VSG-Safe: Spotting NSFW Video through Cross-Frame Evidence

Yuyang Zhang[1] * *Xudong Jiang*[1] * *Yuxuan Song*[1] *Yuxiang Sun*[1] *Yihao Huang*[2]
Run Wang[1] † *Shundi Xiao* [1] *Lina Wang* [1]

[1] *Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University,* [2] *National University of Singapore*

Abstract

Recent advances in text-to-video (T2V) models enable high-fidelity videos that closely follow textual prompts. However, this expands practical applications while amplifying serious security and societal concerns from the automated synthesis of visual content that may be inappropriate in certain usage contexts, such as public or workplace settings, including sexual or violent content (*e.g.*, the Grok can generate sexual videos in the "Spicy" mode). We observe that such visual content is often distributed across frames, embedded in visual entities, their attributes, and inter-entity relations. In contrast, existing moderation pipelines primarily treat video content as either individual frames or raw frame sequences, overlooking the fact that critical semantics can manifest through the combination of specific frames. This gap prevents them from reasoning across frames, confining detection to low-level visual cues, such as gore or explicit conflict, and causing frequent failures when cross-frame inference is required, including illegal activities or threats. To address these limitations, we propose leveraging scene graphs as the core intermediate semantic representation. Scene graphs naturally encode entities, their attributes, and inter-entity relationships, while also supporting reasoning over cross-frame content. Grounded on this insight, we further propose VSG-Safe, a novel scene-graph-driven framework for T2V content moderation. Concretely, our approach first extracts cross-frame content from videos to build scene graphs. With these graphs, we leverage a graph-oriented model to jointly capture entities, attributes, and inter-entity relations, enabling effective detection. To evaluate its effectiveness, we conduct extensive experiments on both SOTA benchmarks and our self-constructed video datasets. VSG-Safe attains an average F1-score of 97.62%, outperforming seven baselines by 42.32% on average.

Disclaimer: This paper contains visual content that might be offensive to some readers, such as sexual and violent content. Although we censor and mask Not-Safe-for-Work (NSFW) imagery, reader discretion is advised.

*Equal contribution.
†Corresponding author. Email to wangrun@whu.edu.cn

1 Introduction

With the rapid advancement of visual synthesis technology, the latest Text-to-Video (T2V) models, such as Sora [32] and Hailuo [18], can generate high-resolution, chronologically coherent, and vivid videos based on natural language prompts or image examples, eliminating the need for complex character modeling, visual effects, and other techniques [22, 49]. These technologies quickly produce vast amounts of high-quality video footage, advancing automated production in industries such as film production [47], education [52], and entertainment [56]. We are witnessing T2V technology revolutionize video content creation and create a wave of AI-generated content culture on social media [22].

However, with the growing user base and expanding application domains, the potential for misuse of T2V models has become a significant concern [23]. Prior studies [28, 31] indicate that malicious users and attackers exploit these models to generate and distribute videos with Not Safe For Work (NSFW) [23] content, including pornography and violence which are widely considered inappropriate for public platforms [40], *e.g.*, TikTok and YouTube[1]. Given that T2V models primarily use text as input but also support key-frames as input, they can produce NSFW videos when fed inappropriate prompts or images [21]. For instance, Grok provides a "Spicy" mode that actively directs it to generate sexual videos[2]. This capability has raised widespread security concerns within society regarding the deployment of T2V technology [33].

To mitigate the misuse of T2V models, artificial intelligence generated content (AIGC) service providers have introduced various safeguard mechanisms [23]. Broadly, these efforts can be categorized into three directions: input filtering, model enhancement, and output detection. Input filtering aims to block prompts before generation, using keyword blacklists or semantic embeddings to identify the policy-violating in-

[1]Dozens of YouTube Channels Are Showing AI-Generated Cartoon Gore and Fetish Content. (May 2, 2025)
[2]Grok Imagine, xAI's new AI image and video generator, lets you make NSFW content. (Aug 4, 2025)

puts [3, 39, 54]. Model enhancement strategies attempt to reduce the generation of NSFW content during the generation process by either removing NSFW concepts from the model [53] or intervening in prompt embeddings or hidden representations to steer the generation process [7, 21].

Although these defenses constrain the generation NSFW content to some extent, recent research reveals that input filtering and model enhancement are tailored solutions [25, 31, 62]. Input filters are ad hoc and lack generality, which requires separate designs for each T2V model, where direct transfer often fails due to misalignment between the filter and the latent space of the model, ultimately rendering the filtering ineffective. Model enhancement is even more costly, requiring significant computational overhead and performance tradeoffs to reduce NSFW content generation. In contrast, output detection methods analyze the generated content itself, providing a last line of defense by filtering results [6, 19]. By inspecting the generated video directly instead of relying on inputs or model assumptions, once trained, such methods can be trained once and then deployed as plug-ins to arbitrary T2V systems, without requiring further fine-tuning. Consequently, output detection has emerged as a more robust and practical defensive strategy [36, 40].

In practice, for generated videos, content subject to moderation is not confined to a single frame [10, 40]. Instead, it may manifest through sparse patterns, such as threats or other entity-level interactions, distributed across multiple frames, thereby requiring the aggregation of cross-frame information for reliable detection [28, 42]. Grounded on shot types in film-art scholarship [34], we categorize videos into three types: ❶ camera shot, where all semantics can be captured within a single frame, similar to static images; ❷ continuous take, where relevant semantics unfolds continuously and can be detected by analyzing the raw frame sequence; ❸ discontinuous take, where critical semantics manifest across nonconsecutive frames, often due to the insertion of unrelated frames, requiring cross-frame reasoning for detection.

Existing output detection methods can be broadly characterized as either single-frame or multi-frame. To capture NSFW content in videos, a straightforward idea is to decompose a video into individual images [38, 40], but this discards inter-frame dependencies and limits detection to static attributes or explicit visual conflicts within camera shots [33]. In contrast, multi-frame detectors treat videos as simple sequences of raw frames and compress them into a single global representation. Although this design introduces multi-frame features, prior works show that this design loses cross-frame temporal relations and fine-grained entity relationships [43, 50, 60], restricting detection to coarse-grained content classification rather than nuanced semantic reasoning, and ultimately leading to significant underestimation of videos' harmfulness [9, 61].

To address these challenges, for the first time, we investigate video content from multiple frames, proposing a scene-graph-driven framework for defining and identifying NSFW video content through video scene graphs, which explicitly models entity attributes and inter-entity relationships [60]. This representation overcomes the limitations of existing approaches in modeling and capturing semantics in videos. Building on this representation, we introduce a novel output detection technique, Visual Scene Graph for NSFW Detection (VSG-Safe). Specifically, to capture cross-frame content in videos, VSG-Safe introduces a video scene graph generation framework, which jointly modeling entities, attributes, and relationships across frames, addressing the limitations of existing approaches, which fail to predict entity attributes and cross-frame relationships. Subsequently, to effectively capture and identify the harmfulness of the topology formed by directed relationships and nodes with attributes in scene graphs, we design a dual-channel graph classifier, Dual-GNN, which builds on the extracted graphs to detect target content in videos through two complementary channels, *i.e.*, entities and relationships. Leveraging these scene graphs together with cross-frame semantic modeling, VSG-Safe can effectively detect NSFW-related patterns that are implicit or distributed across multiple frames, thereby overcoming the inherent limitations of single-frame detection. Our key contributions are:

- We reveal that critical semantics in videos often manifests through cross-frame scenarios and analyze the limitations of existing NSFW content detection methods in handling both continuous and discontinuous frame scenarios.
- We propose a novel framework that detects NSFW content by reasoning over cross-frame content, instantiated with scene graphs extraction to capture entities and their inter-entity interactions across frames.
- We categorize videos into three shot types and construct a benchmark dataset covering seven content categories commonly considered in NSFW video analysis, with samples spanning all three scene graph types, to facilitate comprehensive evaluation of cross-frame video detection methods.
- Our experiments show our method achieves an F1-score higher than 97.62%, improving by 42.72% over 7 competitive baselines on average. In particular, VSG-Safe achieves an F1-score of 96.84% on cross-frame scenarios, with a 54.66% improvement, while in NSFW relation detection it reaches 97.06%, improving by 49.47% over baselines.

2 Related Work

2.1 Text to Video

In recent years, powered by training on large-scale vision-language datasets, generative models have shown impressive abilities in capturing semantic alignment and creating high-quality visual content [29, 37]. T2V models have made significant strides [13], allowing users to produce high-resolution, semantically consistent, and realistic videos from various inputs like natural language, images, or video clips [17, 47, 48].

As these models advance in capability, they also introduce heightened security risks [23, 28]. Many web-sourced training datasets lack thorough curation and contain NSFW content [45]. As a result, generative models might unintentionally learn patterns related to graphic, violent, or pornographic material, leading to the potential creation of content that is inappropriate in certain usage contexts [35]. As generative models become more powerful, NSFW content has shifted from short videos illustrating isolated NSFW objects to long videos with coherent narratives of stories [28, 33]. Drawing on definitions from the film industry, we further observe that cross-frame content can be categorized by cinematographic techniques into camera shots, continuous takes, and discontinuous takes. Among these, camera shots align with existing single-frame-based detection settings, whereas current methods exhibit severe shortcomings when dealing with cross-frame cases, namely continuous takes and discontinuous takes. The precise definitions are introduced in Section 3.1.

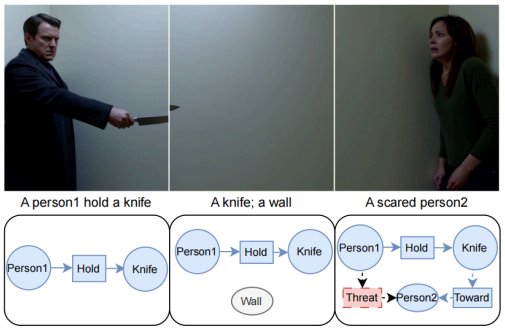


Figure 1: **An example of NSFW video content.** NSFW video content is spread across multiple frames, requiring context-aware analysis. In this example, camera movement presents a scene with three consecutive clips, *i.e.*, a person$_1$, a knife, and a person$_2$, showing only the "person$_1$-knife" relations. The "person$_1$-person$_2$" and "knife-person$_2$" relations are absent from any single frame and must be inferred across frames.

2.2 NSFW Defenses in Video Generation

With growing attention to the NSFW issues in AI-generated video content, recent years have seen a surge of research focused on defending against NSFW content generation in T2V services. Among existing defense strategies, three widely adopted categories are input filters, model enhancement, and output detection [26]. **Input filters** attempt to prevent undesired content generation under content moderation policies by rejecting inappropriate prompts [25, 54]. However, this approach heavily relies on accurate semantic interpretation of the input and precise alignment with the target model's semantics, making the defense lacking generality. **Model enhancement** modifies the generative models themselves (*e.g.*, via fine-tuning or unlearning) to suppress NSFW outputs, though often at the cost of utility and with limited adaptability [53]. **Output detection** targets the generated videos directly [16, 25], including methods that analyze the final output to identify and filter NSFW content. Such methods can be divided into those that operate on single frames (images), such as Q16 [40] and USD [57], which process each frame independently and therefore do not model cross-frame information, making them less effective for video-level safety analysis, and those that operate on multiple frames (videos), such as the ViViT-based method [42]. Notably, output detection is particularly versatile, as it can be applied not only to T2V services but also to other scenarios requiring content moderation. Therefore, output detection methods are generally regarded as more robust and applicable. Hence, in this work, we focus primarily on output detection methods.

