最近在人工智能生成内容（AIGC）技术方面的进展大大简化了高度真实的视频内容的创建。这些复杂的生成模型显著降低了制作成本，并在教育和娱乐等各个领域催生了新的应用。然而，这种快速传播也带来了巨大的社会风险。与合成图像相比，生成视频中固有的更高感知真实性和时间一致性加剧了其在错误信息传播、公众信任侵蚀以及社交和专业平台上的信息安全威胁方面的潜力。因此，迫切需要健全可靠的检测方法，能够有效地区分合成视频与真实视频。

然而，健壮的视频造假检测方法的发展在很大程度上依赖于合适的基准。虽然最近提出了一些用于造假检测的 AIGC 数据集，但它们主要针对静态图像 [20, 21, 39]，因此本质上未能捕捉视频特有的挑战，如时间一致性、真实的运动动态以及跨帧的语义一致性。最近的视频基准，如 VBench [14]、EvalCrafter [26] 和 AIGCBench [8]，主要集中于评估生成质量或感知保真度，而非明确针对真实性检测任务。此外，专门用于视频造假的数据集，如 GenVidBench [29] 和 DeMamba [6]，也存在某些限制，包括简单的动画导向内容、狭窄的生成多样性，以及对真实性和检测复杂性关注不足，从而限制了其全面评估高级检测算法的能力。

为了有效克服这些关键限制，我们提出了 AEGIS，这是一种针对 AI 生成视频序列真实性评估的新基准，经过精心设计以挑战和提升当前对高度欺骗性 AI 生成内容的检测能力。AEGIS 的独特之处在于它专门集合了 5,199 个合成视频，这些视频来源于七种最前沿的生成技术，包括像 Stable Video Diffusion [3]、CogVideoX-5B [44] 和 I2VGen-XL [47]，以及由 KLing [19]、Sora [32] 和 Pika [33] 代表的专有商业系统。这样的知名开源方法。这些多样化生成技术的独特整合确保了当前生成范式的无与伦比的真实感、复杂性和代表性，使 AEGIS 成为一个不可或缺的资源，用于严格评估和显著提升模型在面对新兴高度现实伪造威胁时的鲁棒性。

为了严格地进行基准测试并实质性地推进视频伪造检测，AEGIS 引入了几个重要的创新，这些创新清晰地使其区别于现有的基准。首先，它包括通过 GPT-4o [31] 生成的提示精心制作的具有挑战性的子集，明确设计用于深入评估模型对高度复杂和语义细微的伪造的鲁棒性。这些合成视频通过系统策划的真实视频进行补充，后者以显著的视觉复杂性和多样性为特征，创造现实的评估条件。其次，AEGIS 提供了广泛的多模态注释，包括光流、频域分析和丰富的语义描述，以促进严格的评估并支持多样化的下游法医任务。关键的是，我们使用 SOTA 视觉-语言模型（如 Qwen-VL [1] 和 Video-LLaVA [23]）进行的大量实验，涵盖零次学习、提示引导和微调场景，揭示了显著的性能差距，尤其是在具有挑战性的子集中。这些发现强调了 AEGIS 所带来的巨大挑战性以及其揭示当前检测方法内在关键泛化局限性的有效性。因此，AEGIS 成为一个不可或缺的基准，独特地推动了更为鲁棒和广泛可泛化的视频真实性检测模型的发展。

这项工作的主要贡献包括：

- 我们提出了 AEGIS，这是一个用于视频真实性检测的新型大规模基准，其中包括由六种不同的 SOTA 技术生成的 5,199 个合成视频，涵盖了开源和专有模型。AEGIS 在多样性、真实性和语义复杂性方面显著优于现有基准。
- 我们使用 GPT-4o 优化的提示设计具有挑战性的评估子集，以模拟高度真实、语义细致入微的场景。所得的困难测试集有效地揭示了当前检测模型中的泛化差距。
- 我们整理了具有丰富多模态标注的真实视频，包括语义描述、光流、频域特征和时间一致性指标。与先进视觉语言模型的实验揭示了明显的性能限制，强调了 AEGIS 在稳健且可推广的伪造检测中的价值。

