracy）和AP（average precision）作为主要评价指标。ACC衡量检测器对真伪图像的整体分类正确率；AP反映模型对真伪样本的整体排序区分能力。本文测试结果表格中，各检测器的结果均按照“ACC/AP”的格式给出。

3.1.3　训练数据集

遵循现有方法在图像伪造检测任务中广泛采用的传统训练设置，本文将基础训练集设定为4-class

ProGAN（由car、cat、chair、horse四个类别组成）。该训练设置仅包含单一生成模型和较少语义类别，可能使检测器在训练过程中受到图像语义内容差异的影响，利用语义内容与图像真伪之间的虚假相关性进行判别，而不是稳定地学习生成过程相关的伪造痕迹。

为了分析语义对齐约束对图像伪造检测任务的影响，本文在传统4-class ProGAN训练设置的基础上，引入不同语义对齐粒度的训练数据，并设计三种训练方案：1）传统训练设置（origin）：模型仅使用4-class ProGAN数据集进行训练；2）类别级提示词训练设置（O+C，origin + class-level）：在传统训练设置的基础上加入由类别级提示词生成的SD-v1.4数据，生成图像仅在类别层面与真实图像保持一致，并不以具体真实图像为语义参照；3）实例级语义对齐训练设置（O+A，origin + alignment）：在传统训练设置的基础上加入由实例级图像描述提示词生成的SD-v1.4语义对齐数据，使生成图像以具体真实图像为语义参照，在主体状态、局部细节、背景元素和拍摄风格等维度上形成实例级语义对齐关系。

本文实验中，O+C和O+A均使用猫、车、鱼、鞋子四个语义类别的补充训练数据，并保持新增训练数据量、SD-v1.4生成器和训练策略一致。通过比较O+C和O+A，可以在控制新增数据量、类别数量和生成器来源的条件下，分析实例级语义对齐约束对检测器泛化性能的影响。

3.1.4　评估数据集

李美玲等人（2026）指出，AI生成图像检测器的域外泛化性可通过其在未知生成模型、未知类别等场景下的检测性能进行衡量。为了全面评估检测器的跨生成模型泛化能力，本文选取3个公开基准测试集以及本文构建的AIGI-Align测试集进行实验。其中，公开测试集用于评估检测器在已有基准上的泛化表现，AIGI-Align测试集用于进一步评估检测器在语义对齐场景下的检测能力。

1）ForenSynths包含了BigGAN、CycleGAN、DeepFake、GauGAN、ProGAN、StarGAN、StyleGAN和StyleGAN2共8个测试子集。其真实图像来自LSUN、ImageNet、COCO、FaceForensics++等数据集。

2）UniversalFakeDetect采用了Guided Diffusion、LDM、GLIDE和DALL-E等生成模型，真实图像来源于ImageNet和LAION。本文采用dalle、glide_100_10、glide_100_27、glide_50_27、guided、ldm_100、ldm_200和ldm_200_cfg共8个子集进行测试。

3）GenImage的真实图像来源于ImageNet，包含1000个语义类别，生成器基于类别标签生成图像。伪造图像由BigGAN、GLIDE、VQDM、Stable Diffusion v1.4、Stable Diffusion v1.5、ADM、Midjourney和Wukong共8种生成模型生成。

4）AIGI-Align为本文构建的实例级语义对齐测试集，包含12种生成模型的测试子集，真伪图像保持语义对齐，用于评估检测器在语义捷径被削弱后的跨生成模型泛化能力。

3.1.5　实验参数

本文遵循各个检测器原作者的公开实现，采用原作者建议的超参数设置，保证实验的公平性。所有方法都在3.1.3中提出的三种训练设置下重新训练，训练过程在NVIDIA RTX A5000 GPU上进行。

3.2　语义对齐程度量化分析

为分析不同数据构建方式下真实图像与生成图像之间的语义对齐程度差异，本文以AIGI-Align的SD-v1.4测试子集为基础进行量化分析。该测试子集包含14个语义类别，每类500个实例；每个实例均包含一张真实图像，以及一张由该真实图像的实例级细粒度描述驱动生成的SD-v1.4生成图像。为比较不同语义约束方式对真伪图像语义对齐程度的影响，本文构造四组真伪图像配对关系：

1）实例级语义对齐组（AIGI-Align）。将AIGI-Align的SD-v1.4测试子集中的真实图像与其对应的实例级语义对齐生成图像直接配对。该组用于衡量本文实例级语义对齐构建方式下真实图像与生成图像之间的语义相似度。

2）类别级提示词组（Class-level）。为模拟传统类别级提示词生成方式，本文使用相同的7 000张真实图像作为真实图像锚点，并额外采用SD-v1.4基于类别级提示词生成14类、每类500张图像。这类图像仅与真实图像共享类别标签，并不以某一张具体真实图像的主体姿态、背景元素、光照条件和局部细节为参照。本文将真实图像与同类别的类别级生成图像随机一一配对，构成类别级提示词组，用于衡量仅保持类别一致时的真伪图像语义相似度。

3）跨实例同类组（Cross-instance）。为区分类别一致与实例级对应关系的作用，本文仍使用AIGI-Align的SD-v1.4测试子集中的实例级语义对齐生成

图像，但将真实图像与同类别、非自身对应的生成图像进行错位配对。该组用于衡量同类别、同生成方式，但缺少实例级对应关系时的语义相似度。

4）跨类别随机组（Random）。为获得类别不一致条件下的参照结果，本文将真实图像与不同类别的AIGI-Align实例级生成图像进行跨类别随机配对，构成Random组。该组用于反映跨类别随机匹配时的语义相似度水平。

表3　不同配对方式下真伪图像CLIP语义相似度统计

Table 3　Statistics of CLIP semantic similarity between real and fake images under different pairing strategies

配对方式	语义匹配关系	均值	中位数
随机配对组	跨类别随机匹配	0.5228	0.5239
类别级提示词组	仅类别级一致	0.6622	0.6615
跨实例同类组	同类别但非实例级对应	0.6688	0.6681
实例级语义对齐组	实例级语义对齐	0.7222	0.7261

由此，四组配对数据分别对应实例级语义对齐、仅类别级一致、同类别但非实例级对应和跨类别随机匹配四种语义匹配关系。该实验能够在同一批真实图像锚点和同一生成模型SD-v1.4的基础上，比较不同提示词粒度和不同配对关系对真伪图像语义相似度的影响，从而验证AIGI-Align是否能够提高真实图像与生成图像之间的实例级语义对齐程度。

在语义相似度计算方面，本文采用预训练CLIP ViT-L/14图像编码器提取真实图像和生成图像的图像特征，并对特征进行L2归一化，使用归一化特征之间的余弦相似度作为真伪图像对的CLIP语义相似度。该指标用于衡量两张图像在语义特征空间中的相对接近程度。因此，本文并不要求真实图像与生成图像在光照、背景或局部细节上完全相同，而是关注不同数据构建方式下真伪图像在主体内容、外观属性和场景语义等方面的相对对齐程度。

如表3和图3所示，实例级语义对齐组的总体平均CLIP语义相似度为0.7222，高于跨类别随机组的0.5228、类别级提示词组的0.6622和跨实例同类组的0.6688。与类别级提示词组和跨实例同类组相比，实例级语义对齐组的平均相似度分别提升0.0600和0.0534。逐类别结果如图4所示，实例级语义对齐组在14个类别上的平均相似度均高于其他三组，表明本文方法在不同语义类别中均具有更高的语义对齐程度。