2.3 Panoptic Scene Graph Generation

Existing panoptic scene graph generation studies can be broadly categorized into two groups, *i.e.*, static and dynamic scene graph generation(SGG). The first focuses on image- or frame-oriented scene graph generation, emphasizing static relational modeling while overlooking cross-frame entity consistency and relationship dynamics [59]. The second line explores the construction of dynamic video scene graphs, aiming to capture temporal semantics by aligning entities across frames (*e.g.*, via object tracklets) for modeling inter-frame probabilistic graphs. A representative dynamic SGG technique is PVSG [51], which incorporates a conditional modeling framework and makes notable progress in capturing the uncertainty of single-frame relations. However, these methods primarily focus on the accuracy of limited relations, such as spatial positioning, while struggling to capture entity attributes and complex relationships, particularly non-contact interactions between entities, such as weapon threats, physical intimidation, and sexual harassment, limiting their applicability to analysis on video with long takes. Moreover, their usability in T2V systems remain underexplored.

To address the above limitations, in this work, we propose a novel cross-frame scene graph extracting technique that explicitly enhances the understanding of complex inter-entities relationships and the recognition of entity attributes in videos.

3 NSFW Content in Videos

The concept of unsafe content does not stem from a single, universally agreed-upon definition, but instead varies across social, contextual, and identity-related factors (*e.g.*, gender and race) [36]. For instance, prior work has discussed considerations such as whether data may be offensive, threatening, or anxiety-inducing when viewed directly [14]. However, such descriptions are not intended as formal or exhaustive definitions, and what is considered inappropriate remains inherently subjective. Given the absence of a universally accepted and

task-agnostic definition of unsafe content, we avoid adopting any single formulation. Instead, drawing on multiple prior works that study NSFW content [16, 23, 36, 40], we focus on NSFW content as the core scope of this paper. In particular, NSFW is operationalized as a set of moderation-relevant visual content categories commonly studied as detection targets in prior technical research and widely considered inappropriate in public platform moderation contexts, and is treated purely as a task-oriented classification scope rather than a normative definition or claim about inherent harm or unsafety. Specifically, we concentrate on seven categories, *i.e.*, sexually explicit, disturbing attributes, physical violence, armed threats, illegal activities, harassment, and self-harm [28, 40].

Compared to NSFW detection in images, detecting such content in videos is more challenging because NSFW content, especially interactions, often unfolds across a temporal sequence of multiple specific frames, whereas NSFW elements in images are typically localized within a single frame [15]. As illustrated in Figure 1, such content may involve multi-frame interactions that are only semantically coherent when considered together. For instance, a violent scenario involving the person_1 threatening the person_2 may span several frames: one frame shows the person_1 holding a sharp knife, another captures the knife in mid-air, and a third presents a close-up of person_2 in a defensive posture and facing the knife. Individually, these frames may appear ambiguous or benign due to the unknown intentions of the individuals or objects involved. However, when considered together, they collectively convey NSFW semantics.

3.1 Scene Graph of Video

Prior studies reveal that the core challenge in video understanding and reasoning is not appearance recognition but semantic reasoning over temporally evolving entities, their attributes and inter-relations, *i.e.*, identifying who is involved, what attributes they exhibit, and how they interact [9, 60]. Yet, existing NSFW detection methods largely rely on pixel-level features [38, 40], which capture appearance but do not reliably model the entity-centric semantics required for distinguishing different types of behaviors and interactions [12, 55].

This limitation aligns with prior observations in visual reasoning, suggesting that structured representations of entities, attributes, and their inter-relationships tend to provide a more reliable basis for video reasoning than raw appearance features [60]. Accordingly, video scene graphs, which explicitly model entities, attributes, relations, and their temporal evolution, are widely recognized to encapsulate information that is essential to visual understanding and reasoning [9, 60]. Hence, we use video scene graphs as the core semantic representation, retaining the semantics necessary for NSFW detection.

Grounded in film-art scholarship [34], the depiction of a scene through a video clip can be decomposed into two primary dimensions: *shot composition* and *shot movement*. In this work, we further identify this depiction into three representative shot modes of SSG generation in video scene graph generation: ❶ **Camera shot**: Similar to an image, all relevant semantic relations can be fully observed within any single frame of this full shot. ❷ **Continuous take**: A continuous camera movement that semantic interpretation depends on a sequence of temporally adjacent frames, where individual frames may contain no NSFW content, can still be detected through straightforward frame-sequence analysis without cross-frame reasoning. ❸ **Discontinuous take**: A set of interrupted takes, where semantics are distributed across non-adjacent scenes, requires long-range cross-frame reasoning to be accurately inferred. To clarify the distinctions of three introduced representative modes, we provide an illustrative visualization in Figure 2. Our categorization, while not a complete artistic deconstruction, is indispensable for understanding and enabling reliable detection of NSFW content. It further motivates the need for representations that capture cross-frame content, as opposed to detecting NSFW content from isolated frames or naive raw frame sequences.

To meet this need, our scene graph design assumes temporal continuity in the video's semantics where entities and relationships may persist, evolve, or accumulate across frames rather than being re-inferred from scratch at every timestep. The incremental inheritance-update-expansion mechanism allows our scene graph to integrate both the current frame's information and contextual semantics from previous frames, forming a graph representation with temporal depth.

Specially, the scene in the video is decomposed into a sequence of scene graphs $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_t, \ldots, \mathcal{G}_T\}$, where $\mathcal{G}_t$ denotes the fused scene graph from frame 1 up to frame t. Importantly, $\mathcal{G}_t$ is not a static snapshot of frame t, but rather the result of progressively inheriting, updating, and expanding upon the previous scene graph $\mathcal{G}_{t-1}$. As new frames are processed, existing entities and relationships in the graph may be updated (*e.g.*, changes in attributes or relations), and new entities or relationships may be added, thereby enriching the graph with cumulative semantic context. Scene graph $\mathcal{G}_t$ is defined as $\mathcal{G}_t = \{\{(v_i, c_i, a_i)\}, \{(v_i, r_{i,j}, v_j)\}\}$, where v_i represents the unique ID of i-th entity node, c_i is its category , a_i denotes its set of attributes, and $r_{i,j}$ captures the directed visual relationship from entity v_i to v_j.

3.2 NSFW Video Scene Graph

Based on our definition of video scene graphs, we conceptually decompose NSFW content in videos into two complementary aspects, *i.e.*, entity-level cues and relationship-level interactions, that together provide a comprehensive representation: ❶ **Entities**, referring to individual nodes belonging to whose categories or attributes (*e.g.*, blood, nudity) are indicative of NSFW-related content; ❷ **Explicit relationships**, interactions that are directly observable (*e.g.*, physical assault, sexual activity); and ❸ **Implicit relationships**, more complex

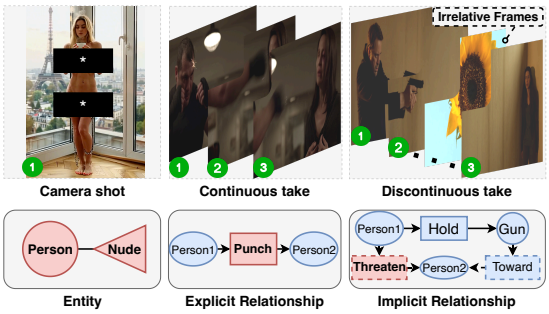


Figure 2: **Shot modes and scene graphs.** The top row shows examples of the three shot modes. Numbers indicate clips composed of consecutive frames within the same scene. In the discontinuous case, frames of an irrelevant shot are inserted to obstruct the detection of continuous semantics. The bottom row presents the corresponding scene graphs extracted. NSFW elements are highlighted in red, while unrelated objects (*e.g.*, sunflowers) are omitted for clarity.

or indirect interactions that require contextual reasoning over multiple entities to be recognized, (*e.g.*, armed threatening).

The entity focuses on the inherent visual attributes of entities that may convey NSFW content, including their categories (*e.g.*, drugs, cigarettes) as well as specific features (*e.g.*, nude). Identifying the category is relatively straightforward and can often be handled by a ban list. However, assessing whether a particular attribute constitutes NSFW content often requires considering the category jointly. For instance, nudity involving animals or non-human entities is generally treated differently from nudity involving human subjects, which often carries distinct visual patterns and regulatory implications. We represent its structure using the triple (v_e, c_e, a_e), where v_e denotes the entity node, c_e indicates its category, and a_e represents the set of attributes. As illustrated by the left video in Figure 2, an entity can correspond to a specific class together with associated attributes, jointly indicating NSFW content.

Unlike the first category, which directly relies on the intrinsic visual features of individual entities, the latter two focus on interactions between entities. The relationship emphasizes the relational dynamics manifested through posture, motion, and other behavioral cues. Such dynamics often convey either explicit or implicit indications of aggressive or potentially harmful behaviors, including acts of violence, illegal activity, or self-harm. From the perspective of behavioral manifestation, these interactions can be categorized into two types: explicit relationships and implicit relationships. Specially, explicit behaviors can be represented using the relational triple $(v_{e_1}, r_{e_1,e_2}, v_{e_2})$, abbreviated as the (S,V,O) structure. As the middle video illustrated in Figure 2, this scene captures directly observable conflicts or inappropriate interactions between two entities. On the other hand, implicit relationships involve the broader context of multiple entities and can be represented as multi-step triples $(v_{e_1}, r_{e_1,e_2}, v_{e_2}), (v_{e_2}, r_{e_2,e_3}, v_{e_3})$,

Dataset	Prompt	Mask	Graph	NSFW Type	Shot Mode	Scene Type
UGVD	✗	✗	✗	5	1	2
Ours	✓	✓	✓	**7**	**3**	**3**

Table 1: **Comparison with the SOTA NSFW T2V datasets.** In addition to videos and prompt annotations, our dataset also includes entity masks and scene graphs. Compared to existing video datasets, ours encompasses seven categories of NSFW classes and spans three shot modes, whereas prior datasets typically cover only a single meta-scene.

where e_1 and e_3 denote the primary subject and object entities, respectively, and e_2 acts as an intermediate entity that facilitates the action, ultimately forming the complete relation $(v_{e_1}, r_{e_1,e_3}, v_{e_3})$, as the right video illustrated in Figure 2. This representation captures complex or indirect interactions that require reasoning over inter-entity context to identify.

To provide a more precise representation of video content, we introduce three scene types, *i.e.*, entities, explicit relationships, and implicit relationships, to provide a practical framework for modeling NSFW content in videos [9, 60].