合成图像生成由于生成对抗网络（GANs）、扩散模型和流匹配技术的进步而显著提升。早期模型如 StyleGAN 显著增强了面部的真实性，而最近的基于扩散和变压器的方法，包括 Stable Diffusion 和 DALLE-2，则显著扩展了一般用途的图像合成。相应地，几个基准测试出现，以严格评估图像生成质量，主要关注感知保真度、语义对齐和精细检测任务，包括 AIGCIQA2023、AGIQA-20K、PKU-AIGIQA-4K 和 FragFake。尽管有这些进展，这些图像级数据集固有地缺乏对视频特定挑战的考虑，如时间一致性和真实运动模式。我们提出的 AEGIS 明确解决了这些关键的以视频为中心的问题，通过整合时间和多模态分析，显著拓展了静态图像基准的范围。

0.1 视频级 AIGC 基准测试

最近的视频级 AI 生成内容（AIGC）基准如 VBench [14]、T2VSafetyBench [28]、EvalCrafter [26] 和 VIDEOPHY [2] 主要集中在评估视频生成质量，而没有明确解决真实性检测任务。同时，现有专门针对视频伪造检测的数据集，如 DF40 [42]、Deepfake-Eval-2024 [4] 和 ExDDV [12]，主要强调面部操控和自然环境中的深度伪造，限制了他们更广泛的泛化能力。最近的大规模基准如 GenVidBench [29] 和 DeMamba [6]，虽然包含了多样的生成源，通常包含较不具有挑战性的动画风格视频或强调数据集规模而非检测复杂性。相比之下，我们提出的 AEGIS 基准明确优先考虑视频真实性检测，专注于超现实的、语义复杂的 AI 生成视频，故意排除易于识别的动画导向内容。与之前的数据集不同，AEGIS 提供了详细的多模态注释，并包含由 GPT-4o 优化的提示增强的特别构建的挑战子集，明确旨在严格评估检测方法的鲁棒性和泛化能力。

1 数据集构建

为了推进视频真实性检测领域的研究，我们构建了 AEGIS，这是一个由 AI 生成和真实世界视频组成的综合数据集。本节详细介绍了我们系统的构建流程，如 Figure 1 所示，包含三个主要阶段：数据收集、数据过滤和数据拆分。AEGIS 数据集最终结构的概述见 **??**。

1.1 数据收集

1.1.1 **真实视频集合**. 为确保较高的真实感、多样性和视觉复杂性，我们从以下三个来源收集真实视频：（1）Vript 数据集 [43]：我们利用了约 12,000 个视频，这些视频来自 YouTube（水平，长格式）和 TikTok（垂直，短格式）。这种跨平台选择捕捉了与不同格式相关的固有内容和风格偏差，从而保留了真实世界视频的多样性特征。（2）DVF 数据集 [35]：由于视觉复杂性和真实感较高，

[0]Parts of the video frame examples in **??** and Figure 1 are sourced from [41] and [43].