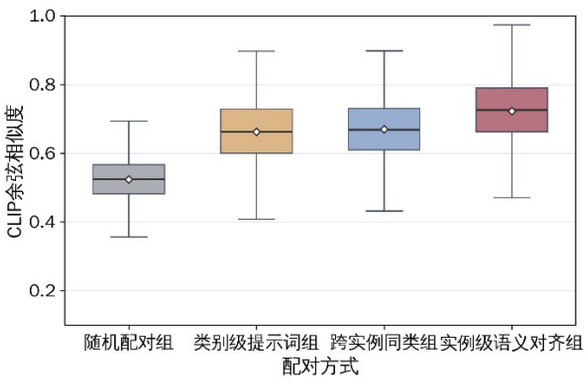

图3　不同配对方式下真伪图像CLIP语义相似度分布

Fig. 3　Distribution of CLIP semantic similarity between real and fake images under different pairing strategies

上述结果表明，类别级提示词生成方式下，真实图像与生成图像之间仍可能存在较明显的语义内容差异；即使使用同类别的实例级生成图像，如果缺少与对应真实图像的匹配关系，真实图像与生成图像之间的语义相似度仍然较低。相比之下，AIGI-Align能够在一定程度上减小这种差异，提高真伪图像之间的语义对齐程度。

3.3　语义偏见分析

3.2节已经表明，不同提示词粒度和不同配对关系会影响真实图像与生成图像之间的语义内容差异。本节进一步分析这种语义差异是否会影响检测器的真伪判别结果。为此，本文在保持真实图像、语义类别、生成模型和检测器权重一致的条件下，仅改变生成图像与真实图像之间的语义对齐程度，并比较检测器在不同测试条件下的性能变化。若检测器在语义差异更明显的测试条件下获得更高性能，则说明其判别结果可能受到真伪图像语义差异的影响。

具体地，本文构造以下两种测试条件：

1）实例级语义对齐测试条件（T-align）。该测试条件使用3.2节实例级语义对齐组中的生成图像，即由对应真实图像的细粒度视觉描述驱动SD-v1.4生成的图像。该条件下，生成图像以具体真实图像为语义参照，与对应真实图像具有更高的实例级语义对齐程度。

2）类别级提示词测试条件（T-class）。该测试条

件使用3.2节类别级提示词组中的生成图像，即由SD-v1.4基于类别级提示词生成的图像。与T-align相比，该条件只保持生成图像与真实图像在类别层面一致，用于模拟传统类别级提示词生成方式下的测试场景。

两个测试条件均使用相同的7 000张真实图像，具有相同的类别构成和每类样本规模；两组生成图像均由SD-v1.4生成。因此，T-align与T-class的主要区别在于生成图像与真实图像之间的语义对齐程度不同。每个测试条件均包含7 000张真实图像和7 000张生成图像。测试时，各检测器均采用origin训练设置下得到的固定模型权重。

表4给出了不同检测器在T-align和T-class两种测试条件下的检测结果。整体来看，在8种代表性检测器中，有6种检测器在T-class条件下的ACC

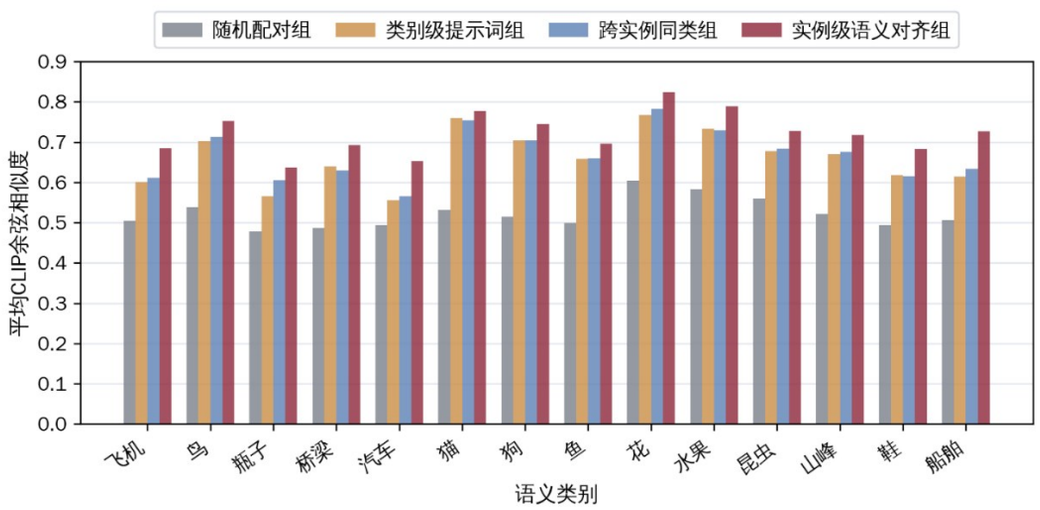

图4　不同语义类别下真伪图像CLIP语义相似度对比

Fig. 4　Comparison of CLIP semantic similarity between real and fake images across different semantic categories

表4　不同语义对齐测试条件下的语义偏见分析结果（ACC/AP，%）

Table 4　Semantic bias analysis results under different semantic alignment test conditions (ACC/AP, %)

模型	T-align测试条件	T-class测试条件	ΔACC	ΔAP
CNN-Spot	50.81/57.54	50.46/56.28	−0.35	−1.26
UnivFD	48.87/43.21	48.68/44.12	−0.19	+0.91
RINE	77.52/98.15	81.12/98.53	+3.60	+0.38
AIDE	83.46/94.71	87.04/96.18	+3.58	+1.47
Effort	65.24/94.13	67.84/95.44	+2.60	+1.31
SAFE	99.02/99.82	99.15/99.84	+0.13	+0.02
DFFreq	94.50/99.36	94.56/99.38	+0.06	+0.02
ForgeLens	90.89/98.19	92.24/98.26	+1.35	+0.07
模型平均	76.29/85.64	77.64/86.00	+1.35	+0.36

注：ΔACC和ΔAP表示T-class测试条件相对于T-align测试条件的性能差值。

高于T-align条件，有7种检测器在T-class条件下的AP高于T-align条件。从模型平均结果来看，T-align条件下平均ACC为76.29%，平均AP为85.64%；T-class条件下平均ACC为77.64%，平均AP为86.00%。

上述结果表明，当真实图像与生成图像之间存在更明显的语义内容差异时，检测器更容易获得较高的检测性能；而当实例级语义对齐削弱这种差异后，检测性能有所下降。由此可见，检测器在该测试条件下表现出一定的语义偏见倾向，即可能利用真伪图像之间的语义内容差异进行捷径判别。

3.4　公开数据集的泛化性评估

本文在公开数据集上分析不同训练设置下各检测器的性能变化，以验证所构建的语义对齐数据对提升检测模型跨生成模型泛化能力的有效性。本文对比了多种基线方法，包括CNN-Spot、UnivFD、RINE、AIDE、Effort、SAFE、DFFreq以及ForgeLens，并按照3.1.3节中的origin、O+C和O+A三种训练设置

进行评估。其中,O+C作为类别级提示词训练设置,用于控制新增训练数据量、类别数量和生成器来源的影响;O+A作为实例级语义对齐训练设置,用于进一步分析实例级语义对齐数据对检测器泛化性能的影响。