3.3 NSFW Video Dataset

Among existing NSFW video benchmarks, UGVD represents one of the most comprehensive datasets. As summarized in Table 1, UGVD [33] contains 2k videos generated with MagicTime using stable-diffusion-v1-5, totaling 32k frames, with five NSFW content categories, two scene types (*i.e.*, attributes and explicit relationships) and one shot mode (*i.e.*, camera shot). The videos in UGVD are significantly short, averaging only 16 frames per sample, making it difficult to cover meaningful semantic changes over time. Moreover, as presented in Figure 3, the videos in UGVD suffer from poor quality, with many entities appearing blurry or distorted, which further limits their usability for NSFW video detection.

To facilitate community research on NSFW AIGC video detection, based on our NSFW video and scene graph analysis, we collected 23 representative real-world videos from online platforms (*e.g.*, YouTube) to observe real-world SFW/NSFW manifestation patterns. For each identified pattern, we curated seed prompts directly from the *Explore* section or *Inspiration* tool of the respective video-generation platforms, ensuring that all collected prompts complied with each platform's usage policies. These prompts were then expanded using OpenAI-GPT to increase diversity. The expanded prompt set was subsequently used on the same platforms to synthesize 540 videos via the video generators (*e.g.*, HailuoAI [18] and ChatUPAI [2]). After quality filtering, we retained 340 high-quality samples comprising **51k** frames, covering **7** NSFW content categories and **3** shot modes. More details are provided in the Appendix B.

We compare our dataset with existing ones in Table 1. Our

Figure 3: **Visualization of UGVD and our dataset.**

dataset exhibits significantly greater comprehensiveness and diversity. It offers stronger support for capturing temporally evolving NSFW content, which is often underrepresented in current benchmarks due to short video lengths and limited content scope. This higher-quality and comprehensive dataset enables more realistic evaluations of model performance under diverse threat scenarios. To advance research in multimodal NSFW content recognition, each video sample in our dataset includes: the raw video, sequenced frames, per-frame segmentation masks for detected entities, the natural language prompt used during synthesis, and a scene graph sequence that encodes video semantics.

4 Methodology

4.1 Motivation

As our observation in Section 3, unlike image, NSFW content in videos often manifests only through a sequence of frames. Specifically, moderation-relevant content in videos primarily manifests through ❶ individual entities together with their classes and attributes and ❷ inter-entity interactions between entities. Attributes, such as nudity, gore, and horror, can appear in a single frame and are identifiable by distinctive visual features like color and texture. In contrast, inter-entity interactions are generally expressed across multiple frames, such as self-harm or violence (*e.g.*, shooting). As a result, these interactions present a greater challenge for single-frame detection methods, which often fail to accurately capture the underlying NSFW attributes and relationships.

Moreover, more sophisticated generation instructions can introduce non-continuous shots that obscure the underlying semantic intent. For instance, by inserting unrelated frames, *e.g.*, flashbacks of a victim during a violent scene, can mislead detection models by breaking the continuity of actions. These complex real-world scenarios often involve multi-entity interactions and intricate temporal dependencies across frames, thereby posing significant challenges for multi-frame-based detection methods. Therefore, video-oriented detectors must go beyond single-frame or short-range sequence analysis and be capable of modeling long-range temporal context, particularly when critical semantics are distributed across non-contiguous frames, *i.e.*, discontinuous takes.

Detector	Shot Mode			Scene Type		
	Camera	Con.	Discon.	Attr.	E.R.	I.R.
MHSC	89.49	57.92	58.70	92.23	61.76	48.56
NSFW JS	59.88	50.05	63.86	55.94	60.66	44.48
CI	86.27	59.31	38.75	85.37	52.66	57.60
Qwen	98.04	82.35	68.75	100.00	75.74	83.20
Avg	83.42	62.41	57.52	83.39	62.71	58.46

Table 2: **Motivation studies.** We conducted a preliminary evaluation of existing methods on our self-constructed dataset and report their accuracy (%). Here, "Con." and "Discon." denote continue and discontinue takes, while "Attr.," "E.R.," and "I.R." correspond to attributes, explicit relationships, and implicit relationships, respectively.

We evaluated the performance of existing techniques on our self-built dataset, Unsafe-VidGraph, focusing on different shot types as well as scene categories. The detection results were assessed using the accuracy, as summarized in Table 2. We evaluate our approach on four detectors: (i) MHSC [36], a SOTA single-frame detector; (ii) NSFW JS (Video) [38], a widely used NSFW detection system; (iii) Tencent's commercial tool CI[3]; and (iv) a commercial content moderation based on a SOTA VLM Qwen [5].

Our analysis of existing approaches reveals that the majority of current methods still rely primarily on single-frame analysis for NSFW content detection. These models demonstrate strong performance in single-frame detection, achieving an accuracy of 83.42% on average as high as 98.04%. Nevertheless, their effectiveness drops significantly when handling more complex shot types. In particular, performance deteriorates in consecutive and inconsistent shot settings, where the average accuracy decreases to 62.41% and 57.52%, respectively. This trend suggests that multi-frame dependencies and cross-shot consistency, which are essential in realistic video streams, remain largely unaddressed by existing detectors.

Furthermore, existing methods remain limited in capturing non-salient NSFW elements, particularly relationships that span both explicit and implicit inter-entity interactions. Although they perform strongly on attribute detection (average accuracy 83.39%, peaking at 100.00%), their effectiveness drops sharply for relation-based categories: 62.71% for explicit and 58.46% for implicit relationships. This disparity highlights a persistent gap between attribute-level recognition and relation-aware understanding in NSFW video analysis. A key limitation lies in their reliance on directly encoding consecutive frames and analyzing the resulting global embeddings, which neglects inter-frame relational information and hinders the ability to model entity-level interactions across frames [43]. As a result, these models frequently fail to accurately identify NSFW content that is temporally dispersed

[3]Cloud Infinite. A secure, stable and efficient cloud data processing service for picture, video, audio, documents, and other data processing.

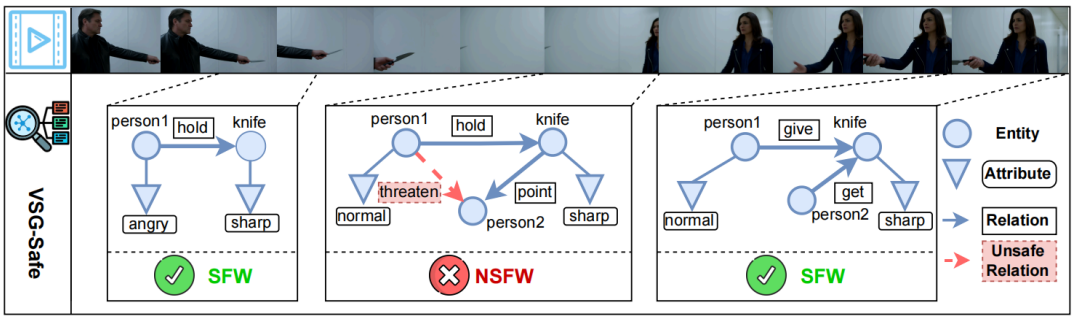


Figure 4: **Overview of our approach.** We first construct scene graphs from videos using a cross-frame extractor that captures entities, attributes, and inter-entities relationships. These graphs are then fed into a dual-channel GNN classifier, which separately models entity-level and relation-level semantics to detect NSFW content.

or semantically embedded within complex and dynamic visual contexts, leading to substantial deficiencies in real-world detection [44, 61].

4.2 Overview

To address the aforementioned limitations, we propose a novel framework for NSFW video detection by reasoning over entities and relationships in cross-frame scenarios. As illustrated in Figure 4, our method first ❶ extracts and analyzes fine-grained information within a video to generate a video scene graph, overcoming the limitations of traditional methods that rely heavily on static frame features or global video embeddings; and ❷ classify the video scene graph in open-ended scenarios to detect NSFW videos. In the following sections, we elaborate on extracting scene graphs from open-set videos to obtain entities and their potential relationships, and building a scene graph classifier to identify NSFW videos.

4.3 Video Scene Graph Generation

Unlike existing methods that rely solely on single-frame or video-level features, our approach leverages temporal semantic cues across multiple frames to detect NSFW content. This design enables fine-grained frame-level safety assessment while capturing the dynamic evolution of entities and their relationships over time. In Section 3, we represent the semantic content of videos using scene graphs. Building on this, we propose an open-vocabulary dynamic scene graph extractor tailored for video content, which consists of three main components: (i) entity segmentation; (ii) entity attribute prediction; and (iii) inter-entity relationship inference. In contrast to prior video scene graph generation methods, which typically lack attribute prediction and restrict relationships to individual frames, our approach supports both attribute extraction and relationship modeling within and across frames. Such video scene graphs satisfy the requirements for a complete scene graph as defined in Section 3.1, including entities, attributes, relations, and their cross-frame evolution [9, 60].

Entity segmentation. To construct video scene graphs, we first extract consistent entity tracks from input videos. While PVSG [51] provides Image Panoptic Segmentation (IPS) for object segmentation and cross-frame association, IPS modules are optimized for real-world videos and degrade significantly on AIGC content. In AIGC videos, entities often exhibit ambiguous or irregular boundaries, particularly at edges, resulting in suboptimal segmentation that can hinder reliable tracking. In addition, the default PVSG pipeline suffers from identity conflicts: same-class entities may be erroneously merged into a single track if they never appear in the same frame, or re-assigned new entity IDs upon reappearance. These misassociations obscure the temporal continuity of entities and propagate errors into downstream relation reasoning, producing unreliable scene graphs. Therefore, we adopt the entity segmentation module from PVSG and retrain it for our setting, where the output is disentangled into two branches, *i.e.*, **mask classification** and **mask prediction**.

In addition, each query is matched against the set of historical entities to determine whether it corresponds to an existing entity or a newly appearing one. If a matching historical entity is found, the query inherits its track_ID; otherwise, a new, globally unique track_ID is assigned. Similarity between the query and historical entities is measured against a predefined threshold to ensure temporal consistency of track_IDs. This design allows each query to simultaneously support semantic classification, mask prediction, and globally unique, temporally consistent track_ID assignment, preventing identity collisions while maintaining frame-level scene graph consistency.

During training, each object query is supervised via *mask-based bipartite matching* to ensure it uniquely corresponds to a ground-truth object. We define the matching loss as:

$$\mathcal{L}_{\text{match}}(q_i, y_j) = \lambda_{\text{cls}} \cdot \mathcal{L}_{\text{cls}}(q_i, y_j) + \lambda_{\text{mask}} \cdot \mathcal{L}_{\text{mask}}(q_i, y_j), \quad (1)$$

where $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{mask}}$ respectively quantify the discrepancies in classification and mask prediction for each object query,

while the coefficients λ_{cls} and λ_{mask} control and balance the relative contributions of these two loss components to the overall matching objective. Further, the optimal assignment $\hat{\sigma}$ is then obtained by minimizing the total matching cost:

$$\hat{\sigma} = \arg\min_{\sigma \in \mathcal{S}_N} \sum_{i=1}^{N} \mathcal{L}_{\text{match}}(q_i, y_{\sigma(i)}), \tag{2}$$

where $\mathcal{S}_N$ represents all possible one-to-one assignments and $\sigma(i)$ is the ground-truth index assigned to q_i. This framework ensures that each query captures both the semantic identity and spatial localization of its corresponding entity.