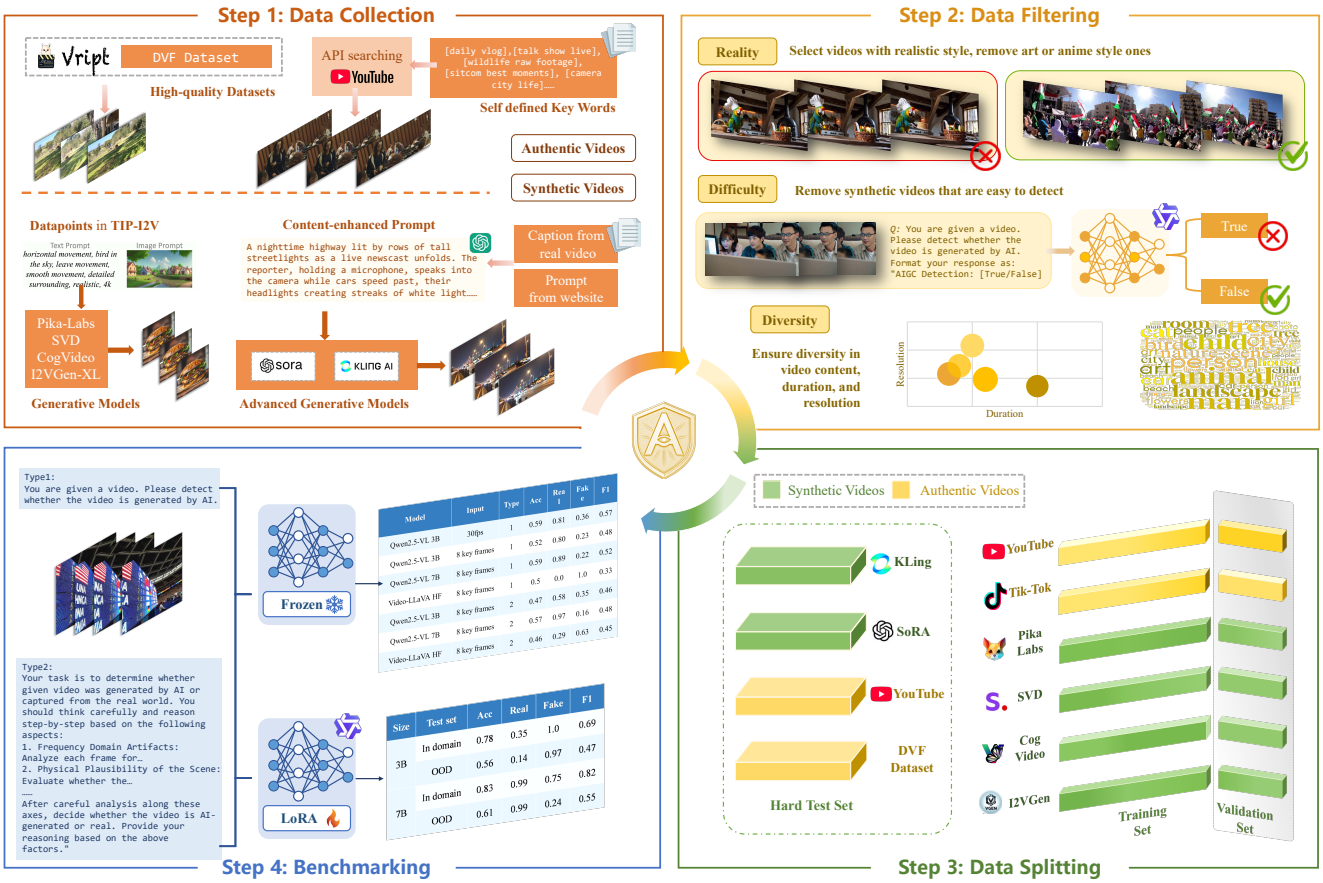

Figure 1: AEGIS 数据集构建流程。步骤 1：数据收集——从不同来源收集真实和合成视频。步骤 2：数据过滤——应用三个关键原则：真实性（去除非真实感内容）、难度（剔除易于检测的伪造品）和多样性（确保内容、分辨率和时长的变化）。步骤 3：数据划分——创建平衡的训练、验证和难度测试集。步骤 4：基准测试——在不同设置下评估视觉语言模型，包括零样本推断、结构化推理提示和低秩适应（LoRA）微调。

该数据集特意选择由人类记录的各种视频片段，这些视频捕捉了细微的细节，并模仿了真实人类生成内容的复杂性。(3) 补充的 YouTube 集合：为了进一步增强我们数据集的真实感和实用性，我们独立收集了较少编辑的视频，包括原始街头采访、野生动物纪录片和真实的博客视频。我们系统性地采用了 30 个预定义的搜索查询，旨在最大化多样性并确保真实性。通过手动检查严格验证每个收集的视频，并通过将剪辑修整到 2.4 到 10 秒的时长范围内，删除音轨，并保持分辨率多样性从 360p 到 4K 的标准化过程，显著提高了数据集的真实感、代表性和复杂性，有效促进了稳健的模型评估。

1.1.2 **合成视频收集**. 我们的合成子集严格整合了公开可用的高级数据集和独立生成的合成内容，采用先进的生成模型（SOTA）以确保多样性和真实复杂性。(1) TIP-I2V 数据集 [41] [1]：该数据集提供了 500,000 个合成视频片段，这些片段是使用五个先进的视频生成模型（SOTA）从 100,000 个提示生成的：Stable Video Diffusion [3] 、CogVideoX-5B [44] 、I2VGen-XL [47] 、Open-Sora [49] 以及 Pika [33] 。(2) 通过 KLing 和 Sora 的专有模型生成：为了在挑战性场景下进一步评估和增强模型的鲁棒性，我们独立地生成了合成视频，利用专有的、先进的生成模型（SOTA），即 KLing [19] 和 Sora [32] 。高质量的文本提示来源于广泛的 HD-VG-130M 数据集 [40] ，并参考官方 KLing [19] 和 Sora [32] 展示中提供的示例进行精心设计。每个提示经过额外的细化过程，利用 GPT-4o [31] ，确保了细节的增强、语义的精确性和真实性。我们系统地生成了一个平衡的 218 个合成视频的集合，仔细控制视觉内容的多样性，分辨率从 360p 到 1080p 不等，持续时间从 5 到 10 秒不等。