在ForenSynths测试集上,本文评估了现有检测器在多种GAN(generative adversarial network)架构上的泛化性能。在ProGAN域内测试中,由于origin、O+C和O+A三种训练设置都采用4-class ProGAN作为基础训练数据,各方法在ProGAN测试子集上都取得了优异表现。对于未见的GAN测试子集,O+C和O+A设置下部分检测器取得了较好的效果,比如DFFreq在BigGAN和GauGAN等子集上的检测性能有明显的提升,SAFE也在多数未见子集上实现了AP的提升,基于CLIP的ForgeLens和Effort在多数GAN子集上检测能力都有所提高。从整体平均性能来看,O+A设置下ForgeLens在所有检测器中表现最优,其平均ACC为96.37%,平均AP达到99.63%;与O+C相比,Effort和ForgeLens在O+A设置下的平均ACC和平均AP均有所提升,DFFreq的平均AP也有所提升,这说明实例级语义对齐数据能有效增强部分模型对未知GAN架构的泛化能力。然而,与O+C相比,O+A在ForenSynths上的总体优势并不明显,部分检测器如CNN-Spot和UnivFD在O+A设置下的平均ACC和平均AP均低于O+C。这可能与ForenSynths主要由GAN和DeepFake测试子集构成有关,其伪造痕迹分布与本文新增的SD-v1.4扩散模型训练数据存在差异,因此实例级语义对齐数据在该测试集上的作用受到测试域差异限制。

本文在UniversalFakeDetect测试集上评估了检测器对多种扩散模型的泛化表现。相较于传统训练设置,实例级语义对齐数据的引入增强了多数模型在DALLE与LDM系列测试中的检测能力。不过,部分基于CLIP的方法(如UnivFD、RINE和AIDE)在ADM子集上性能下降。这说明实例级语义对齐数据虽能削弱语义差异对检测器判别过程的干扰,但也可能会使检测器对训练域内的扩散去噪机制产

表5　ForenSynths测试结果(ACC/AP,%)

Table 5　Results on ForenSynths (ACC/AP, %)

模型	方法	ProGAN	StyleGAN	StyleGAN2	BigGAN	CycleGAN	StarGAN	GauGAN	DeepFake	Mean
CNN-Spot	origin	97.90/99.85	72.73/94.63	71.94/93.75	58.30/72.35	71.12/82.02	73.19/89.01	67.08/80.56	52.41/62.83	70.58/84.37
	O+C	96.86/99.56	73.94/93.90	71.68/92.82	59.75/74.02	72.52/87.97	86.94/95.25	59.65/80.68	54.34/80.13	71.96/88.04
	O+A	97.81/99.73	73.70/92.26	69.29/90.77	54.23/68.03	67.26/81.07	79.84/90.08	55.33/75.42	58.26/72.23	69.46/83.70
UnivFD	origin	99.29/99.98	74.79/95.50	69.07/95.10	95.23/99.13	95.69/99.71	91.05/99.79	98.99/99.98	78.54/89.11	87.83/97.29
	O+C	98.40/99.91	73.74/90.85	71.61/92.07	93.08/98.49	94.13/99.62	87.27/99.30	98.19/99.94	78.65/87.70	86.88/95.98
	O+A	98.41/99.92	74.08/90.17	71.98/91.36	92.20/98.39	92.66/99.60	86.07/99.27	97.16/99.93	68.18/88.51	85.09/95.89
RINE	origin	100.00/100.00	92.02/99.77	94.13/99.95	99.62/99.98	99.55/99.98	99.92/100.00	99.86/99.96	62.85/97.69	93.49/**99.67**
	O+C	99.99/100.00	90.26/99.72	92.95/99.98	98.22/99.95	99.70/100.00	99.87/100.00	99.86/100.00	66.42/97.19	93.41/99.61
	O+A	100.00/100.00	90.91/99.72	91.85/99.85	99.10/99.96	99.62/99.99	99.72/99.99	99.46/99.92	67.31/97.60	93.50/99.63
AIDE	origin	99.88/100.00	99.57/99.99	97.10/99.94	90.93/98.13	98.22/99.83	99.67/99.99	86.82/99.89	55.10/79.53	90.91/97.16
	O+C	99.85/100.00	95.08/99.93	95.32/99.93	95.70/99.36	96.52/99.78	98.42/99.99	96.81/99.84	56.74/84.29	91.80/97.89

表5续表

模型	方法	ProGAN	StyleGAN	StyleGAN2	BigGAN	CycleGAN	StarGAN	GauGAN	DeepFake	Mean
Effort	O+A	99.73/ 100.00	98.46/ 99.97	95.82/ 99.87	93.98/ 98.57	94.97/ 99.53	98.70/ 99.99	90.79/ 99.17	57.87/ 79.22	91.29/ 97.04
	origin	99.98/ 100.00	88.78/ 98.77	92.66/ 99.42	99.40/ 99.99	99.92/ 100.00	99.87/ 100.00	99.86/ 100.00	74.12/ 96.29	94.32/ 99.31
	O+C	99.94/ 100.00	88.91/ 99.51	90.70/ 99.42	99.55/ 99.99	100.00/ 100.00	100.00/ 100.00	99.72/ 99.99	84.53/ 97.95	95.42/ 99.61
	O+A	99.99/ 100.00	92.54/ 99.84	93.89/ 99.62	99.15/ 99.99	99.96/ 100.00	100.00/ 100.00	99.56/ 99.99	81.63/ 97.71	95.84/ 99.64
SAFE	origin	99.83/ 100.00	97.86/ 99.92	98.37/ 99.98	86.60/ 90.69	98.15/ 99.42	99.92/ 100.00	87.65/ 95.17	90.32/ 94.60	94.84/ 97.47
	O+C	99.95/ 100.00	97.41/ 99.85	99.19/ 99.99	89.30/ 95.60	97.92/ 99.85	100.00/ 100.00	91.46/ 96.44	73.60/ 95.55	93.60/ 98.41
	O+A	99.79/ 100.00	94.20/ 99.79	97.15/ 99.96	90.33/ 97.04	98.26/ 99.54	99.97/ 100.00	86.52/ 96.36	90.75/ 94.95	94.62/ 98.46
DFFreq	origin	99.62/ 99.99	99.67/ 99.99	99.80/ 100.00	84.60/ 97.31	95.84/ 99.33	87.02/ 95.90	82.72/ 99.09	68.01/ 74.64	89.66/ 95.78
	O+C	99.50/ 99.99	99.71/ 99.99	99.64/ 100.00	92.07/ 97.58	93.07/ 99.27	98.80/ 99.94	94.14/ 99.09	63.70/ 69.56	92.58/ 95.68
	O+A	99.28/ 99.98	99.62/ 99.99	99.29/ 99.99	92.30/ 97.65	89.52/ 98.48	91.35/ 99.35	95.16/ 99.37	69.42/ 76.71	91.99/ 96.44
ForgeLens	origin	99.95/ 100.00	90.49/ 98.74	96.18/ 99.40	97.15/ 91.11	99.47/ 99.86	99.12/ 99.64	97.57/ 92.80	93.38/ 97.45	**96.66**/ 97.38
	O+C	99.96/ 100.00	87.98/ 99.10	90.70/ 99.34	97.45/ 99.76	99.96/ 100.00	100.00/ 100.00	98.31/ 99.66	90.12/ 98.57	95.56/ 99.55
	O+A	99.99/ 100.00	90.54/ 99.50	93.98/ 99.81	96.28/ 99.95	99.89/ 100.00	100.00/ 100.00	97.69/ 99.91	92.60/ 97.84	96.37/ 99.63

