VidLeaks: Membership Inference Attacks Against Text-to-Video Models

Li Wang[1,3,4], *Wenyu Chen*[1], *Ning Yu*[2], *Zheng Li*[1,3,4]*, *Shanqing Guo*[1,3,4]*

[1]*School of Cyber Science and Technology, Shandong University;* [2]*Eyeline Labs*
[3]*State Key Laboratory of Cryptography and Digital Economy Security, Shandong University*
[4]*Shandong Key Laboratory of Artificial Intelligence Security, Shandong University*

Abstract

The proliferation of powerful Text-to-Video (T2V) models, trained on massive web-scale datasets, raises urgent concerns about copyright and privacy violations. Membership inference attacks (MIAs) provide a principled tool for auditing such risks, yet existing techniques—designed for static data like images or text—fail to capture the spatio-temporal complexities of video generation. In particular, they overlook the sparsity of memorization signals in keyframes and the instability introduced by stochastic temporal dynamics.

In this paper, we conduct the first systematic study of MIAs against T2V models and introduce a novel framework *VidLeaks*, which probes sparse-temporal memorization through two complementary signals: 1) Spatial Reconstruction Fidelity (SRF), using a Top-K similarity to amplify spatial memorization signals from sparsely memorized keyframes, and 2) Temporal Generative Stability (TGS), which measures semantic consistency across multiple queries to capture temporal leakage. We evaluate *VidLeaks* under three progressively restrictive black-box settings—supervised, reference-based, and query-only. Experiments on three representative T2V models reveal severe vulnerabilities: *VidLeaks* achieves AUC of 82.92% on AnimateDiff and 97.01% on InstructVideo even in the strict query-only setting, posing a realistic and exploitable privacy risk. Our work provides the first concrete evidence that T2V models leak substantial membership information through both sparse and temporal memorization, establishing a foundation for auditing video generation systems and motivating the development of new defenses. Code is available at: https://zenodo.org/records/17972831.

1 Introduction

The advent of powerful Text-to-Video models (T2V), such as Sora [1], Kling [2], Luma [3], and Gen-3 [4], marks a new frontier in generative AI, enabling the creation of high-fidelity, dynamic video content directly from text prompts [5–11].

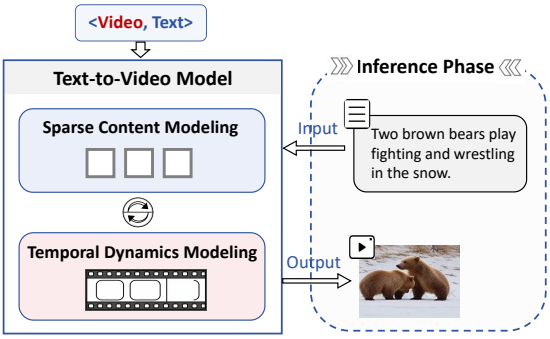


Figure 1: Illustration of the T2V generation process.

However, this capability critically depends on training with massive datasets, often containing billions of internet-scraped videos that inevitably include private and unauthorized data [12–15]. The use of such data has already sparked significant controversy [16]. For instance, in 2024, *YouTube*'s CEO publicly stated that training OpenAI's *Sora* with YouTube videos would "clearly violate" the platform's terms of service, while OpenAI's CTO declined to confirm whether such data had been used (reported by *Bloomberg*[1]). The controversy has also spread internationally — after OpenAI launched Sora in the UK, debates intensified over AI training on copyrighted works and artist rights (reported by *The Guardian*[2]). This tension is mirrored in the open-source community, where a large video dataset curated for the Open-Sora project was removed from Hugging Face following a DMCA takedown request by the stock footage platform Pexels[3]. Therefore, significant concerns have been raised regarding such data risks, motivating us to address a fundamental question, "Is a given video used

*Corresponding authors.