为支持真实性检测模型的鲁棒训练和评估，我们提出了一个统一的过滤框架，基于现实、难度和多样性这三个原则，应用于真实和合成子集。该框架确保真实世界的样本呈现高保真、未修改的内容，而合成样本则是照片级真实但不易检测的。它还促进内容多样性，这对于在视觉复杂的场景中实现泛化至关重要。

[1]我们使用了在 https://huggingface.co/datasets/WenhaoWang/TIP-I2V/tree/main/subset_videos_tar 提供的官方整理的原始 TIP-I2V 数据集的子集

（1）真实性。我们确保所有包含的视频——无论是真实还是合成的——都呈现出照片级写实风格，方法是移除艺术风格、低质量或经过大量编辑的内容。对于真实视频，由于数据集如 Vript [43] 和附加的 YouTube Collection 包含来源于网络的素材，许多片段要么是 AI 生成的，要么是经过过度编辑的。为了解决这个问题，我们使用 Qwen2.5-VL [1] 将约 9,000 个片段分类为摄像头拍摄（真实）、大量编辑和 AI 生成的类别，丢弃了约 4,000 个不真实样本。对于 DVF 数据集，我们进行了额外的人工审查，以确保视觉质量和真实性。对于来自 TIP-I2V 数据集 [41] 的合成视频，我们首先从官方提供的 100,000 个中选择高质量、照片级写实的提示。具有抽象、不可信或风格不一致内容的提示被排除。我们随后确认描述的动作在视觉上是可实现的且语义上连贯的，过滤掉物理上不现实或静态的场景。在这个两阶段过滤过程之后，保留了 17,000 个提示。对于每个提示，我们从以上提到的五个候选模型中随机选择一个生成的视频。由于 Open-Sora 的输出视觉质量相对较低 [49]，其生成的视频进一步从我们的子集中移除。此外，我们进行了人工过滤，以确保 KLing [19] 和 Sora [32] 生成的视频的视觉真实性。

(2) 难度。在现实性过滤之后，来自 TIP-I2V 数据集的剩余合成视频通过使用 Qwen2.5-VL 模型 [1] 的零样本分类进行了额外一轮筛选。该步骤旨在进一步排除显示出明显合成生成迹象的样本。具体来说，我们使用 Qwen2.5-VL 将每个视频分类为“AI 生成”（真）或“非 AI 生成”（假）。那些被自信地预测为“AI 生成”的视频被丢弃。相比之下，模型置信度较低而被标记为“非 AI 生成”的视频则被保留。该程序生成了一个经过精心筛选的大约 5000 个合成视频片段的子集，这些片段具有更高的视觉真实性，且不易被轻易检测为生成。这一额外的精炼步骤确保用于我们评估的合成数据在挑战性和感知质量上都很高。

(3) 多样性。我们确保在真实和合成视频中的场景类型、主题和视觉条件上的多样性。真实片段涵盖了室内和室外环境、人类、动物和物体主题，以及城市和乡村场景，时长从 2.4 秒到 10 秒不等，分辨率从 360p 到 4K。合成视频是使用四种 SOTA 生成器从多样的提示生成的，加入了动作、演员、背景及光照条件的变化。场景标签分布（通过词云可视化）和分辨率-时长散点图确认了在语义和视觉维度上的广泛覆盖。

总体而言，我们的综合过滤流程确保了生成的数据集既高质量又具代表性，从而能够对视频真实性检测方法进行稳健且现实的评估。经过所有过滤程序后，最终的 AEGIS 数据集由大约 5,199 个合成视频和 5,271 个真实视频组成，系统地编制以支持在各种具有挑战性的场景中进行可靠的基准测试。