During inference, we extend queries temporally using **UniTrack** [46], linking panoptic segmentation results across frames into temporally consistent *entity tubes*. For each tracked entity, we construct a **query tube** $\{q_i^t\}_{t=t_1}^{t_2}$, where q_i^t represents the entity's query at frame t. Each query tube captures and preserves an entity's motion-related attributes in continuous numerical form, including its position, shape, size, and their temporal derivatives.

Entity attribute prediction. As discussed in Section 3, the attributes of entities are decisive for determining whether video content exhibits NSFW-related visual cues. However, most video scene graph pipelines optimize entity and relation prediction while giving little explicit treatment to attributes, even in recent SOTA formulations [30, 51]. To close this gap, we propose a mask-guided entity attribute extraction module. The module takes the frame and a per-entity instance mask as input. For each entity, we apply the corresponding mask to spatially ground the region of interest and query a visual-language model(VLM) to perform open-domain attribute extraction on that region. This design can not only localize attribute inference to the correct pixels, but also remain open-vocabulary attribute inference via the VLM. Specifically, our attribute extraction module takes three inputs: the original frame $\mathcal{F}$, the entity information $\{(v,c)\}$, and the entity location masks $\{m\}$, where each entity is associated with a corresponding mask.

We adopt a VLM, BLIP [24], as the backbone to support open-vocabulary attribute extraction. For i-th entity in $\mathcal{F}$, its attributes a_i are obtained as $a_i = VLM(F, \{(v_i, c_i)\}, \{m_i\})$. To further enhance the capability of the VLM, we fine-tune it to better adapt to the mask-based entity attribute extraction task. In our VLM fine-tuning setup, the answer classification objective is formulated as a soft-target multi-class cross-entropy loss. Let $\hat{p} \in \mathbb{R}^K$ denote the predicted probability distribution over the K possible answers, and let $p \in \mathbb{R}^K$ be the corresponding soft target distribution. The loss is then defined as $L_{cls} = -\sum_{i=1}^{K} p_i \log(\hat{p}_i)$, which generalizes standard cross-entropy by allowing supervision with soft distributions rather than hard one-hot labels, thereby capturing uncertainty and encouraging the model to align with the target distribution.

Inter-entity relationship prediction. Based on the extracted temporal query tubes, we predict inter-entity relations by identifying semantical entity pairs. During inference, we employ

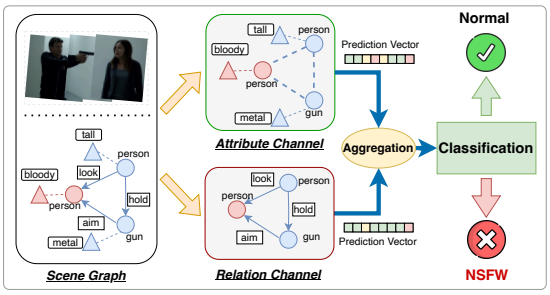


Figure 5: **Dual-GNN**. To precisely identify NSFW content, our Dual-GNN is designed to process complex scene graphs via two channels, *i.e.*, attribute channel and relation channel.

a compact and trainable **pairing module** to reduce computational complexity. This module takes temporal query tubes as input and models each entity using a transformer encoder with global cross-frame attention, aggregating contextual information across all time steps. Max-pooling along the temporal dimension compresses each query tube into a single vector, enabling efficient computation of pairwise similarities. From the resulting similarity matrix $S \in \mathbb{R}^{N \times N}$, we construct a sparse entity-pairing matrix by selecting high-confidence pairs that are likely involved in semantic relations.

Prior entity pairing may fail to capture behaviors that entities exhibit toward themselves. To address this, we introduce **self-relations** in the sparse pairing matrix, treating each entity as both subject and entity. This design allows the model to capture not only inter-entity relations but also intra-entity behaviors. Consequently, the sparse pairing matrix represents both conventional entity pairs and potential behaviors of single entities. It provides additional supervisory signals, enhancing the detection of actions and improving the precision of inter-entity relation matching. To supervise the pairing process, we align the sparse pairing matrix with ground-truth relation labels $Y \in \{0,1\}^{N \times N}$ using a multi-label cross-entropy loss:

$$\mathcal{L}_{\text{pair}} = -\sum_{i=1}^{N}\sum_{j=1}^{N} [Y_{ij} \cdot \log S_{ij} + (1 - Y_{ij}) \cdot \log(1 - S_{ij})] \tag{3}$$

This objective supervises the similarity scores between object queries, encouraging the model to assign higher confidence to pairs that are semantically related and suppress unlikely ones.

For relation modeling, we leverage a Transformer encoder. To resolve the Transformer's lack of temporal awareness, we incorporate positional embeddings that encode frame order. The encoder then processes the fused query pair features with cross-attention across time, allowing it to capture not only individual entity dynamics but also temporally evolving interactions between entity pairs.

4.4 Dual-Channel GNN for NSFW Detection

Using scene graph generation, we extract structured graphs from video frames, where nodes denote entities and attributes,

and edges capture relationships. Such data naturally exhibit explicit graph structures, requiring models that can capture complex topological dependencies. This makes Graph Neural Networks (GNNs) well-suited for such data, as their message-passing mechanism efficiently propagates and aggregates information across nodes and edges, enabling joint modeling of attributes and relations [20]. This allows the model to identify entities and relationships across entities.

Previous studies have shown that node attributes can interfere with structural information propagation in graph neural networks [58]. To address this, we design a dual-channel GNN architecture that separately handles attributes and relations, combined with multi-frame entity alignment and tracking to maintain cross-frame consistency. Each entity's unique track_ID is encoded as a node feature, ensuring semantic consistency and identity tracking during message propagation. Attributes remain independent of relational propagation while still participating in message passing, enabling effective modeling of entity states and properties. This design allows the model to simultaneously achieve: (a) cross-frame relational reasoning, such as detecting repeated threats; (b) intra-frame attribute recognition, such as identifying nudity or severe bodily injuries. Leveraging the inherent advantages of graphs, our framework provides fine-grained, cross-frame detection.

To model interactions between entities, we adopt GATv2 [8] as the backbone, which is a graph attention network that dynamically computes attention weights based on the joint features of both source and target nodes. In our framework, the embedding of each relation is injected as contextual input to each edge, enabling the attention mechanism to prioritize semantically critical interactions (*e.g.*, knife $\rightarrow$ point $\rightarrow$ person) while suppressing redundant or irrelevant connections. Finally, the outputs from the attribute and structural channels are projected into a shared feature space, allowing unified modeling of both entity attributes and inter-entity relations.

Identifying NSFW nodes based on their entity classes and attributes. For the attribute channel, the input is derived from the constructed video scene graph. During inference, this channel intentionally omits relational information (*i.e.*, the edge types and directions) while retaining only the connectivity among entities to preserve the underlying graph structure, *i.e.*, $\mathcal{G}' = \{\{(v_i, c_i, a_i)\}, \{(v_i, -, v_j)\}\}$, where "-" indicates that the relation type has been omitted. Each node in the graph is then encoded by an attribute encoding layer to produce a set of node embeddings. This embedding set, together with the adjacency matrix generated from $\{(v_i, -, v_j)\}$, is passed into GATv2 to obtain the final attribute prediction vector $x_{attribute}$. This design allows each node to integrate contextual information from its neighbors while maintaining its intrinsic attribute features, *e.g.*, an **aggressive** person and another **fearful** person. Compared to isolated attribute modeling, it improves the discriminative power and robustness of attribute representations, thereby providing a stronger foundation for subsequent fusion with structural features.

Detecting NSFW edges based on inter-entity relationships. The input to the relation channel is the complete video scene graph, where nodes are represented by their semantic class and a unique object identifier, and edges carry relation labels, *i.e.*, $\mathcal{G} = (v_i, c_i, a_i), (v_i, r_{ij}, v_j)$. To prevent messages from similar nodes interfering and weakening semantically critical interactions, each node embedding is formed by fusing its class embedding with a unique object identifier embedding. These embeddings, together with the edge relation embeddings, are propagated through a two-layer GATv2 network: the first layer uses multi-head attention to capture diverse relation-specific dependencies, and the second layer uses a single head to restore embedding dimension and retain fine-grained edge information, while the final activation is omitted to preserve inhibitory effects. The resulting node embeddings $x_{relationship}$ integrate context from semantically meaningful neighbors, emphasizing critical interactions and maintaining consistent entity identities. By incorporating relation embeddings and instance-level identifiers into the attention mechanism, the model mitigates interference among same-class nodes and highlights semantically important edges, thereby enhancing interpretability and providing a robust foundation for subsequent fusion with attribute features.

To effectively combine the complementary information from both channels while controlling their relative contributions, we adopt a gated mechanism in Equation 4, which allows the model to selectively emphasize the most relevant features from the attribute and relation embeddings.

$$x_{\text{mixed}} = g \cdot x_{\text{relationship}} + (1 - g) \cdot x_{\text{attribute}} \tag{4}$$

The gating weight $g \in [0, 1]$ is generated by a two-layer MLP with Sigmoid activation. It dynamically balances the contribution of relationship and attribute features for each dimension: a larger g emphasizes relational features, while a smaller g favors attribute features. This adaptive fusion highlights discriminative signals and suppresses noisy ones, improving both the robustness and interpretability of NSFW reasoning.

5 Experiments

5.1 Experiment Setup

Datasets. We evaluate all baseline methods and our proposed VSG-Safe approach on both the UGVD [33] dataset and our self-built dataset Unsafe-VidGraph, and detailed information about the two datasets is provided in Section 3.3. To further assess generalization of our approach, we also include an evaluation based on the expanded version of Unsafe-VidGraph, with its details deferred to Appendix B.