注：加粗字体表示最优结果。

生依赖，导致其在面对未见的、更具挑战性的架构时泛化能力受限。在整体平均指标上，ForgeLens在传统训练设置下已经取得了较高性能，引入实例级语义对齐数据后平均ACC从96.79%提升至97.95%，平均AP达到99.91%，取得在该测试集上的最佳性能。同时，基于传统的卷积网络（CNN-Spot）、基于图像变换的方法（SAFE）、基于预训练CLIP的Effort和RINE，在实例级语义对齐训练设置下整体平均指标均有所提升，说明实例级语义对齐数据能够提升现有检测器在扩散模型测试集上的泛化表现。进一步比较O+C和O+A可以发现，O+A在多数检测器上的平均性能高于O+C。具体而言，8种检测器在O+A设置下的平均ACC均高于O+C，6种检测器的平均AP高于O+C；其中，ForgeLens的平均ACC由O+C的96.14%提升至O+A的97.95%，Effort由94.28%提升至95.50%，SAFE由95.59%提升至96.64%。这说明，在扩散模型测试场景中，实例级语义对齐数据相较类别级提

表6　UniversalFakeDetect测试结果（ACC/AP，%）

Table 6　Results on UniversalFakeDetect (ACC/AP, %)

模型	方法	DALLE	Glide_ 100_10	Glide_ 100_27	Glide_ 50_27	ADM	LDM_100	LDM_200	LDM_ 200_cfg	Mean
CNN- Spot	origin	51.95/ 54.61	56.70/ 75.58	55.30/ 73.77	57.00/ 78.14	56.60/ 70.14	53.25/ 64.84	53.40/ 64.25	53.45/ 65.00	54.71/ 68.29

表6续表

模型	方法	DALLE	Glide_100_10	Glide_100_27	Glide_50_27	ADM	LDM_100	LDM_200	LDM_200_cfg	Mean
	O+C	51.75/59.72	52.80/65.43	51.90/62.98	51.85/62.50	54.25/67.63	55.80/70.73	55.80/70.87	56.20/71.78	53.79/66.45
	O+A	55.25/68.43	52.65/67.84	52.80/65.56	53.95/68.72	55.30/68.55	58.50/76.23	59.55/77.37	65.85/84.54	56.73/72.15
UnivFD	origin	91.20/98.08	83.20/94.85	83.05/94.58	85.30/95.87	78.60/91.83	96.75/99.51	96.75/99.61	80.85/93.73	86.96/96.01
	O+C	86.65/95.39	85.30/95.06	85.75/94.92	86.95/95.57	72.00/84.62	92.50/98.28	93.35/98.52	76.25/89.44	84.84/93.97
	O+A	89.90/96.76	88.55/96.57	88.00/96.24	90.25/97.04	73.70/84.88	93.35/98.56	94.30/98.86	78.95/91.06	87.13/95.00
RINE	origin	94.20/99.42	81.30/97.63	78.90/97.42	83.50/98.16	75.55/97.31	98.65/99.94	98.35/99.89	90.75/99.00	87.65/98.60
	O+C	98.05/99.94	77.25/98.73	75.50/98.55	79.60/99.22	68.05/97.45	99.20/99.99	99.20/99.97	95.90/99.85	86.59/99.21
	O+A	98.65/99.92	84.95/98.72	82.15/98.50	87.00/99.12	70.90/96.82	99.25/99.95	98.85/99.90	95.55/99.59	89.66/99.06
AIDE	origin	96.20/98.58	97.40/99.29	97.35/99.19	97.60/99.31	82.20/94.08	98.20/99.31	97.80/99.27	96.05/98.70	95.35/98.47
	O+C	93.85/98.90	97.45/99.49	97.20/99.51	97.65/99.56	78.50/97.68	98.65/99.76	98.40/99.69	96.15/99.37	94.73/99.24
	O+A	94.55/98.87	97.60/99.60	98.25/99.61	98.45/99.66	80.10/96.26	98.15/99.67	97.85/99.61	96.30/99.29	95.16/99.07
Effort	origin	92.50/99.64	86.05/99.34	86.40/99.32	86.30/99.46	62.05/90.81	97.10/99.85	97.30/99.89	85.75/98.71	86.68/98.38
	O+C	99.55/100.00	96.20/99.89	95.15/99.74	96.05/99.92	70.70/97.07	99.55/99.99	99.45/99.98	97.60/99.92	94.28/99.56
	O+A	100.00/100.00	96.20/99.88	95.45/99.73	97.35/99.95	75.95/97.47	99.85/100.00	99.80/99.98	99.40/99.99	95.50/99.62
SAFE	origin	96.80/99.45	96.55/98.95	95.40/98.80	95.95/98.96	80.30/94.11	98.25/99.80	98.25/99.80	98.00/99.73	94.94/98.70
	O+C	97.65/99.34	97.20/99.06	94.85/98.51	96.05/98.92	82.65/97.53	98.80/99.54	98.75/99.55	98.75/99.54	95.59/99.00
	O+A	98.80/99.86	98.10/99.35	96.70/98.93	97.25/99.07	85.25/96.13	99.05/99.92	99.00/99.90	98.95/99.87	96.64/99.13
DFFreq	origin	91.30/97.90	98.20/99.88	98.30/99.78	98.40/99.78	93.55/98.84	99.65/99.99	99.45/99.97	99.50/99.91	97.29/99.51
	O+C	89.90/98.61	97.65/99.90	97.25/99.73	98.10/99.82	93.55/99.21	99.80/100.00	99.55/99.99	99.35/99.88	96.89/99.64
	O+A	95.65/99.75	97.15/99.78	95.95/99.43	96.75/99.73	92.90/99.21	99.55/100.00	99.40/99.98	99.35/99.97	97.09/99.73
ForgeLens	origin	98.60/99.32	98.60/99.33	98.30/99.24	98.80/99.32	82.55/95.26	99.35/99.27	99.50/99.32	98.60/99.13	96.79/98.77

表6续表

模型	方法	DALLE	Glide_100_10	Glide_100_27	Glide_50_27	ADM	LDM_100	LDM_200	LDM_200_cfg	Mean
	O+C	99.80/100.00	98.15/99.99	97.65/99.92	98.40/99.96	77.95/99.14	99.45/99.93	99.40/99.97	98.35/99.83	96.14/99.84
	O+A	99.90/100.00	99.40/100.00	98.95/99.99	99.25/99.99	87.05/99.33	99.75/100.00	99.80/100.00	99.50/99.99	**97.95/99.91**

注：加粗字体表示最优结果。

示词数据更有利于提升多数检测器的泛化表现。

在GenImage测试集上，实例级语义对齐数据的引入对检测器的泛化能力产生了较大影响。在SD-v1.4域内测试子集及同源的SD-v1.5测试子集上，各类检测器在引入实例级语义对齐数据后性能接近饱和。在跨域泛化测试中，实例级语义对齐数据提高了检测器对Midjourney、Glide、Wukong等未见扩散模型的检测能力。同时，实验结果也反映了本文策略在特定生成架构上的局限性，部分模型在面对ADM和VQDM子集时性能下降。在整体平均性能方面，多数检测器在实例级语义对齐训练设置下实现了检测水平的提升。相较于传统训练设置，引入实例级语义对齐数据后，DFFreq的平均ACC由87.47%提升至93.23%，平均AP由94.25%提升至99.22%，取得了最优的平均AP；SAFE取得了最优的平均ACC；基于CLIP的ForgeLens和Effort取得了较高的平均AP。上述结果表明，实例级语义对齐数据有助于减少检测器对语义差异的依赖，缓解语