1.2 数据分割

为了在现实部署场景下有效地对模型进行基准测试并严格评估其泛化能力，我们系统地将过滤过的 AEGIS 数据集划分为三个子集：训练集 、验证集 和 难测试集 。训练集和验证集主要包含来自 Vript 数据集 [43] 的过滤后的真实视频和来自 TIP-I2V 数据集 [41] 的高质量合成视频。训练集有助于学习将真实视频与合成视频区分开来的判别特征，而验证集支持超参数调优和初步模型评估。困难测试集专门用于评估模型在更具挑战性条件下的鲁棒性和泛化能力。它包括从 DVF 数据集 [35] 和补充的 YouTube 集合中获取的多样化真实视频，以及由专有模型 KLing [19] 和 Sora [32] 生成的高级合成视频。这些为复杂性和微妙性而挑选出的样本为在涉及复杂伪造和细微视觉细节的现实场景中评估模型的能力提供了关键基准。

1.3 多模态注释

有效区分 AI 生成的视频和真实视频需要在多个维度上捕捉和表示互补的视觉线索，正如最近的研究所强调的那样 [5, 11] 。为支持这一目标，AEGIS 为每个视频提供丰富的多模态注释，涵盖语义真实性描述、运动特征和低级视觉特征。

(1) 语义-真实性描述。为了捕捉高级语义和与真实性相关的线索，我们为每个视频提供两种类型的文本描述：语义描述和真实性推理描述。对于合成视频，我们直接使用来自 TIP-I2V 数据集的原始提示作为语义描述，指定预期的场景、物体和动作。对于没有提示的真实视频，我们使用 CLIP [34] 提取帧级别嵌入，并应用 k -均值聚类（$k = 8$ ）来识别代表性关键帧。然后，我们查询 GPT-4V [30] 以生成总结这些关键帧内容的语义描述。此外，对于真实和合成视频，我们提供真实性推理描述。对于每个视频，我们将其真实标签（真实或 AI 生成）告知 GPT-4V，并提示其仅根据视觉内容解释标签背后的推理。这些解释可能突出显示时间平滑性、光线一致性或视觉伪影的存在，为真实性线索提供人类可解释的见解。

(2) 运动特征。真实的运动往往是时间上平滑且物理上连贯的，而合成视频通常会表现出微妙的伪影或自然动态的违背。为了捕捉这些运动不一致性，我们使用 RAFT 算法提取密集光流场，从而能够对帧与帧之间的运动模式进行细粒度的表征。

(3) 低级视觉特征。低级视觉特征处理微妙但显露的像素级和频域差异，例如边缘锐度、压缩伪影、过于平滑的纹理或重复的图案以及动态范围变化。我们计算每个灰度关键帧的二维快速傅里叶变换（FFT），并应用径向积分操作（RIO）来总结各个方向上的频率能量。

1.4 AEGIS 的独特贡献

提出的 AEGIS 数据集通过明确解决与逼真的 AI 生成视频带来的挑战推进了视频真实性检测，这些视频与真实的人类创作内容非常相似。与通常包含风格化动画或易于检测的场景的现有基准不同，AEGIS 仅专注于视觉上细微且情境丰富的视频，经过精心策划以反映现实检测任务的复杂性。

此外，AEGIS 利用 GPT-4o 精炼的提示和最先进的专有生成模型——例如 KLing 和 Sora——来合成高度逼真且具有欺骗性的伪造品。这些伪造品与经过精心挑选的真实样本相结合，形成了“困难测试集”。这是一个严格的基准，旨在评估模型在具有挑战性的真实世界条件下的鲁棒性和泛化能力。

此外，AEGIS 提供了可直接使用的多模态视觉提示，以支持合成视频检测和可解释推理中的下游任务，帮助深入了解模型行为和失败情况。在本节中，我们设计了评估策略，以基准测试我们 AEGIS 数据集上 SOTA 视觉-语言模型的真实性检测性能。我们评估两个子集。(i) 域内测试集：从验证集中随机抽样出的子集，与训练数据共享相同的分布；(ii) 难测试集（详见 Sec. 1.2 ）：一个明确设计用于评估模型在具有挑战性的合成视频上的鲁棒性和广泛性的跨域数据集。

基线模型。我们在 AEGIS 数据集上评估了两个最新的视觉语言模型：Qwen2.5-VL [1] 和 Video-LLaVA [23] 。Qwen2.5-VL 是一个强大的通用模型，具有稳健的多模态理解能力，并在以视频为中心的任务上表现出竞争力；我们考虑了它的 3B 和 7B 两个版本。Video-LLaVA 是一个具有代表性的自回归 transformer 模型，旨在将图像和视频的理解统一在一个框架中。