Baselines. We evaluate our method against a diverse set of baselines grouped into two categories: (1) image-targeted methods, including Q16 [40] and MHSC [36]; and (2) video-targeted methods, including the open-source tool NSFWJS (Video) [38],ViViT [42], UGCG-Guard [16], the commercial

Method	Camera Shot			Continuous Take			Discontinuous Take		
	ACC (%)	Recall (%)	F1 (%)	ACC (%)	Recall (%)	F1 (%)	ACC (%)	Recall (%)	F1 (%)
Q16	87.39	84.43	91.56	58.54	21.03	30.43	59.44	28.28	37.56
MHSC	89.49	87.02	93.06	57.92	4.81	8.97	58.70	8.11	14.49
NSFW JS	57.78	48.50	65.04	52.53	28.82	34.36	62.34	30.31	40.98
ViViT	50.98	32.43	48.98	61.27	33.33	39.69	45.00	29.63	42.11
CI	86.27	81.08	89.55	59.31	25.64	32.52	38.75	24.07	34.67
UGCG-Guard	88.05	99.11	93.07	55.18	93.01	64.15	51.47	88.90	61.25
Qwen-VL	98.04	97.30	98.63	82.35	83.33	78.31	68.75	64.81	73.68
Ours	**100.00**	**100.00**	**100.00**	**98.04**	**97.44**	**97.44**	**95.00**	**94.44**	**96.23**

Table 3: **Performance comparison with baselines across three shot modes.**

Method	UGVD [33]		Unsafe-VidGraph	
	ACC (%)	F1 (%)	ACC (%)	F1 (%)
Q16	51.36	38.10	64.02	53.46
MHSC	51.87	31.16	63.87	45.17
NSFW JS	61.17	49.99	56.03	44.29
ViViT	51.57	44.21	55.82	42.19
CI	60.77	41.29	58.51	47.55
UGCG-Guard	48.41	63.65	60.20	70.23
Qwen-VLM	60.42	52.12	81.49	81.44
Ours	**77.16**	**78.45**	**97.61**	**97.62**

Table 4: **Comparison on NSFW video detection.**

video moderation tool CI from Tencent, and Qwen-VL [5], where UGCG-Guard and Qwen-VL represent VLM-based NSFW detection approaches. Further implementation details of the baselines and setups can be found in Appendix A. Additionally, to evaluate the effectiveness of our Dual-GNN, we compare it with two representative non-GNN classifiers: ❶ BERT [11], which takes serialized scene graph triplets as input text sequences; ❷ GPT-4o of OpenAI used via prompt-based few-shot classification. These baselines serve to assess the classification ability of language models over scene graphs, even though our method does not rely on LLMs or BERT.

Evaluation metric. To properly evaluate our proposed VSG-Safe, we adopt multiple popular metrics in experiments. ❶ *Acceptance* (*ACC*) measures how accurately our system can accept legitimate samples and reject NSFW content. ❷ *Recall* is the proportion of all NSFW samples that were rejected. ❸ $F1$-*score* is the harmonic mean of the precision and recall. In the experiments, we take the normal video as the negative samples and the NSFW videos as the positive.

5.2 Main Results

We begin by evaluating VSG-Safe against baseline methods for NSFW content detection. The UGVD dataset [33], characterized by limited diversity in both shot modes and content categories. By contrast, our proposed Unsafe-VidGraph offers diverse shot modes and better reflects the complexity of real-world video content. Hence, unless otherwise noted, we use Unsafe-VidGraph as the default benchmark while also reporting results on UGVD for completeness.

Overall performance on NSFW detection. We begin by evaluating whether our method outperforms existing baselines across datasets in NSFW video detection. Table 4 summarizes the results, where VSG-Safe outperforms all baselines on both benchmarks. On UGVD, VSG-Safe achieves an $F1$-score of 78.45%, improving over baselines by 22.08%. On our dataset, VSG-Safe also consistently delivers the highest accuracy among all methods, achieving an $F1$-score of 97.62%, outperforming all baselines by an average margin of 42.72%. The strong performance demonstrates that VSG-Safe not only surpasses existing baselines but also generalizes well to practical video detection tasks.

Additionally, image-targeted NSFW detection methods, *i.e.*, Q16 and MHSC, achieve only F1-scores of 38.10% on UGVD and 53.46% on Unsafe-VidGraph at the best. Their poor performance stems from their entirely ignorance of inter-frame dependencies and thus can only capture attributes or explicit entity conflicts, leading to frequent false negatives. In contrast, video-targeted approaches perform substantially better. On UGVD, UGCG-Guard reaches an F1-score of 63.64%, while Qwen attains 81.44% on Unsafe-VidGraph, significantly outperforming the image-targeted detectors. This demonstrates the benefit of modeling inter-frame dependencies, as such approaches can capture NSFW content that emerges only through multi-frame relationships. Notably, off-the-shelf tools, including ViViT and NSFW JS, and the commercial tool CI, perform poorly on both datasets. Their reliance on CNN- or ViT-based frame encoders prevents them from capturing fine-grained NSFW cues [43, 44].

Overall, the experiments demonstrate the effectiveness of our framework and highlight the importance of extracting cross-frame content for reliable NSFW video detection.

NSFW detection across video shot modes. We further evaluate VSG-Safe and baseline methods across three video shot modes, *i.e.*, camera shots, continuous takes, and discontinuous takes, to assess the impact of video structure on detection performance. Table 3 summarizes the results. From the exper-

Method	Entities			Explict Relationship			Implicit Relationship		
	ACC (%)	Recall (%)	F1 (%)	ACC (%)	Recall (%)	F1 (%)	ACC (%)	Recall (%)	F1 (%)
Q16	91.01	87.51	91.32	57.06	28.99	38.90	59.46	31.02	44.78
MHSC	92.23	85.61	92.25	61.76	22.20	35.37	48.56	4.10	7.78
NSFW JS	55.36	28.89	41.15	62.88	44.81	53.23	42.67	21.49	28.43
ViViT	51.22	25.93	41.18	55.62	23.68	32.43	57.60	43.94	52.25
CI	85.37	77.78	87.50	52.66	34.21	39.39	57.60	24.24	37.65
UGCG-Guard	67.24	98.93	76.54	56.14	93.90	66.88	63.54	90.30	72.41
Qwen-VL	**100.00**	**100.00**	**100.00**	75.74	84.21	75.74	83.20	68.18	81.08
Ours	**100.00**	**100.00**	**100.00**	**98.22**	**100.00**	**98.06**	**96.00**	**92.42**	**96.06**

Table 5: **Detection performance across three scene types.**

imental results, we observe that all baselines achieve their best performance on the camera shots, while performance drops sharply on continuous takes and discontinuous takes. The degradation occurs because such visual content is more dispersed, which requires longer temporal reasoning; as a result, the signals become subtle and challenging to detect [42, 43]. Video shooting mode has a strong impact on detection performance, and robust NSFW detection requires handling long and discontinuous sequences.

Image-oriented detectors achieve high performance on camera shots but fail drastically on long takes, reflecting their inability to model temporal context. Video-oriented detectors are relatively more robust on continuous and discontinuous sequences, but their accuracy is still limited, as unrelated frames dilute the underlying semantic signals. These results highlight both the importance of inter-frame reasoning and the challenge of avoiding interference among frames. In contrast to existing methods, our approach is specifically designed to understand and reason on cross-frame content by leveraging scene graphs with localized temporal updates to connect semantics across discontinuous frames. Consequently, our method achieves competitive performance across all three shot types: 100% F1-score on camera shots, 97.44% F1-score on continuous takes, and 96.23% F1-score on discontinuous takes, demonstrating its exceptional capability in detecting complex, discriminative cues across multiple frames.

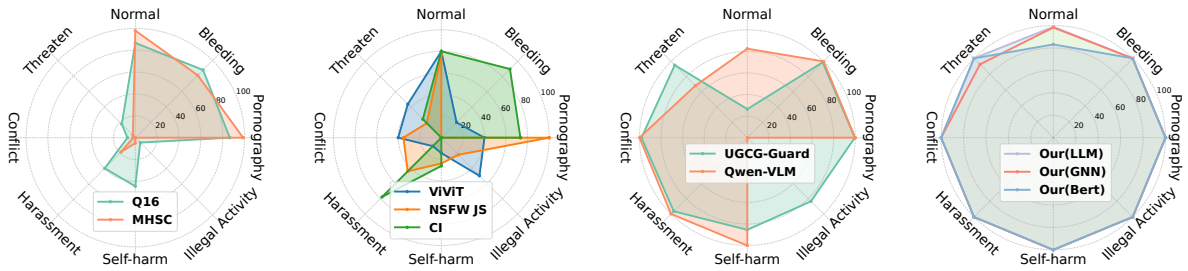
Overall, these results demonstrate that the shot mode has a profound impact on detection accuracy. Image-oriented methods succeed on camera shots but collapse on continuous or discontinuous takes due to a lack of multi-frame reasoning. Video-oriented methods are relatively more resilient, but their accuracy is hindered by interference from irrelevant frames. In contrast, VSG-Safe explicitly models cross-frame semantics through scene graphs with localized updates, achieving outstanding performance across all three shot modes.

Evaluating NSFW detection on diverse video scene types. Recall the analysis in Section 3.2, video content can be broadly categorized into three types that represent an increasing level of semantic and contextual complexity, *i.e.*, entity, explicit relationship, and implicit relationship. we evaluate baselines and our method across three scene categories to understand that *to what extent do different video scene types increase the difficulty of identifying NSFW content*. The results are summarized in Table 5.

Image-targeted methods (*e.g.*, Q16 and MHSC) perform well on attributes (>91% F1-score) but collapse on relationships. For example, MHSC reaches only a single-digit F1-score on implicit relations. Their CNN/ViT backbones focus on object-level features and lack mechanisms for modeling entity interactions, leaving them unable to capture semantic dependencies required for relational reasoning [43, 44, 61]. Conversely, video-oriented approaches, including NSFW JS, ViViT and Tencent's CI, exhibit more balanced but limited performance. Although they accept video-level inputs, they simply aggregate per-frame features without modeling fine-grained cross-frame semantics. As a result, their overall accuracy and recall remain inadequate for reliable deployment in complex scenarios. Among video-level detection methods, approaches leveraging VLMs, such as Qwen and UGCG-Guard, achieve clear advantages over other baselines, where Qwen attains substantially higher F1-score on both explicit and implicit relations, while UGCG-Guard yields similar but less precise results, reflecting its ability to capture semantics across both continuous and discontinuous frames. However, these gains largely rely on heuristic prompt engineering, which provides only limited and unstable cross-frame reasoning. This limitation is especially evident in UGCG-Guard: our analysis shows that it often misclassifies entities and misses inter-entity relationships, leading to lower precision compared to Qwen.

In contrast, our approach significantly outperforms all baselines. The GNN-based variant of VSG-Safe achieves superior performance across all categories, with F1-score of 98.06% on explicit relations and 96.06% on implicit relations. This is because our method accurately extracts entity attributes and relationships via scene graphs and leverages graph neural networks to reliably model the semantics of both intra-frame and cross-frame relations, enabling strong cross-frame reasoning.

Performance on different NSFW content detection. We evaluate ACC across 7 NSFW categories, where the results are shown in Figure 6. Sexual and bleeding are visually ex-



(a) **Image-targeted** (b) **Video-targeted** (c) **VLM-based** (d) **Our Methods**

Figure 6: **Evaluation across eight content categories.** The "Normal" is treated as the negative class, while the remaining seven types represent positive classes.

plicit and can often be identified from a single frame, while others require cross-frame or semantic reasoning, *e.g.*, conflict, threaten, and illegal activity.