[1] https://www.bloomberg.com/news/articles/2024-04-04/youtube-says-openai-training-sora-with-its-videos-would-break-the-rules

[2] https://www.theguardian.com/technology/2025/feb/28/openai-sora-video-generation-uk-amid-copyright-row?utm_source=chatgpt.com

[3] https://huggingface.co/datasets/hpcai-tech/open-sora-pexels-45k

in training a T2V model?"

In this paper, we take the first step toward studying data risks of T2V models through membership inference attacks. The goal is to infer whether a given video is part of a T2V model's training dataset. While MIAs have been widely explored for classifiers [17, 18] and more recently for large language models (LLMs) [19] and text-to-image (T2I) models [20], extending them to T2V models introduces fundamentally new challenges. As illustrated in Figure 1, T2V models are trained on video–text pairs to generate temporally coherent video sequences [21–23]. To this end, they adopt specialized mechanisms such as spatial-temporal sparse attention [9, 24] to compress redundant frames, preserve spatial fidelity, and stabilize temporal dynamics. While these designs improve generation quality, they also complicate membership inference, creating unique attack surfaces beyond those in text- or image-based MIAs:

- **Challenge ❶: Sparsity of Content Memorization.** Video data is highly redundant, with many visually similar frames. To train efficiently, T2V models selectively memorize only sparse, informative anchors (e.g., keyframes [25, 26]). This sparsity weakens membership signals, as the few memorized anchors are easily overwhelmed by the noise from generalized, non-memorized frames. Consequently, naive frame-wise similarity fails because it averages over all frames, drowning out the sparse memorization signal (see Section 8.4).

- **Challenge ❷: Dynamics of Temporal Memorization.** Beyond static appearance, videos encode motion over time. T2V models thus learn temporal dynamics in addition to spatial content [21, 27]. However, they generate stochastic motion textures rather than fixed trajectories. Natural motion variation introduces substantial pixel-level noise that overwhelms the subtle temporal cues distinguishing members from non-members. As a result, pixel-level tools such as optical flow [28] fail to isolate reliable membership signals under this variability (see Section 8.3).

To address these challenges, we propose a novel sparse-temporal MIA framework, *VidLeaks*, tailored to T2V models, which focuses on: (1) the fidelity of reconstructing sparsely memorized key content and (2) the stability of its temporal dynamics. Concretely, we design two complementary signals:

- **Signal ❶: Sparse Reconstruction Fidelity (SRF).** To overcome sparsity, SRF evaluates whether the model memorizes key content by introducing a *Top-K Reconstruction Fidelity* metric. This metric compares generated frames against the most relevant keyframes of the target video, effectively acting as a matched filter that amplifies weak memorization signals otherwise obscured by frame redundancy.

Table 1: Adversary's knowledge under three threat models.

Threat Model	Member Data		Non-Member Data	
	Video	Text	Video	Text
Supervised	✓	×	✓	×
Reference-based	×	×	✓	×
Query-only	×	×	×	×

- **Signal ❷: Temporal Generative Stability (TGS).** To address temporal dynamics, TGS measures the stability of scene-level semantics across repeated generations via a *Multi-Q Generative Stability* metric. This metric captures semantic consistency across generated frames over multiple queries, providing a reliable probe into whether the model has memorized temporal patterns of a video.

Building on SRF and TGS, we formalize our *VidLeaks* and systematically evaluate it under three progressively restrictive black-box threat models (see Table 1). Given a target video, an adversary first leverages a public video captioning model to obtain a surrogate text, which is then used to query the target T2V model. The generated videos are analyzed with SRF and TGS, and the results are fed into an inference module to determine membership. Depending on the adversary's capability, the module is instantiated in three scenarios (formally defined in Section 4.1): (1) a trained classifier in the **supervised** setting, (2) a statistical anomaly detector in the **reference-based** setting, or (3) an unsupervised fusion in the **query-only** setting.