为了系统地检查模型在不同任务条件下的表现，我们实施了三种评估策略：(i) 零样本推理，(ii) 结构化推理提示，以及 (iii) 低秩自适应 (LoRA) [13] 微调。为了在推理过程中利用提取的多模态线索，预提取的关键帧被以模型接口支持的 <image> 令牌格式输入到视觉-语言模型中。

(1) 零样本推理。我们使用一个最小提示来从模型中获取二元决策："你是 AI 生成内容 (AIGC) 检测的专家。给定一个视频，判断它是真实的还是 AI 生成的。"此设置评估模型在仅有任务描述的情况下执行真实性检测的默认能力。

(2) 结构化推理提示。我们构建了一个多步骤提示，引导模型通过几个视觉维度进行详细的推理过程，包括频率伪影、光照一致性、压缩噪声和物理合理性。我们增强推理的提示模板在补充链接中提供。

(3) LoRA 微调。为了探索特定任务的适应性，我们在训练集上使用 LoRA [13] 对 Qwen2.5-VL [1] 进行微调（学习率 $1e^{-4}$ ，秩为 8，3 个 epoch）。我们利用广泛使用的框架 llama-factory [48] 进行有效训练。此设置用于量化轻量级监督的潜在收益，并评估模型在超出训练分布之外的泛化能力。

对于每个设置，我们报告四个指标：Acc_{all} ：总体分类准确率（从 0 到 1）。Acc_{real} ：真实视频的准确率。Acc_{ai} ：合成视频的准确率。Macro- $F1$ ：跨两个类别的无权平均 F1 得分。

1.5 基准测试结果

在 AEGIS 数据集上进行的实验有两个目的：（i）展示当前的 VLMs 在 AEGIS 上的视频真实性评估存在困难，以及（ii）测试在 AEGIS 上进行额外训练是否能提升它们的性能。

AEGIS 揭示了 VLMs 零样本检测中的缺陷。如 Table 1a 所示，像 Qwen2.5-VL [1] 这样的 SOTA 模型在 AEGIS Hard 测试集上的零样本设置下实现了低合成视频检测准确率（Acc_{ai} 从 0.22 到 0.23）。这突显出现有模型能力与 AEGIS 样本的高视觉保真度之间的巨大差距。此外，基于提示的推理几乎没有改善。如 Table 1b 所示，当将直接的文本提示应用于 Qwen2.5-VL 7B 时，准确率进一步从 0.22 下降到 0.16。该意外表现表明传统的提示策略未能捕捉到 AEGIS 中高质量伪造的细微视觉和语义线索。

Table 1: 困难测试集上的检测准确性

(a) 零样本推理

Model	Acc_{all}	Acc_{real}	Acc_{ai}	Macro F1
Qwen2.5-VL 3B	0.52	0.80	0.23	0.48
Qwen2.5-VL 7B	0.59	0.89	0.22	0.52
Video-LLaVA-HF 7B	0.5	0.0	1.0	0.33

(b) 结构化推理提示

Model	Acc_{all}	Acc_{real}	Acc_{ai}	Macro F1
Qwen2.5-VL 3B	0.47	0.58	0.35	0.46
Qwen2.5-VL 7B	0.57	0.97	0.16	0.48
Video-LLaVA-HF 7B	0.46	0.29	0.63	0.45

在 AEGIS 上的训练提升了真实性检测。通过 LoRA 微调，在域内测试集上取得了显著的性能提升，例如，Qwen2.5-VL 7B 的宏 F1 从 0.43 增加到 0.82。然而，正如 Table 2 所示，在困难测试集上的改进仍然有限，宏 F1 仅略微上升从 0.52 到 0.55。这突显了在实现对真实且高保真伪造品的泛化方面持续的挑战。

在领域数据和困难测试集上的表现之间的鲜明对比，强调了由 AEGIS 独特提出的重要泛化挑战。尽管进行了有针对性的微调，但当前模型在面对困难测试集中故意包含的微妙而真实的视频时，仍然难以有效泛化所学到的真实性线索。