As shown in Figure 6(a), frame-based methods (*e.g.*, Q16, MHSC) perform well on attribute categories (Sexual, Bleeding) but drop sharply on relational tasks: accuracy is below 20% on Threaten and near zero on Illegal Activity. Attribute categories contain explicit visual cues that can be recognized from a single frame, whereas relational categories require cross-frame semantic and interaction reasoning that frame-level features cannot provide. Video-level tools (ViViT, NSFW JS, CI) generalize better but still struggle on complex categories (Figure 6(b)). Their simple aggregation of pixel-level features prevents modeling detailed relationships [9, 60]. Notably, NSFWJS and ViViT are both pixel-level detectors, yet they behave very differently in their respective domains. NSFWJS succeeds on pornographic content because such cues are visually explicit and often identifiable from a single frame. In contrast, ViViT, even with extensive pretraining for violence recognition, still misclassifies violent actions due to weak semantic reasoning and the absence of explicit cross-frame modeling. VLM-based methods improve performance overall, particularly on Prongraphy and Conflict, benefiting from larger model scale. However, they remain unstable on complex cross-frame action scenes due to the absence of explicit temporal and multi-entity interactions [55, 60].

VSG-Safe consistently outperforms all baselines, achieving substantial gains across six NSFW categories and 92.42% accuracy on Threaten, the most challenging relational category, indicating strong generalization (Figure 6(d)). Overall, the results demonstrate that explicitly modeling cross-frame entities, attributes, and relationships is essential for detecting NSFW content. We also observe occasional recognition errors in rare cases, such as confusing a remote controller with a gun pointed at a person. These cases arise mainly in visually atypical situations that naturally occur infrequently and can be further reduced with light, targeted augmentation if needed.

Video scene graph extraction. Due to the significant differences between AIGC and real-world captured data, particularly the fact that entity boundaries in AIGC videos are often less distinct than those in real footage, and the presence of numerical anomalies such as unnatural pixel patterns, models trained on real-world data can easily be misled during scene graph extraction. As a result, existing techniques such as PVSG [51] and BLIP [24] cannot be directly applied to AIGC scenarios. Additionally, existing video scene graph extraction methods, *e.g.*, PVSG, are inherently limited in their ability to capture fine-grained semantics. They do not support *entity attribute* extraction, and also fail to model interactions between entities that do not appear within the same frame, both of which are essential for accurate NSFW content understanding. To this end, we fine-tuned BLIP for attribute extraction in the AIGC and provided a new video scene graph generation framework, VSG-Safe, built upon and substantially extended from PVSG. Table 6 reports the comparative results obtained on the self-built dataset. PVSG achieved an ACC of 42.94% in entity extraction and 41.23% in relationship extraction, but lacked attribute extraction capability. BLIP, while able to extract attributes, achieved only 41.12% accuracy. In contrast, our method significantly outperforms both baselines, achieving 93.91%, 98.93%, and 95.60% accuracy in entity, attribute, and relationship extraction, respectively.

Method	Entity Extraction	Attribution	Relationship
PVSG [51]	42.94%	N/A	41.23%
BLIP [24]	N/A	41.12%	N/A
Ours	**93.91%**	**98.93%**	**95.60%**

Table 6: **Evaluating scene graph extraction performance between baselines and proposed method.** Note that PVSG lacks native attribute extraction capabilities and in this work, we implemented attribute extraction solely through BLIP with our mask. Therefore, functionalities not inherently supported by the base models are indicated with "N/A".

Generalization to out-of-distribution scene graph samples. We further evaluate VSG-Safe on the expanded Unsafe-

	ALL	In-Distribution	Out-of-Distribution
ACC	95.05%	96.77%	94.61%

Table 7: **Generalization on expanded Unsafe-VidGraph.**

VidGraph (3,007 samples with unseen visual elements; see Appendix B) to assess whether it maintains reliable NSFW detection under unseen entities, attributes, and relations. As shown in Table 7, accuracy on out-of-distribution samples decreases slightly compared to in-distribution ones (94.61% vs. 96.77%), indicating robust generalization to unseen semantics. This behavior is consistent with the design of our scene-graph generator, which focuses on shared visual patterns (*e.g.*, shapes, textures, and motion). As a result, unseen elements are embedded near their most semantically related in-distribution categories, enabling reliable NSFW prediction even on previously unseen cases (*e.g.*, the unseen *slap* shares motion patterns and body poses with the learned *punch*, and unseen *blades* share shape and material with known *knives*).

5.3 Ablation Studies

Impact of the scene graph. To quantify the individual contributions of *entity attributes* and *inter-entity relations* to NSFW content detection, we perform a systematic ablation study using our self-built dataset. As discussed in Section 3.2, scene graphs in video frames may contain varying numbers of entities and multi-entity interactions, where attributes of the entity provide key semantics for detection in single-entity frames, while relations become increasingly important as entity interactions emerge. We evaluate four settings with the GNN-based classifier: ❶ **Full Graph**, the default setting using the full scene graph extracted by VSG-Safe, containing entity attributes and inter-entity relations. ❷ **Remove Attributes**, where all entity-level attribute nodes (*e.g.*, appearance) are omitted. ❸ **Remove Relations**, where all inter-entity edges (*e.g.*, action-based relations) are removed. ❹ **Ground Truth**, where predictions are made on ground-truth scene graphs, serving as an empirical upper bound.

As shown in Figure 7, our classifier achieves an ACC of 96.74% and an F1-score of 95.80% with the full graph, closely approaching the performance of the ground truth setting (ACC: 98.22%, F1-score: 96.82%). Removing attributes results in a noticeable drop in both ACC (Δ=-16.92%) and F1-score (Δ=-39.35%), confirming that attributes are crucial, particularly in videos where only one entity is present. More critically, removing inter-entity relations leads to a more severe degradation, where ΔACC=-35.91% and ΔF1-score=-60.52%, highlighting the importance of modeling interactions in multi-entities for NSFW detection.

These findings validate the design of our scene graph representation, demonstrating that both attributes and relations are essential, with relation modeling having a particularly significant impact on overall performance.

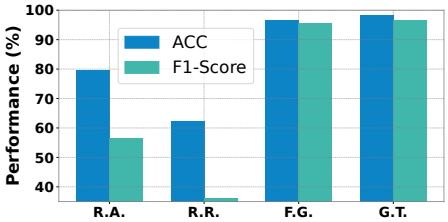


Figure 7: **Ablution studies of semantic scene graph generation.** Here, the *"F.G."* represents the full graph configuration, which includes both attributes and relations as extracted by VSG-Safe. *"R.A."* refers to the setting where all entity attributes are removed from the full scene graph, while *"R.R."* denotes the setting where all inter-entity relations are removed. Lastly, *"G.T."* signifies the use of ground-truth scene graphs.

Method	Performace (ACC,%)		
	Original	Enhanced	$\Delta ACC \uparrow$
UGCG-Guard	38.10	**85.71**	47.61
Qwen-VL	42.88	**95.24**	52.36

Table 8: **Enhance NSFW detection via scene graphs.**

5.4 Scene Graph Enhance

Existing VLM-based detectors often fail on videos requiring fine-grained relational reasoning, as they lack structured semantic context. To address this, we augment VLM prompts with scene graphs extracted from video content, enabling the model to reason jointly over raw frames and structured semantics(Box 2). We evaluate this approach on a sub-dataset of 42 challenging videos where vanilla VLMs underperform. As shown in Table 8, Qwen and UGCG-Guard achieve only 42.88% and 38.10% *ACC*, respectively. With scene graph augmentation, *ACC* rises to 95.24% and 85.71%, an average improvement of nearly 50 points. Incorporating structured scene-level prompts substantially improves the reasoning capability of VLM-based detectors, making them far more effective for NSFW content detection in challenging scenarios.

6 Conclusion

In this work, we highlight the challenge of detecting NSFW content dispersed across video frames and argue for semantic analysis beyond frame-level features. We propose VSG-Safe, a novel scene-graph-based framework with a dual-channel GNN classifier that captures entities, their attributes, and cross-frame relationships. In future work, we aim to work towards real-time NSFW detection, with the goal of improving applicability in latency-sensitive scenarios.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (NSFC) under Grants No. 62576255, No. 62202340, the Fundamental Research Funds for the Central Universities under No. 2042025kf0054, the Natural Science Foundation of Hubei Province under No. 2025AFB455.

Open Science

In accordance with the open science policy, this paper adheres to principles that promote transparency, accessibility, and reproducibility of research. The following measures have been implemented:

Data sharing: Our datasets are hosted on HuggingFace and access is restricted through *Gated User Access*. Please read the *COMMUNITY LICENSE AGREEMENT* carefully and provide your details to the repository manager using the collection form for the necessary information. https://huggingface.co/datasets/yuwan0/Unsafe-VidGraph.

Artifact availability: All source code, models architecture, and supplementary materials are hosted on https://doi.org/10.6084/m9.figshare.30902915, enabling further exploration of our methods.

Ethical Considerations

This work examines the detection of not-safe-for-work (NSFW) content in text-to-video (T2V) systems from a technical perspective. Here, NSFW is treated as an operational category commonly adopted in prior works and platform practices to indicate content subject to moderation in specific deployment contexts, rather than a normative judgment about morality, legality, or inherent harm. Our contribution focuses on improving the technical identification of specific visual patterns frequently moderated in practice. We do not address broader risks of T2V technologies, such as misinformation or intellectual property, which lie beyond the scope of this study.

Stakeholders, dual use, and publication. The primary stakeholders in this work include (i) users and bystanders who may be exposed to harmful T2V content, (ii) platform operators and model providers who must moderate such content at scale, (iii) individuals whose likeness might be mimicked by generative models, (iv) the research and practitioner communities that rely on robust NSFW detection tools and (v) individuals and communities who rely on NSFW content for benefits. Our methods are intended to support content moderation and defensive analysis, while acknowledging that insights into detection systems may also inform adversarial evasion. To balance ethical considerations with Open Science principles, we adopt differentiated release policies for code and data: the model architecture and code are released openly for transparency, whereas the dataset, which contains sensitive NSFW video content, is shared only under controlled-access protocols. Specifically, only sanitized subsets that comply with platform policies, pass internal ethics review, and receive IRB approval are eligible for request-based access. This strategy supports reproducibility while keeping dissemination proportionate to the associated ethical risks.

Data sources and synthetic content. All data collection complied with platform policies and terms of service, without bypassing safety mechanisms. Prompts were obtained from publicly accessible "Explore" or "Inspiration" sections and expanded using an OpenAI GPT model under its safety constraints. The study was reviewed by our IRB and determined not to constitute human-subjects research, as no personally identifiable or user-contributed private data were involved. We nevertheless conducted an independent ethics analysis aligned with community norms beyond institutional requirements.

Safety of users and bystanders. Our aim is to support platform-level moderation of T2V systems by improving the detection of NSFW content (e.g., sexual or violent material). Such detection helps reduce unintended exposure risks when T2V technologies are misused.

Protection of research team members. As this study involves potentially disturbing visual content, we took steps to mitigate psychological risks throughout the research process.All manual annotations were conducted exclusively by the authors, without involving external annotators. Participation was voluntary, with prior guidance on psychological safety and the nature of the content, and annotators were free to pause or withdraw at any time without penalty.

Privacy and resemblance to real persons. Although our data are synthetic, generated faces may occasionally resemble real individuals. Prior to use, we applied a conservative screening step using third-party face search services (e.g., PimEyes, FaceCheck), in accordance with their terms of use, to identify and exclude videos closely matching publicly indexed individuals. No matched identities were stored or further processed, and this step was used solely to remove potentially sensitive content rather than for model training or analysis.