表7　GenImage测试结果（ACC/AP，%）

Table 7　Results on GenImage（ACC/AP，%）

模型	方法	Midjourney	SD-v1.4	SD-v1.5	ADM	Glide	Wukong	VQDM	BigGAN	Mean
CNN-Spot	origin	51.42/59.84	50.37/56.07	50.24/56.28	55.88/67.85	52.91/64.43	50.40/55.42	53.77/65.97	51.31/70.38	52.03/62.03
	O+C	52.49/68.79	66.27/88.37	66.50/88.30	53.81/68.99	55.26/76.97	59.45/80.37	50.13/58.31	55.43/81.31	57.42/76.43
	O+A	56.10/76.25	75.78/92.88	76.26/92.98	52.99/66.99	53.57/71.56	68.15/87.74	51.85/63.11	50.33/60.02	60.63/76.44
UnivFD	origin	49.52/44.44	51.33/53.98	51.26/54.29	74.58/90.12	65.60/81.64	55.86/67.71	90.56/97.68	87.48/96.49	65.77/73.30
	O+C	56.97/66.85	66.11/80.06	65.94/79.56	69.61/82.25	72.23/85.23	67.59/81.19	84.42/93.75	87.03/95.17	71.24/83.01
	O+A	60.17/67.59	70.45/80.09	70.22/80.20	71.90/82.33	73.66/84.33	70.83/81.18	85.56/93.90	86.33/94.53	73.64/83.02
RINE	origin	56.27/88.14	82.08/98.54	81.58/98.44	71.50/96.39	65.76/95.80	81.71/98.50	87.69/99.24	90.51/99.61	77.14/96.83
	O+C	65.19/96.83	98.73/99.97	98.54/99.91	64.89/97.25	63.75/97.87	96.61/99.92	83.93/99.49	85.37/99.65	82.13/98.86
	O+A	71.31/96.78	99.53/99.98	99.28/99.89	71.68/96.81	77.00/98.58	98.35/99.93	89.61/99.57	91.46/99.76	87.28/98.91
AIDE	origin	66.13/83.14	82.87/93.79	82.04/93.57	87.50/96.14	91.44/97.26	84.60/94.96	93.48/98.57	93.93/98.80	85.25/94.53
	O+C	64.88/89.80	95.36/99.64	95.18/99.63	82.70/98.11	93.28/99.37	93.60/99.50	91.99/99.47	92.40/99.47	88.67/98.12

表7续表

模型	方法	Midjourney	SD-v1.4	SD-v1.5	ADM	Glide	Wukong	VQDM	BigGAN	Mean
Effort	O+A	79.38/ 93.30	95.47/ 99.46	95.34/ 99.42	84.59/ 97.08	94.91/ 99.34	94.22/ 99.31	91.71/ 98.94	93.87/ 99.20	91.18/ 98.26
	origin	50.77/ 74.14	72.63/ 96.93	72.44/ 96.83	62.38/ 90.52	77.83/ 97.46	71.62/ 96.56	75.64/ 97.05	97.47/ 99.96	72.60/ 93.68
	O+C	61.58/ 93.19	99.65/ 100.00	99.61/ 99.99	78.47/ 97.18	93.95/ 99.85	98.82/ 99.99	94.65/ 99.85	98.80/ 99.99	90.69/ 98.76
	O+A	71.01/ 95.24	99.95/ 100.00	99.88/ 99.99	82.40/ 97.61	93.35/ 99.77	99.78/ 100.00	98.28/ 99.95	99.70/ 100.00	93.04/ 99.07
SAFE	origin	94.07/ 99.00	98.68/ 99.81	98.58/ 99.73	83.66/ 96.18	96.16/ 98.93	98.23/ 99.65	95.24/ 98.97	97.06/ 99.42	95.21/ 98.96
	O+C	96.22/ 99.87	99.83/ 99.96	99.76/ 99.94	71.71/ 92.52	96.55/ 99.48	99.57/ 99.96	96.43/ 99.62	97.27/ 99.91	94.67/ 98.91
	O+A	93.64/ 99.49	99.71/ 99.99	99.62/ 99.91	74.77/ 89.10	97.86/ 99.48	99.53/ 99.89	98.74/ 99.82	97.87/ 99.73	**95.22**/ 98.43
DFFreq	origin	93.62/ 99.09	93.03/ 98.88	93.08/ 98.85	93.09/ 98.09	93.63/ 99.08	93.12/ 98.47	94.97/ 99.24	45.23/ 62.29	87.47/ 94.25
	O+C	97.00/ 99.62	98.11/ 99.81	98.04/ 99.76	88.61/ 98.57	96.57/ 99.65	97.10/ 99.55	95.88/ 99.61	49.18/ 70.78	90.06/ 95.92
	O+A	98.12/ 99.84	99.11/ 99.96	99.01/ 99.90	90.67/ 99.00	97.38/ 99.73	98.35/ 99.88	90.86/ 99.06	72.35/ 96.42	93.23/ **99.22**
ForgeLens	origin	59.37/ 80.48	95.23/ 98.20	95.19/ 97.86	83.91/ 96.29	97.20/ 98.18	92.81/ 97.87	87.62/ 96.61	99.41/ 98.57	88.84/ 95.51
	O+C	64.14/ 92.38	99.70/ 100.00	99.58/ 99.91	84.35/ 99.22	97.72/ 99.93	99.15/ 99.98	94.54/ 99.89	98.87/ 99.89	92.26/ 98.90
	O+A	72.68/ 93.23	99.92/ 100.00	99.84/ 99.97	89.50/ 99.25	99.01/ 99.95	99.80/ 99.99	98.58/ 99.96	99.82/ 99.99	94.89/ 99.04

注：加粗字体表示最优结果。

义偏见问题，并提升其在多生成模型测试场景下的跨域泛化能力。从类别级提示词训练设置和实例级语义对齐训练设置的对比结果来看，在实例级语义对齐训练设置下，8种检测器的平均ACC均高于类别级提示词训练设置，7种检测器的平均AP也高于类别级提示词训练设置。具体地，RINE的平均ACC由82.13%提升至87.28%，Effort由90.69%提升至93.04%，DFFreq由90.06%提升至93.23%，ForgeLens由92.26%提升至94.89%。这一结果说明，与类别级提示词数据相比，实例级语义对齐约束能够进一步提升检测器在GenImage测试集上的跨生成模型泛化表现。

3.5 AIGI-Align数据集的泛化性评估

为了进一步研究现有检测器在面对前沿生成架构和语义对齐测试场景时的表现，本文评估了各检测器在本文构建的AIGI-Align测试集上的性能。表8、表9和表10分别展示了各检测器在传统训练设置（origin）、类别级提示词训练设置（O+C）和实例级语义对齐训练设置（O+A）下的性能表现。

在传统训练设置下，不同类型的检测器在各测试子集和整体平均性能上表现出明显的差异，如表8所示。传统的基于CNN的检测器（CNN-Spot）在多数未见架构上泛化性受限，在Kolors、Hunyuan-DiT、FLUX.1-schnell等子集上的ACC较低，整体平均ACC仅为51.15%。基于预训练视觉模型的UnivFD在多数未见架构上存在泛化问题，整体平均ACC为50.24%；RINE、Effort、ForgeLens表现出一定的跨域检测能力，但在Kolors和FLUX.1-schnell上未能取得理想的检测效果。基于频域与图像变换的检测器（DFFreq、SAFE）性能较强，整体平均ACC分别达到

了93.42%和96.59%；不过SAFE在DeepFloyd IF子集上的ACC出现了明显下降。上述结果说明，仅依赖4-class ProGAN训练数据难以充分适应AIGI-Align中前沿生成模型和语义对齐测试场景带来的泛化挑战。