Table 2: 微调后在两个测试集上的检测准确率

M	T	Eval	Acc_{all}	Acc_{real}	Acc_{ai}	Macro- $F1$
3B	ID	ZS	0.65	0.87	0.55	0.65
7B	ID	ZS	0.45	0.50	0.20	0.43
3B	ID	LoRA	0.78	0.35	1.00	0.69 (+0.04)
7B	ID	LoRA	0.83	0.99	0.75	0.82 (+0.41)
3B	HT	ZS	0.52	0.80	0.23	0.48
7B	HT	ZS	0.59	0.89	0.22	0.52
3B	HT	LoRA	0.56	0.14	0.97	0.47 (-0.01)
7B	HT	LoRA	0.61	0.99	0.24	0.55 (+0.03)

M : Model size (3B = Qwen2.5-VL-3B, 7B = Qwen2.5-VL-7B); T : Test set (ID = In-domain, HT = Hard test set); Eval : Evaluation Type (ZS = Zero-shot, LoRA = After LoRA fine-tuning).

这一限制强烈表明需要进行未来研究，以探索更先进和鲁棒的微调或领域适应策略，这些策略专门针对提高模型对 AEGIS 级伪造复杂性的泛化能力。所有这些见解共同强调了 AEGIS 所呈现的独特价值和重大挑战，明确地将其建立为推进强大、现实且高度可泛化的 AI 生成视频检测研究的关键资源。

在这项工作中，我们提出了 AEGIS ，这是一种新颖的大规模视频真实性基准，专门针对复杂的 AI 生成视频。与现有数据集不同，AEGIS 优先考虑超现实的场景，排除简单或易于检测的样本，显著增强了检测的复杂性和真实性。通过严格的数据筛选、战略性的数据集划分，以及来自先进生成模型（例如，Sora，KLing）的欺骗性样本的加入，AEGIS 提高了合成视频检测的标准。实验评估显示，即使是当前最先进的视觉语言模型在零样本设置中，特别是在困难测试集上，也难以泛化。此外，AEGIS 提供了多维度的视觉线索和丰富的多模态注释。这些不仅支持下游检测任务，还促进可解释推理，使对模型失败和决策边界的分析更加细致。我们相信，AEGIS 通过提供一个具有挑战性、多样化和可解释性导向的基准，为合成视频检测研究建立了基础性的转变，这对于开发强大和值得信赖的多模态 AI 系统至关重要。

2

致谢 本研究得到了新加坡国家研究基金会在其 AI 新加坡计划下的支持（AISG 奖项编号：AISG3-RP-2024-033）。

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025).

[2] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. 2024. Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520 (2024).

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023).

[4] Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, et al. 2025. Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. arXiv preprint arXiv:2503.02857 (2025).

[5] Chirui Chang, Zhengzhe Liu, Xiaoyang Lyu, and Xiaojuan Qi. 2024. What Matters in Detecting AI-Generated Videos like Sora? arXiv preprint arXiv:2406.19568 (2024).

[6] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. 2024. Demamba: Ai-generated video detection on million-scale genvideo benchmark. arXiv preprint arXiv:2405.19707 (2024).

[7] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[8] Fanda Fan, Chunjie Luo, Wanling Gao, and Jianfeng Zhan. 2023. Aigcbench: Comprehensive evaluation of image-to-video content generated by ai. BenchCouncil Transactions on Benchmarks, Standards and Evaluations (2023).

[9] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. 2024. Discrete flow matching. In NeurIPS .

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets.

[11] Xinan He, Yue Zhou, Bing Fan, Bin Li, Guopu Zhu, and Feng Ding. 2025. VLForgery Face Triad: Detection, Localization and Attribution via Multimodal Large Language Models. arXiv preprint arXiv:2503.06142 (2025).

[12] Vlad Hondru, Eduard Hogea, Darian Onchis, and Radu Tudor Ionescu. 2025. ExDDV: A New Dataset for Explainable Deepfake Detection in Video. arXiv preprint arXiv:2503.14421 (2025).

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. ICLR (2022).

[14] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In CVPR .

[15] Yoori Hwang, Ji Youn Ryu, and Se-Hoon Jeong. 2021. Effects of disinformation using deepfake: The protective effect of media literacy education. Cyberpsychology, Behavior, and Social Networking (2021).

[16] Xinyi Jin, Zhuoyue Zhang, Bowen Gao, Shuqing Gao, Wenbo Zhou, Nenghai Yu, and Guoyan Wang. 2025. Assessing the perceived credibility of deepfakes: The impact of system-generated cues and video characteristics. new media & society (2025).

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In CVPR .