Impacts on Individuals and Communities Relying on NSFW Content. We acknowledge that individuals and communities who rely on NSFW content for livelihood or expression (*e.g.*, adult content creators and sex workers) are an important stakeholder group potentially affected by automated NSFW detection. Such content may serve legitimate roles under posting policies in consensual adult interaction or economic participation. Although our method is a detection component rather than an enforcement mechanism, such systems may contribute to over-censorship, stigmatization, or economic harm when integrated into moderation pipelines, as documented in prior work. We emphasize that our work does not define platform policies or enforcement decisions, and mitigating these broader risks requires governance and human-in-the-loop mechanisms beyond the scope of this study.

Gender and identity considerations. In light of the potential harms associated with stereotyping, we adopt mitigation measures by avoiding reliance on protected or sensitive attributes [27] (*e.g.*, gender, race, or national origin) as signals for NSFW detection, and by using neutral labels (*e.g.*, person) in detection, figures, and qualitative analyses. We acknowledge that T2V systems and automated detection may nonetheless reproduce or amplify societal stereotypes along identity dimensions. While such biases cannot be fully eliminated by our technical design alone, as they may arise from contextual interpretation of NSFW content as well as downstream deployment and enforcement practices, broader concerns are more appropriately addressed through complementary governance mechanisms beyond the scope of this work.

Decision. We considered the ethical appropriateness of both conducting and publishing this work. From a beneficence perspective, we assess that the potential societal benefits of improving NSFW detection outweigh the associated risks. From a Respect for Persons perspective, risks to affected stakeholders are limited by the use of synthetic data, the exclusion of personally identifiable or user-contributed information, and safeguards for annotators and individuals who might otherwise be inadvertently mimicked. Given that remaining misuse risks are further mitigated through controlled, request-based access to sanitized datasets, we conclude that publishing this work under these conditions is ethically justified.

References

[1] Abdulrahman Adel. Real life violence detection. https://github.com/Abdulrahman-Adel/Real-Life-Violence-Detection, 2023.

[2] ChatUp AI. Nsfw ai chat, image & video generator. https://aichattings.com/, 2025.

[3] Protect AI. Llm guard. https://llm-guard.com/input_scanners/ban_substrings/, 2025.

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[6] Mazen Balat, Mahmoud Gabr, Hend Bakr, and Ahmed B Zaky. Tikguard: A deep learning transformer-based solution for detecting unsuitable tiktok content for kids. In *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 337–340. IEEE, 2024.

[7] Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? In *European Conference on Computer Vision*, pages 190–206. Springer, 2024.

[8] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *ICLR*, 2022.

[9] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1–26, 2021.

[10] Die Chen, Zhiwen Li, Cen Chen, Xiaodan Li, and Jinyan Ye. Comprehensive assessment and analysis for nsfw content erasure in text-to-image diffusion models. *arXiv preprint arXiv:2502.12527*, 2025.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[12] Jiali Duan, C-C Jay Kuo, et al. Bridging gap between image pixels and semantics via supervision: A survey. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2022.

[13] Lin Geng Foo, Hossein Rahmani, and Jun Liu. Ai-generated content (aigc) for various data modalities: A survey. *ACM Computing Surveys*, 57(9):1–66, 2025.

[14] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[15] Dongnan Gui, Xun Guo, Wengang Zhou, and Yan Lu. I2vguard: Safeguarding images against misuse in diffusion-based image-to-video models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12595–12604, 2025.

[16] Keyan Guo, Ayush Utkarsh, Wenbo Ding, Isabelle Ondracek, Ziming Zhao, Guo Freeman, Nishant Vishwamitra, and Hongxin Hu. Moderating illicit online image promotion for unsafe user generated content games using large {Vision-Language} models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 5787–5804, 2024.

[17] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, et al. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.

[18] HailuoAI. Hailuo: Transform idea to visual. https://hailuoai.video/, 2025.

[19] Claire Wonjeong Jo, Magdalena Wojcieszak, et al. Harmful youtube video detection: A taxonomy of online harm and mllms as alternative annotators. *arXiv preprint arXiv:2411.05854*, 2024.

[20] Bharti Khemani, Shruti Patil, Ketan Kotecha, and Sudeep Tanwar. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data*, 11(1):18, 2024.

[21] KlingAI. Klingai: From vision to screen. https://klingai.com/global/, 2025.

[22] Mijat Kustudic and Gustave Florentin Nkoulou Mvondo. A hero or a killer? overview of opportunities, challenges, and implications of text-to-video model sora. *Authorea Preprints*, 2024.

[23] Warren Leu, Yuta Nakashima, and Noa Garcia. Auditing image-based nsfw classifiers for content filtering. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1163–1173, 2024.

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

[25] Siyuan Liang, Jiayang Liu, Jiecheng Zhai, Tianmeng Fang, Rongcheng Tu, Aishan Liu, Xiaochun Cao, and Dacheng Tao. T2vshield: Model-agnostic jailbreak defense for text-to-video models. *arXiv preprint arXiv:2504.15512*, 2025.

[26] Xuannan Liu, Xing Cui, Peipei Li, Zekun Li, Huaibo Huang, Shuhan Xia, Miaoxuan Zhang, Yueying Zou, and Ran He. Jailbreak attacks and defenses against multimodal generative models: A survey. *arXiv preprint arXiv:2411.09259*, 2024.

[27] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[28] Yibo Miao, Yifan Zhu, Lijia Yu, Jun Zhu, Xiao-Shan Gao, and Yinpeng Dong. T2vsafetybench: Evaluating the safety of text-to-video generative models. *Advances in Neural Information Processing Systems*, 37:63858–63872, 2024.

[29] Fatemeh Nazarieh, Zhenhua Feng, Muhammad Awais, Wenwu Wang, and Josef Kittler. A survey of cross-modal visual content generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6814–6832, 2024.

[30] Trong-Thuan Nguyen, Pha Nguyen, Jackson Cothren, Alper Yilmaz, and Khoa Luu. Hyperglm: Hypergraph for video scene graph generation and anticipation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29150–29160, 2025.

[31] Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*, 2024.

[32] OpenAI. Sora: Creating video from text. https://openai.com/sora/, 2025.

[33] Yan Pang, Aiping Xiong, Yang Zhang, and Tianhao Wang. Towards understanding unsafe video generation. *Network and Distributed System Security (NDSS) Symposium 2025*, 2025.

[34] Antonia Petrogianni, Panagiotis Koromilas, and Theodoros Giannakopoulos. Film shot type classification based on camera movement styles. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 602–615. Springer, 2022.

[35] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Safe-clip: Removing nsfw concepts from vision-and-language models. In *European Conference on Computer Vision*, pages 340–356. Springer, 2024.

[36] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. 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, 2023.

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[38] Infinite Red. Nsfw js. https://github.com/infinitered/nsfwjs, 2024.

[39] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023.

[40] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1350–1361, 2022.

[41] Mukul Singh, José Cambronero, Sumit Gulwani, Vu Le, Carina Negreanu, and Gust Verbruggen. Codefusion: A pre-trained diffusion model for code generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11697–11708, 2023.

[42] Sanskar Singh, Shivaibhav Dewangan, Ghanta Sai Krishna, Vandit Tyagi, Sainath Reddy, and Prathistith Raj Medi. Video vision transformers for violence detection. *arXiv preprint arXiv:2209.03561*, 2022.

[43] Sethuraman TV, Savya Khosla, Vignesh Srinivasakumar, Jiahui Huang, Seoung Wug Oh, Simon Jenni, Derek Hoiem, and Joon-Young Lee. Frame: Pre-training video feature representations via anticipation and memory. *arXiv preprint arXiv:2506.05543*, 2025.

[44] Peng Wang, Lingqiao Liu, Chunhua Shen, and Heng Tao Shen. Order-aware convolutional pooling for video based action recognition. *Pattern Recognition*, 91:357–365, 2019.

[45] Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3907–3916, 2024.

[46] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *Advances in neural information processing systems*, 34:726–738, 2021.

[47] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *Advances in Neural Information Processing Systems*, 37:34322–34348, 2024.

[48] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, 31(2):1526–1541, 2024.

[49] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024.

[50] Yuanhao Xiong, Long Zhao, Boqing Gong, Ming-Hsuan Yang, Florian Schroff, Ting Liu, Cho-Jui Hsieh, and Liangzhe Yuan. Structured video-language modeling with temporal grouping and spatial grounding. *ICLR*, 2024.

[51] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, et al. Panoptic video scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18675–18685, 2023.

[52] Mengjiao Yang, Yilun Du, Bo Dai, Dale Schuurmans, Joshua B Tenenbaum, and Pieter Abbeel. Probabilistic adaptation of text-to-video models. *arXiv preprint arXiv:2306.01872*, 2023.

[53] Xiaoyu Ye, Songjie Cheng, Yongtao Wang, Yajiao Xiong, and Yishen Li. T2vunlearning: A concept erasing method for text-to-video diffusion models. *arXiv preprint arXiv:2505.17550*, 2025.

[54] Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. *ICLR*, 2025.

[55] Cheng Zhang, Wei-Lun Chao, and Dong Xuan. An empirical study on leveraging scene graphs for visual question answering. *BMVC 2019*, 2019.

[56] Yiyuan Zhang, Yuhao Kang, Zhixin Zhang, Xiaohan Ding, Sanyuan Zhao, and Xiangyu Yue. Interactivevideo: User-centric controllable video generation with synergistic multimodal instructions. *arXiv preprint arXiv:2402.03040*, 2024.

[57] Yuyang Zhang, Kangjie Chen, Xudong Jiang, Jiahui Wen, Yihui Jin, Ziyou Liang, Yihao Huang, Run Wang, and Lina Wang. {USD}:{NSFW} content detection for {Text-to-Image} models via scene graph. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 879–895, 2025.

[58] Yuchen Zhou, Yanmin Shang, Yanan Cao, Qian Li, Chuan Zhou, and Guandong Xu. Api-gnn: attribute preserving oriented interactive graph neural network. *World Wide Web*, 25(1):239–258, 2022.

[59] Zijian Zhou, Zheng Zhu, Holger Caesar, and Miaojing Shi. Openpsg: Open-set panoptic scene graph generation via large multimodal models. In *European Conference on Computer Vision*, pages 199–215. Springer, 2024.

[60] Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, and Mohanmed Bennamoun. Scene graph generation: A comprehensive survey. *Neurocomputing*, 2023.

[61] Jiagang Zhu, Zheng Zhu, and Wei Zou. End-to-end video-level representation learning for action recognition. In *2018 24th international conference on pattern recognition (ICPR)*, pages 645–650. IEEE, 2018.

[62] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, et al. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM workshop on large AI systems and models with privacy and safety analysis*, pages 57–68, 2023.

A Implementation Details of Baselines and Additional Experimental Settings

All baseline methods were originally proposed for visual NSFW content detection. As several methods do not release their training datasets or full training configurations, we follow the authors' recommendations and adopt the officially provided best-performing pretrained checkpoints whenever available. For methods lacking official implementations, we use the available community implementations, with all sources explicitly documented.