引入类别级提示词生成数据后，多数检测器的跨域泛化性能有所提升，如表9所示。与传统训练设置相比，CNN-Spot的平均ACC由51.15%提升至59.92%，UnivFD由50.24%提升至80.84%，RINE由67.72%提升至88.88%，Effort由60.33%提升至88.24%，ForgeLens由82.09%提升至91.63%。这说明，在传统训练设置基础上引入SD-v1.4类别级生成数据，能够增强检测器对AIGI-Align中部分未见生成模型的检测能力。需要注意的是，类别级提示词生成图像仅与真实图像保持类别层面一致，真伪图像之间仍可能存在实例级语义差异。

引入本文构建的实例级语义对齐数据后，多数检测器的跨域泛化性能进一步提升，如表10所示。传统的CNN-Spot在引入实例级语义对齐数据后，在SD-v1.4和SD-v1.5子集上的ACC提升至87%以上，整体平均ACC提升到65.94%。对于基于CLIP的检测器，UnivFD和Effort的整体平均ACC分别提升至84.81%和93.40%。ForgeLens、RINE和AIDE在多数生成子集上实现了指标增长，其中RINE和Effort的整体平均AP达到了99.74%。在跨模态的CogVideoX子集上，Effort和ForgeLens的ACC均超过了99.50%。对于基于频域与图像变换的方法，DFFreq的平均ACC提升至97.08%；SAFE的整体平均ACC达到96.80%。上述结果表明，实例级语义对齐数据能够提升检测器在前沿生成模型和语义对齐测试场景下的泛化能力。

3.6 训练数据控制消融分析

为进一步区分性能提升是否仅来自新增训练数据量、类别数量或生成器样本，本文汇总表5至表10中8种检测器在不同测试集上的平均性能，结果如表11所示。

引入补充训练数据后，多数测试集上的总体平均性能有所提升。其中，类别级提示词训练设置相较传统训练设置在GenImage和AIGI-Align测试集上提升较为明显，说明在传统训练数据基础上加入SD-v1.4类别级生成数据，能够增强检测器对部分未见生成模型的检测能力。这也表明，新增生成数据本身会对检测器的跨域泛化性能产生影响。

在新增训练数据量、类别数量和生成器来源保持一致的条件下，实例级语义对齐训练设置在UniversalFakeDetect、GenImage和AIGI-Align测试集上均取得高于类别级提示词训练设置的总体平均性能，说明实例级语义对齐约束本身仍然能够发挥作用。其中，AIGI-Align测试集上的提升最为明显，表明实例级语义对齐训练数据在语义对齐测试场景中具有更突出的作用。因此，本文方法的性能提升不能简单归因于训练数据数量、语义类别或生成器来源的变化。

需要注意的是，实例级语义对齐训练设置在ForenSynths测试集上的总体平均性能未高于类别级提示词训练设置。这可能是由于ForenSynths主要包含GAN和DeepFake类型图像，其伪造痕迹分布与本文新增的SD-v1.4扩散模型训练数据存在差异。因此，实例级语义对齐训练数据的作用并不一定在所有生成模型类型上都体现，其优势主要体现在语义差异可能干扰检测判断的测试场景以及扩散模型相关的跨域测试场景中。总体而言，在训练阶段引入实例级语义对齐的真实与伪造样本，有助于减少检测器对语义差异的依赖，缓解语义偏见问题，引导模型更加关注生成过程相关的伪造痕迹，从而提升其在未知生成模型和语义对齐测试场景下的泛化能力。

4 结 论

本文围绕AIGI检测中的语义偏见问题展开研究。现有数据集缺乏对真实图像与伪造图像之间细粒度语义差异的控制，容易使语义内容与真伪标签之间形成虚假关联，导致检测器依赖图像语义内容而非伪造痕迹进行判别。为缓解这一问题，本文构建了语义对齐训练与评估数据集AIGI-Align，并围绕语义对齐程度、语义偏见表现和跨生成模型泛化能力进行了实验。实验结果表明，AIGI-Align能够提高真实图像与伪造图像之间的实例级语义对齐程度，削弱语义内容差异对检测器判别过程的干扰。在语义偏见分析中，部分检测器在语义差异更明显的测试条件下取得更高检测性能，表明现有检测器可能利用真伪图像之间的语义差异进行捷径判别；在引入实例级语义对齐训练数据后，多数检测器的

表8　传统训练设置(origin)下AIGI-Align测试结果(ACC/AP,%)

Table 8　Detection results on AIGI-Align under the traditional training setting (origin) (ACC/AP, %)

生成模型	CNN-Spot	UnivFD	RINE	AIDE	Effort	SAFE	DFFreq	ForgeLens
Kolors	49.65/48.80	48.90/42.04	59.11/89.12	79.39/93.98	53.16/75.19	99.14/99.87	94.19/99.08	69.49/82.49
Lumina-Image-2.0	56.05/66.15	51.84/60.78	68.79/95.90	85.97/96.54	65.14/91.47	99.18/99.91	94.91/99.56	85.64/94.57
Hunyuan-DiT	52.62/56.50	49.22/50.03	66.22/95.19	84.86/95.67	59.59/89.83	98.87/99.84	94.46/99.22	80.64/96.00
Stable Cascade	51.44/54.61	50.84/58.23	90.55/99.47	95.01/98.74	61.04/98.28	99.22/99.92	93.57/98.88	87.88/98.58
PixArt-α	51.47/56.04	51.06/59.96	70.39/96.70	75.35/91.97	59.69/91.47	98.84/99.82	92.89/98.52	80.61/93.15
Kandinsky 3.0	49.17/45.10	51.36/62.05	67.45/95.05	78.04/93.72	59.61/92.46	98.96/99.86	93.93/99.27	85.10/96.41
DeepFloyd IF	51.78/62.44	54.09/66.94	70.92/97.25	87.61/97.02	60.69/92.72	71.59/89.81	90.17/96.85	74.17/88.15
FLUX.1-schnell	49.64/45.34	49.51/46.55	56.40/88.76	77.79/93.16	52.39/78.55	99.19/99.93	94.21/99.20	70.74/83.19
SD-v1.4	50.81/57.54	48.87/43.21	77.52/98.15	83.46/94.71	65.24/94.13	99.02/99.82	94.50/99.36	90.89/98.19
SD-v1.5	51.14/58.17	48.86/43.61	78.14/98.17	84.18/94.98	64.76/94.31	98.98/99.83	94.49/99.33	90.40/98.21
SDXL	48.31/47.26	48.65/43.52	53.20/85.69	75.25/88.22	56.41/89.90	99.10/99.90	94.79/99.50	73.04/90.68
CogVideoX	51.78/53.79	49.66/50.66	53.91/88.94	73.54/90.75	66.18/95.53	96.94/99.37	88.98/96.43	96.51/98.98
Mean	51.15/54.31	50.24/52.30	67.72/94.03	81.70/94.12	60.33/90.32	**96.59/98.99**	93.42/98.77	82.09/93.22

注:加粗字体表示最优结果。

表9　类别级提示词训练设置(O+C)下AIGI-Align测试结果(ACC/AP,%)

Table 9　Detection results on AIGI-Align under the class-level prompt training setting (O+C) (ACC/AP, %)