[18] Leon Klein, Andreas Krämer, and Frank Noé. 2023. Equivariant flow matching. In NeurIPS .

[19] Kuaishou. 2024. Kling: AI Video Generation Model. https://https://klingai.kuaishou.com//.

[20] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, et al. 2024. Aigiqa-20k: A large database for ai-generated image quality assessment. In CVPR .

[21] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. 2023. Agiqa-3k: An open database for ai-generated image quality assessment. IEEE Transactions on Circuits and Systems for Video Technology (2023).

[22] Xiaoming Li, Xinyu Hou, and Chen Change Loy. 2024. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In CVPR .

[23] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023).

[24] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. 2024. Detecting multimedia generated by large ai models: A survey. arXiv preprint arXiv:2402.00045 (2024).

[25] Xiao Liu, Xinhao Xiang, Zizhong Li, Yongheng Wang, Zhuoheng Li, Zhuosheng Liu, Weidi Zhang, Weiqi Ye, and Jiawei Zhang. 2024. A survey of ai-generated video evaluation. arXiv preprint arXiv:2410.19884 (2024).

[26] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024. Evalcrafter: Benchmarking and evaluating large video generation models. In CVPR .

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

[28] Yibo Miao, Yifan Zhu, Lijia Yu, Jun Zhu, Xiao-Shan Gao, and Yinpeng Dong. 2024. T2vsafetybench: Evaluating the safety of text-to-video generative models. In NeurIPS .

[29] Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. 2025. GenVidBench: A Challenging Benchmark for Detecting AI-Generated Video. arXiv preprint arXiv:2501.11340 (2025).

[30] OpenAI. 2023. GPT-4V. https://openai.com/index/gpt-4v-system-card/.

[31] OpenAI. 2024. GPT-4o. https://platform.openai.com/docs/models/gpt-4o.

[32] OpenAI. 2024. Sora: AI Video Generation Model. https://openai.com/sora.

[33] Pika Labs. 2024. Pika: AI Video Generation Platform. https://www.pika.art/.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning .

[35] Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. 2024. On learning multi-modal forgery representation for diffusion generated video detection. arXiv preprint arXiv:2410.23623 (2024).

[36] Zhen Sun, Ziyi Zhang, Zeren Luo, Zeyang Sha, Tianshuo Cong, Zheng Li, Shiwen Cui, Weiqiang Wang, Jiaheng Wei, Xinlei He, Qi Li, and Qian Wang. 2025. FragFake: A Dataset for Fine-Grained Detection of Edited Images with Vision Language Models. arXiv preprint arXiv:2505.15644 (2025).

[37] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16 . Springer, 402–419.

[38] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social media+ society (2020).

[39] Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. 2023. Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. In CAAI ICAI .

[40] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. 2023. VideoFactory: Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. arXiv preprint arXiv:2305.10874 (2023).

[41] Wenhao Wang and Yi Yang. 2024. TIP-I2V: A Million-Scale Real Text and Image Prompt Dataset for Image-to-Video Generation. arXiv preprint arXiv:2411.04709 (2024).

[42] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. 2024. Df40: Toward next-generation deepfake detection. arXiv preprint arXiv:2406.13495 (2024).

[43] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. 2024. Vript: A video is worth thousands of words. Advances in Neural Information Processing Systems 37 (2024), 57240–57261.

[44] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. arXiv preprint arXiv:2408.06072 (2024).

[45] Jiquan Yuan, Fanyi Yang, Jihe Li, Xinyan Cao, Jinming Che, Jinlong Lin, and Xixin Cao. 2024. PKU-AIGIQA-4K: A Perceptual Quality Assessment Database for Both Text-to-Image and Image-to-Image AI-Generated Images. arXiv preprint arXiv:2404.18409 (2024).
[46] Ruihan Zhang, Borou Yu, Jiajian Min, Yetong Xin, Zheng Wei, Juncheng Nemo Shi, Mingzhen Huang, Xianghao Kong, Nix Liu Xin, Shanshan Jiang, et al. 2025. Generative AI for Film Creation: A Survey of Recent Advances. arXiv preprint arXiv:2504.08296 (2025).
[47] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. 2023. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 (2023).
[48] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. arXiv preprint arXiv:2403.13372 (2024).
[49] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. 2024. Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024).
[50] Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. 2023. Shifted diffusion for text-to-image generation. In CVPR .