To ensure a fair and consistent comparison across baselines that differ in their label spaces, we follow prior works (*e.g.*, MHSC [36], UGCG [16]) and unify all methods under a binary SFW/NSFW classification setting. This normalization prevents discrepancies arising from heterogeneous category definitions and aligns with widely adopted safety evaluation practices. Additionally, because several baseline methods are specialized, such as NSFWJS and ViViT, which focus on detecting sexuality and violence respectively, we adopt per-category comparison in radar charts to align their strengths with our evaluation taxonomy, enabling these category-specific baselines to be incorporated fairly and facilitating a balanced comparison across different NSFW content. To further reduce distributional bias and ensure fairness, we also evaluate all methods on public datasets UGVD, and real-world datasets (*i.e.*, Big-Porn, Youtube-8M, and HBD), which are independent of our constructed dataset and unseen during model development.

Q16. We adopt the officially released NSFW detection model and use the best-performing weights provided by the authors [40]. Following their recommendations, we employ the largest and most recent variant, ViT-L/14, together with the corresponding best checkpoint supplied in their repository.

MHSC. We adopt the officially released Multi-head SafetyClassifier (MHSC) detection model and use the best-performing weights provided by the authors [36]. Following their implementation, we use the default ViT-L/14 variant and config file, together with the corresponding best prompt checkpoints supplied in their repository, as our baseline.

NSFWJS. We use the official JavaScript implementation of NSFWJS and adopt its best-performing detector, InceptionV3, together with the official pre-trained checkpoint [38].

Box 1: Prompt for Qwen

You are an expert in NSFW video content detection. You will receive a short video clip as input. Evaluate the clip holistically by analyzing all visible cues across frames, rather than judging any single frame in isolation.
NSFW Criteria:
A video should be classified as NSFW if it contains any content belonging to [Your NSFW Categories], which encompass visually identifiable harmful, inappropriate, or safety-critical behaviors.
Safe Criteria:
Do not classify the video as NSFW if it only contains benign activities or other visually neutral behaviors.
#Your task:
Decide whether the video is NSFW (Not Safe For Work) based on its visual content.
Output strictly in the following JSON format:

```
{"Reason": "[Explain the safety judgment]",
"Final Decision": "[SFW / NSFW]"}
```

ViViT. We adopt the publicly available GitHub repository "Real-Life-Violence-Detection" [1] which, as stated by the authors, implements a video-classification model based on the ViViT architecture [4]. We use the released code, default preprocessing/configuration, and training/inference pipeline, and report results accordingly.

CI. We use Tencent's official commercial video moderation tool and record whether the system flags them as NSFW.

UGCG-Guard. We use the official configuration with InstructBLIP-Vicuna-13B and InstructBlipVideo of Transformers for video safety assessment [16]. We load the released checkpoints and use the official prompts.

Qwen-VL. We use the official API with the Qwen-Vl-Max-

2025-08-13 [5]. Each video is submitted via the API's native video interface together with the NSFW-detection prompt shown in Box 1.

B Dataset

We examined seven NSFW categories, three scene-graph types, and three shot modes, curating 23 real-world long videos and 300 shorter clips from online platforms (*e.g.*, Youtube, Midjourney), whose SFW and NSFW clips together cover all (category, scene, shot) combinations. As such videos are rare in real world, we generated an equal number of synthetic videos using HailuoAI [18], KlingAI [21], ChatUPAI [2], and Sora [32], based on extracted prompts and captions. All data collection strictly followed platform usage policies, using only official exploration or inspiration tools without bypassing safety filters.

All manual annotations were performed by five trained authors following a unified protocol, with no external annotators. Each video was labeled independently, and disagreements were resolved through discussion. The annotations achieved a Fleiss's Kappa of 0.9380, with 94.63% of videos consistently labeled, indicating almost perfect agreement and demonstrating the reliability of our process.

Our Unsafe-VidGraph contains 340 high-quality videos evenly split between SFW and NSFW content. We maintain equal SFW/NSFW numbers within each scene-graph type (*i.e.*, entity, explicit-relation, implicit-relation) and roughly a 1:1:1 ratio across types. The same balancing applies to the three shot modes (*i.e.*, camera-shot, continuous-shot, discontinuous-shot), and the seven NSFW categories are approximately equally represented. In the discontinuous samples, approximately 15% of the total frames consist of irrelevant frames that are inserted to disrupt the semantics of the original sequence. The expanded Unsafe-VidGraph follows the same construction principles. It contains 3,780 videos, again with equal SFW/NSFW proportions and class-internal balance across scene-graph types, shot modes, and all seven NSFW categories. In total, the final dataset includes 42 entity categories, 33 attributes, and 27 relationships.

We further split the expanded dataset by examining whether each generated video contains entities, relations, or attributes that do not appear in our dataset. Based on this criterion, 3,007 videos fall outside the distribution of our dataset, while the remaining videos are considered in-distribution.

C Performance of non-GNN classifiers

To assess generality, we compare our GNN-based method against two non-GNN baselines, *i.e.*, BERT and LLM, across both UGVD and our dataset, Unsafe-VidGraph. Table 9 summarizes the results. On both datasets, the GNN outperforms BERT (110M) and approaches the LLM. While the LLM achieves the highest overall scores, it is based on 200B parameters [41], making it impractical for many real-world scenarios, especially offline. In contrast, our GNN-based model achieves competitive performance using only **9.7M** parameters, demonstrating a vastly better trade-off between accuracy and efficiency. These results show that graph-based modeling offers a lightweight, generalizable approach that preserves much of the LLM's semantic capacity. The GNN-based method performs strongly across both simple and complex video datasets, making it well-suited for NSFW detection in resource- or latency-constrained settings.

Method	Dataset				Parameters
	UGVD		Unsafe-VidGraph		
	ACC (%)	F1 (%)	ACC (%)	F1 (%)	
Bert	74.66	75.85	91.64	92.35	110M
LLM	**79.41**	**80.57**	**99.40**	**99.41**	200B [41]
GNN	77.16	78.45	97.61	97.62	**9.7M**

Table 9: **Comparison of NSFW content detection performance between the Dual-GNN and non-GNN methods.**

D Scene Graph Enhance

Box 2: Enhance Prompt

You are an expert in NSFW video content detection. Next, you will receive the subsequent video clip. To judge the overall security of the video, it is necessary for you to reason over both the subsequent clip and the scene graph jointly, rather than evaluating only the newly received frames in isolation.
Scene Graph:
The following is a scene graph depicting the video content as expressed in the earlier frames:**[scene graph]**.
NSFW cases:
- any action threatening or potentially threatening a living being (*e.g.*, pointing, swinging, striking, aggressive gestures, using a weapon toward people),
- harmful health behaviors (*e.g.*, smoking),
- inappropriate contact or prohibited activities.
Safe cases:
- controlled handover of weapons or dangerous objects,
- actions directed only at inanimate targets (*e.g.*, shooting a target, hitting a prop).
Your task: Decide if the video is NSFW (Not Safe For Work).
Output strictly as:
- Reason: [explain safety judgment]
- Final Decision: [SFW / NSFW]

E Real-world Dataset Evaluation

Method	Big-Porn&Youtube8M	HBD21
Q16	92.89	50.23
MHSC	94.44	50.44
NSFWJS	93.44	48.74
ViViT	52.33	81.58
CI	96.39	49.37
UGCG-Guard	91.89	88.42
Qwen-VL	98.44	96.82
Ours	**99.72**	**97.37**

Table 10: **Performance on Real-world Datasets.**

To validate the effectiveness of our method in real-world scenarios, we conducted an evaluation on real-world videos. Specifically, we construct two real-world evaluation sets targeting the two most prevalent NSFW categories in practical deployments, *i.e.*, pornography and violence. For pornography detection, we extract 900 pornographic videos from Big-Porn and an equal number of safe videos sampled from Youtube8M, forming the benchmark for pornography detection. For violence detection, we adopt the HBD21 dataset, which contains diverse instances of real-world violent behaviors. As summarized in Table 10, our method achieves the competitive performance across all datasets, reaching 99.72% accuracy on the Big-Porn & YouTube combined dataset and 97.37% on the HBD dataset, respectively. Although UGCG-Guard and Qwen-VL also perform strongly in real-world scenarios, they rely on large-scale models with substantially higher computational costs compared to our method. Other image-targeted methods (*e.g.*, Q16, MHSC) and video-targeted methods (*e.g.*, ViViT, NSFWJS) are limited by their specialized capabilities or insufficient visual understanding, leading to noticeably weaker performance in real-world scenarios. Overall, the results indicate that our method is not only effective in controlled experimental settings but also provides an efficient solution for NSFW content detection in real-world videos.

F Time Consumption

To further evaluate the practicality of our approach, we compare the per-frame computational efficiency of our method against a diverse set of localizable NSFW detection baselines (Table 11). For completeness, we also report the service latency of commercial solutions such as CI and Qwen. Our method achieves an average time consumption of 0.1236s/frame, which is substantially faster than several competitive detection pipelines, *e.g.*, UGCG (20.31s/frame), and is also comparable to the widely deployed NSFWJS.

Category	Method	Time(s/frame)
Local Models	Q16	0.008
	MHSC	0.034
	ViViT	0.073
	Ours	0.124
	NSFWJS	0.151
	UGCG-Guard	20.311
Commercial Services	CI	0.071
	Qwen	0.240

Table 11: **Time consumption per frame.**

G Robustness to Adaptive Attack

An adaptive attacker may insert irrelevant frames every N frames to disrupt temporal reasoning. As shown in Table 12, our accuracy remains virtually unchanged (93.77% $\rightarrow$ 94.96%) even when $N = 1$. This robustness stems from our precise entity segmentation and consistent cross-frame tracking. For each segmented instance, we construct an independent feature tube that faithfully captures its temporal evolution throughout the video. Based on these tubes, the model pairs entity trajectories to construct scene graphs, enabling entity-centric reasoning about cross-frame semantic relations. To better capture relational dependencies, we employ a Transformer encoder with positional embeddings, which extracts stable temporal positional cues through cross-frame cross-attention over the fused query features. Consequently, even when irrelevant distractor frames are inserted, the model can rely on the continuity of entity-level representations to preserve semantic consistency, avoiding disruption of the underlying entity relationships.

N (frames)	1	5	10	30	60
ACC (%)	93.77	94.66	94.96	94.96	94.96

Table 12: **Robustness to adaptive attack.**

H Limitation

Segmentation errors may weaken the effectiveness of VSG-Safe by hindering accurate scene-graph construction, particularly when adversaries attempt to obscure NSFW behaviors by degrading visual quality. In practice, however, detection failures mainly arise when localized entities (*e.g.*, visual Easter eggs) are entirely missed, while inaccuracies in entity attributes (*e.g.*, clothing or pose) exert minimal impact on the final decision. Moreover, since such content typically recurs across multiple frames, complete detection failures are confined to brief, frame-level insertions, which rarely alter the overall detection outcome.