We conduct extensive experiments on three representative T2V models: AnimateDiff [11], Mira [29], and InstructVideo [25], covering diverse architectural paradigms. Empirical evaluations show that *VidLeaks* achieves strong performance across all settings, achieving AUC of 82.92% on AnimateDiff and 97.01% on InstructVideo, even under the most restrictive query-only scenario. This demonstrates that T2V models inevitably leak membership information through both sparse reconstruction fidelity and temporal generative stability. In summary, our main contributions are as follows:

- We conduct the first comprehensive investigation of membership inference attacks against text-to-video models. We identify and formalize two domain-specific challenges: the sparsity of content memorization and the dynamics of temporal memorization.

- We propose a novel MIA framework, *VidLeaks*, introducing two complementary signals—Sparse Reconstruction Fidelity (SRF) and Temporal Generative Stability (TGS). These signals expose membership leakage by targeting keyframe fidelity and temporal stability, which conventional holistic or motion-based methods fail to capture.

- We design an attack pipeline applicable under progressively restrictive black-box threat models, starting only

from a target video without its ground-truth caption. Extensive experiments on three representative T2V models reveal severe vulnerabilities: *VidLeaks* achieves strong performance across all scenarios, with AUC of 82.92% on AnimateDiff and 97.01% on InstructVideo, even in the most restrictive query-only setting.

2 Background & Related Work

2.1 Text-to-Video Generation Models

Text-to-Video (T2V) generation aims to synthesize temporally coherent videos from textual prompts [5,10,25,30,31]. Recent breakthroughs are largely driven by diffusion models [32,33], particularly in latent space [34], where a denoiser is trained to iteratively recover clean video representations from noisy inputs conditioned on text [11,35].

Extending diffusion from static images to the spatio-temporal domain of video has given rise to three representative paradigms [22, 36, 37]: *1) T2I adaptation with motion modules.* This paradigm builds upon powerful text-to-image backbones by freezing spatial layers and inserting lightweight temporal modules (e.g., temporal attention) for motion modeling [38]. AnimateDiff [11] is a canonical example, explicitly decoupling spatial appearance from temporal motion modeling. *2) End-to-end spatio-temporal training.* Instead of reusing T2I backbones, these models train large spatio-temporal transformers directly on web-scale video–text datasets [13–15]. Systems such as Mira [29], OpenSora [30], and CogVideoX [8] belong to this category. They often employ factorized spatio-temporal attention to jointly capture visual and motion representations. *3) Reward- or instruction-based fine-tuning.* Inspired by alignment techniques in LLMs [10], this line of work fine-tunes pre-trained T2V models with reward signals or human feedback. InstructVideo [25] is a notable example, using image-based reward models to enhance visual appeal and text alignment.

These paradigms—adaptation, end-to-end training, and fine-tuning—capture the major strategies in today's T2V landscape. For our study, we select AnimateDiff, Mira, and InstructVideo as representative and auditable cases. Unlike recent closed-source systems such as CogVideoX [8], HunyuanVideo [39], and Wan [40], whose training datasets are proprietary or undisclosed, our chosen models provide transparent data provenance, which is essential for verifiable membership inference studies. Importantly, despite their architectural differences, all share the common goal of modeling both static spatial fidelity and dynamic temporal evolution [21,27,41], creating the attack surfaces we exploit.

2.2 Membership Inference Attacks

Membership Inference Attacks (MIAs) aim to determine whether a given sample was part of a model's training set [42]. Early studies focused on classification models, exploiting the observation that models typically yield higher confidence (or lower loss) on member samples they were trained on [17,18]. The scope later has broadened to generative models, such as GANs [43,44] and diffusion models [45]. For GANs, attacks exploited signals from the discriminator's output or the generator's ability to reconstruct a sample, as explored in GAN-Leaks [46]. For diffusion models, particularly in the T2I domain, researchers have focused on the reconstruction error between queried images and their generated counterparts [47]. These works established that reconstruction fidelity is a viable, albeit sometimes noisy, signal for membership.