生成模型	CNN-Spot	UnivFD	RINE	AIDE	Effort	SAFE	DFFreq	ForgeLens
Kolors	60.75/79.28	79.85/89.66	88.81/99.59	83.37/97.90	78.96/98.28	99.57/99.79	97.04/99.63	81.38/95.42
Lumina-Image-2.0	56.40/73.19	83.14/92.10	87.13/99.63	88.84/98.87	89.71/99.70	99.66/99.81	98.26/99.81	93.89/99.58
Hunyuan-DiT	60.62/78.35	84.21/92.04	87.59/99.57	88.20/98.85	87.31/99.71	99.28/99.78	96.91/99.60	91.20/99.73
Stable Cascade	53.26/68.20	84.96/92.60	94.12/99.90	98.69/99.92	98.52/100.00	99.80/99.78	97.29/99.74	97.79/99.99
PixArt-α	61.78/81.51	87.75/94.39	92.81/99.80	82.85/98.14	88.30/99.77	99.31/99.79	96.14/99.46	92.94/99.68
Kandinsky 3.0	54.17/66.98	84.66/92.56	87.03/99.61	84.75/98.51	93.88/99.93	99.72/99.81	96.95/99.64	96.62/99.95
DeepFloyd IF	54.75/76.58	86.27/93.60	78.92/99.31	89.00/99.14	63.85/97.62	64.61/93.65	91.28/98.66	70.51/95.46
FLUX.1-schnell	53.67/64.87	75.90/87.52	79.96/99.02	81.89/97.42	70.89/98.13	99.79/99.83	97.14/99.55	81.02/96.97
SD-v1.4	72.52/91.68	74.94/87.24	99.87/100.00	99.06/99.95	99.94/100.00	99.73/99.81	98.89/99.93	99.94/100.00
SD-v1.5	72.11/91.37	75.18/87.18	99.86/100.00	99.11/99.96	99.89/100.00	99.74/99.81	98.85/99.92	99.92/100.00
SDXL	50.20/49.46	74.79/86.43	76.17/98.95	83.82/98.18	88.59/99.80	99.45/99.80	98.47/99.89	94.45/99.75
CogVideoX	68.84/86.20	78.36/88.48	94.33/99.86	87.49/98.63	99.04/100.00	98.73/99.74	90.37/98.31	99.87/100.00
Mean	59.92/75.64	80.84/90.32	88.88/**99.60**	88.92/98.79	88.24/99.41	**96.62**/99.28	96.47/99.51	91.63/98.88

注:加粗字体表示最优结果。

跨生成模型泛化表现得到提升。进一步的训练数据控制消融结果表明,在控制新增训练数据量、类别数量和生成器来源后,实例级语义对齐数据仍能带来性能提升,说明语义对齐约束对AIGI检测任务具有实际意义。

本文的研究还存在一定局限。一方面,不同生成模型在架构、生成质量和语义对齐能力上存在差异,检测结果可能受到生成模型特性和语义因素的

表10　实例级语义对齐训练设置（O+A）下AIGI-Align测试结果（ACC/AP，%）

Table 10　Detection results on AIGI-Align under the instance-level semantic alignment training setting (O+A) (ACC/AP, %)

生成模型	CNN-Spot	UnivFD	RINE	AIDE	Effort	SAFE	DFFreq	ForgeLens
Kolors	62.70/81.98	83.91/89.69	95.19/99.76	96.36/99.56	90.07/99.58	99.49/99.82	98.05/99.76	85.81/97.58
Lumina-Image-2.0	56.36/75.30	86.47/92.54	94.24/99.71	97.50/99.70	96.06/99.89	99.32/99.88	98.72/99.87	96.54/99.75
Hunyuan-DiT	61.04/78.40	87.59/92.82	95.34/99.79	96.93/99.71	93.45/99.89	99.38/99.85	94.97/99.42	94.81/99.79
Stable Cascade	56.84/75.50	89.19/94.28	99.30/99.95	99.16/100.00	99.91/100.00	99.60/99.85	97.21/99.66	99.84/99.99
PixArt-α	70.76/88.27	89.34/94.58	96.38/99.84	94.95/99.42	94.77/99.85	99.38/99.82	97.89/99.74	95.61/99.81
Kandinsky 3.0	62.09/81.04	87.60/93.37	93.96/99.77	95.81/99.52	99.22/99.99	99.45/99.82	97.62/99.73	98.36/99.95
DeepFloyd IF	60.58/82.74	88.25/93.96	87.30/99.54	96.43/99.60	70.99/98.54	68.51/90.55	91.76/98.88	75.44/97.31
FLUX.1-schnell	58.19/75.20	81.04/88.02	89.42/99.26	94.28/99.12	83.23/99.31	99.51/99.89	98.15/99.76	82.68/97.14
SD-v1.4	88.33/97.16	79.74/87.41	99.84/99.95	98.77/99.94	99.96/100.00	99.48/99.85	99.09/99.93	99.93/100.00
SD-v1.5	87.31/96.88	79.41/87.40	99.86/99.95	98.91/99.96	99.96/100.00	99.46/99.87	99.06/99.93	99.93/100.00
SDXL	50.79/51.38	82.59/88.87	88.81/99.45	95.11/98.47	93.67/99.90	99.31/99.81	98.55/99.87	96.33/99.82
CogVideoX	76.33/91.91	82.53/89.09	97.76/99.87	94.75/99.33	99.55/99.99	98.71/99.65	93.89/99.16	99.89/99.99
Mean	65.94/81.31	84.81/91.00	94.78/**99.74**	96.58/99.53	93.40/**99.74**	96.80/99.06	**97.08**/99.64	93.76/99.26

注：加粗字体表示最优结果。

表11　训练数据控制消融的总体性能对比（ACC/AP，%）

Table 11　Overall performance comparison of training data control ablation (ACC/AP, %)

测试集	传统训练设置	类别级提示词训练设置	实例级语义对齐训练设置	ΔC	ΔA
ForenSynths	89.79/96.05	90.15/96.85	89.77/96.30	+0.36/+0.80	−0.38/−0.55
UniversalFakeDetect	87.55/94.59	87.86/94.61	89.48/95.46	+0.31/+0.02	+1.62/+0.85
GenImage	78.04/88.64	83.39/93.61	86.14/94.05	+5.35/+4.97	+2.75/+0.44
AIGI-Align	72.91/84.51	86.44/95.18	90.39/96.16	+13.53/+10.67	+3.95/+0.98

注：表中结果为8种检测器在对应测试集上的平均性能。ΔC表示类别级提示词训练设置与传统训练设置之间的性能差值；ΔA表示在新增训练数据量、类别数量和生成器来源保持一致时，实例级语义对齐训练设置与类别级提示词训练设置之间的性能差值。

共同影响，仍需进一步分析各因素的独立作用；另一方面，AIGI-Align基于14个语义类别构建，类别覆盖范围仍然有限；同时，数据构建过程中的负向提示词虽有助于减少严重偏离真实图像语义参照的低质量样本，但仍可能在一定程度上影响生成图像的原始分布。后续工作将从以下方向展开：第一，从更细粒度的语义维度约束真实图像与伪造图像之间的内容差异，进一步提升数据集质量；第二，从模型结构设计角度引入显式去语义偏见机制，减少检测器对语义内容的依赖，并将实例级语义对齐约束扩展至多模态伪造检测场景；第三，扩展真实图像类别数量和视觉内容覆盖范围。

参考文献（References）

Arkhipkin V, Filatov A, Vasilev V, Maltseva A, Azizov S, Pavlov I, et al. 2023. Kandinsky 3.0 technical report[EB/OL]. [2026-05-20]. https://arxiv.org/abs/2312.03511

Black Forest Labs. 2024. FLUX.1-schnell[EB/OL]. [2026-05-20]. https://huggingface.co/black-forest-labs/FLUX.1-schnell

Cavia B, Horwitz E, Reiss T and Hoshen Y. 2024. Real-time deepfake detection in the real-world[EB/OL]. [2026-05-20]. https://arxiv.org/abs/2406.09398

Chen H, Hu H, Hu J, Huang X, Li W, Lin G, et al. 2023. GenImage: A million-scale benchmark for detecting AI-generated image//

Advances in Neural Information Processing Systems 36. New Orleans, USA: Neural Information Processing Systems Foundation: 77771-77782 [DOI: 10.52202/075280-3398]

Chen J, Yu J, Ge C, Yao L, Xie E, Wang Z, et al. 2024. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis//International Conference on Learning Representations. Vienna, Austria: ICLR: 57611-57640 [DOI: 10.48550/arXiv.2310.00426]

Chen Y, Zhang L and Niu Y. 2025. ForgeLens: Data-efficient forgery focus for generalizable forgery image detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 16270-16280 [DOI: 10.1109/iccv51701.2025.01510]

DeepFloyd Lab. 2023. DeepFloyd IF: A novel state-of-the-art open-source text-to-image model with a high degree of photorealism and language understanding[EB/OL]. [2026-05-20]. https://www.deepfloyd.ai/deepfloyd-if

Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L. 2009. ImageNet: A large-scale hierarchical image database//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Gemma Team, Kamath A, Ferret J, Pathak S, Vieillard N, Merhej R, et al. 2025. Gemma 3 technical report[EB/OL]. [2026-05-20]. https://arxiv.org/abs/2503.19786

Karras T, Aila T, Laine S and Lehtinen J. 2017. Progressive growing of GANs for improved quality, stability, and variation [EB/OL]. [2026-05-20]. https://arxiv.org/abs/1710.10196

Kolors Team. 2024. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis[EB/OL]. [2026-05-20]. https://github.com/Kwai-Kolors/Kolors/blob/master/imgs/Kolors_paper.pdf

Koutlis C and Papadopoulos S. 2024. Leveraging representations from intermediate encoder-blocks for synthetic image detection//European Conference on Computer Vision. Milan, Italy: Springer Nature Switzerland: 394-411 [DOI: 10.1007/978-3-031-73220-1_23]

Li M L, Qian Z X and Zhang X P. 2026. Survey on artificial intelligence-generated image detection. Journal of Image and Graphics, 31(1): 0013-0044 (李美玲, 钱振兴, 张新鹏. 2026. 人工智能生成图像检测技术综述. 中国图象图形学报, 31(1): 0013-0044) [DOI: 10.11834/jig.250053]

Li O, Cai J, Hao Y, Jiang X, Hu Y and Feng F. 2025. Improving synthetic image detection towards generalization: An image transformation perspective//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1. Toronto, Canada: ACM: 2405-2414 [DOI: 10.1145/3690624.3709392]

Li Z, Yan J, He Z, Zeng K, Jiang W, Xiong L, et al. 2026. Is artificial intelligence generated image detection a solved problem? // Advances in Neural Information Processing Systems 38. Sydney, Australia: Neural Information Processing Systems Foundation [DOI: 10.48550/arXiv.2505.12335]

Li Z, Zhang J, Lin Q, Xiong J, Long Y, Deng X, et al. 2024. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained Chinese understanding[EB/OL]. [2026-05-20]. https://arxiv.org/abs/2405.08748

Liu H, Tan Z, Tan C, Wei Y, Wang J and Zhao Y. 2024. Forgery-aware adaptive transformer for generalizable synthetic image detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10770-10780 [DOI: 10.1109/cvpr52733.2024.01024]

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023. Visual instruction tuning// Advances in Neural Information Processing Systems 36. New Orleans, USA: Neural Information Processing Systems Foundation: 34892-34916 [DOI: 10.48550/arXiv.2304.08485]

Ojha U, Li Y and Lee Y J. 2023. Towards universal fake image detectors that generalize across generative models//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Vancouver, Canada: IEEE: 24480-24489 [DOI: 10.1109/cvpr52729.2023.02345]

Pernias P, Rampas D, Richter M L, Pal C and Aubreville M. 2024. Würstchen: An efficient architecture for large-scale text-to-image diffusion models//International Conference on Learning Representations. Vienna, Austria: ICLR: 25097-25109 [DOI: 10.48550/arXiv.2306.00637]

Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, et al. 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis//International Conference on Learning Representations. Vienna, Austria: ICLR: 1862-1874 [DOI: 10.48550/arXiv.2307.01952]

Qin Q, Zhuo L, Xin Y, Du R, Li Z, Fu B, et al. 2025. Lumina-image 2.0: A unified and efficient image generative framework//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 20031-20042 [DOI: 10.48550/arXiv.2503.21758]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//International Conference on Machine Learning. Virtual: PMLR: 8748-8763

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10684-10695 [DOI: 10.1109/CVPR52688.2022.01042]

Wang S Y, Wang O, Zhang R, Owens A and Efros A A. 2020. CNN-generated images are surprisingly easy to spot… for now//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle, USA: IEEE: 8692-8701 [DOI: 10.1109/

cvpr42600.2020.00872

Wang Y B, Hong X P and Huang Z W. 2025. Benchmark dataset and framework for continual AI-generated image detection. Journal of Image and Graphics, 30(11): 3438-3450 (王亚斌, 洪晓鹏, 黄智武. 2025. 面向AI生成图像持续检测基准数据集与框架研究. 中国图象图形学报, 30(11): 3438-3450) [DOI: 10.11834/jig.250167]

Wang Z, Bao J, Zhou W, Wang W, Hu H, Chen H, et al. 2023. DIRE for diffusion-generated image detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22388-22398 [DOI: 10.1109/iccv51070.2023.02051]

Yan J, Li Z, Wang F, He Z and Fu Z. 2026. Dual frequency branch framework with reconstructed sliding windows attention for AI-generated image detection. IEEE Transactions on Information Forensics and Security, 21: 2313-2325 [DOI: 10.1109/TIFS.2026.3664000]

Yan S, Li O, Cai J, Hao Y, Jiang X, Hu Y, et al. 2025. A sanity check for AI-generated image detection//International Conference on Learning Representations. Singapore: ICLR: 70702-70720 [DOI: 10.48550/arXiv.2406.19435]

Yang Y, Li F, Kong S, Diao Y, Gao X, Shi Z, et al. 2026. Layer consistency matters: Elegant latent transition discrepancy for generalizable synthetic image detection[EB/OL]. [2026-05-20]. https://arxiv.org/abs/2603.10598

Yan Z, Wang J, Jin P, Zhang K Y, Liu C, Chen S, et al. 2024. Orthogonal subspace decomposition for generalizable AI-generated image detection[EB/OL]. [2026-05-20]. https://arxiv.org/abs/2411.15633

Yang Z, Teng J, Zheng W, Ding M, Huang S, Xu J, et al. 2025. CogVideoX: Text-to-video diffusion models with an expert transformer//International Conference on Learning Representations. Singapore: ICLR: 83048-83077 [DOI: 10.48550/arXiv.2408.06072]

Zhong N, Xu Y, Li S, Qian Z and Zhang X. 2023. Patchcraft: Exploring texture patch for efficient AI-generated image detection [EB/OL]. [2026-05-20]. https://arxiv.org/abs/2311.12397

作者简介

杨锦旭，男，硕士研究生，主要研究方向为多媒体内容安全。E-mail: yangjx@mail.hfut.edu.cn

陈雁翔，通信作者，女，教授，主要研究方向为多媒体信息处理和多媒体内容安全。E-mail: chenyx@hfut.edu.cn

王志远，男，博士研究生，主要研究方向为信息安全。E-mail: zhiyuanwang@mail.hfut.edu.cn

赵鹏铖，男，讲师，主要研究方向为视听多媒体信息处理。E-mail: stevenzhao1001@gmail.com