With the rise of large-scale models, MIAs have faced new challenges. For LLMs [48,49] and VLMs [50,51], which are often accessed only via black-box APIs, traditional loss-based signals are unavailable. This has spurred the development of "label-only" attacks that rely on model outputs alone [52,53]. These attacks often probe for subtle behavioral differences, such as a model's consistency, robustness, or sensitivity to specific parameters [19,20]. This evolution towards analyzing subtle behavioral artifacts inspires our design for T2V models.

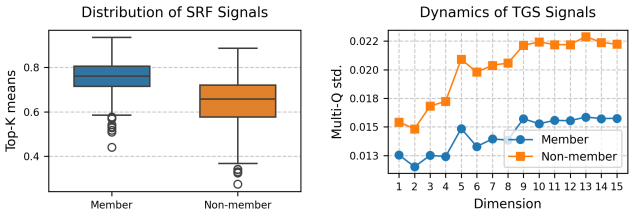
Despite these advances, the vulnerability of modern T2V models to MIAs has not been systematically studied. Prior works against T2I models are insufficient as they neglect the temporal dimension [21,27], while techniques for other modalities fail to capture the spatio-temporal generative process of video. To our knowledge, this work is the first to bridge this gap by systematically studying MIAs against T2V systems across diverse architectural paradigms.

3 Key Insights & Signal Design

To design an effective MIA on T2V models, we first revisit how these models memorize training data and identify signals that can expose such memorization. As discussed in Section 2.1, T2V models jointly optimize two objectives: (1) generating high-fidelity visual content within frames and (2) maintaining coherent motion across frames. These dual objectives imply two complementary perspectives of memorization: (1) fidelity of visual details in key frames and (2) stability of temporal evolution. We now develop concrete signals that capture both aspects.

3.1 Sparse Memorization: Reconstruction Fidelity

A direct approach to detecting memorization is to compare a target video with one generated from its text prompt. However, naive frame-wise similarity is ineffective (see Section 8.4), since most frames contain redundant content that obscures weak memorization signals. T2V models are more likely to memorize distinctive *key anchors*—salient keyframes or regions—rather than entire sequences.



(a) (b)

Figure 2: Differences between members and non-members under SRF and TGS signals. (a) Distribution of SRF scores computed from Top-K similarities. (b) Per-dimension TGS instability across repeated generations.

Signal ❶: Sparse Reconstruction Fidelity (SRF). We design SRF to focus on these key anchors. We first extract keyframes from the target video using FFmpeg's standard frame-selection tools [54], which provides a reproducible set of structural anchors in the video. For each generated frame, we compute its CLIP [55] similarity with all extracted keyframes of the target video and average the Top-K scores. This acts as a matched filter that emphasizes memorized anchors while ignoring redundant frames. The overall SRF score is then averaged across generated frames. A higher SRF indicates stronger memorization of member videos. As shown in Figure 2a, SRF produces a clear distributional shift between members and non-members, validating its effectiveness in probing sparse memorization.

Formally, let v_t be a target video with M keyframes $\mathcal{F}_{keys} = \{\mathbf{f}_1, \ldots, \mathbf{f}_M\}$, and $\tilde{v}_g$ be the generated video with N frames $\tilde{\mathcal{F}} = \{\tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N\}$. For each generated frame $\tilde{\mathbf{f}}_i$, the SRF score is:

$$SRF_i = \frac{1}{K} \sum_{j=1}^{K} \max_{k \in \{1..M\}}^{(j)} \cos(\tilde{\mathbf{f}}_i, \mathbf{f}_k) \tag{1}$$

where $\max^{(j)}$ denotes the j-th largest value, and $\cos(\cdot,\cdot)$ denotes the cosine similarity.

The overall SRF score is $S_{SRF} = \frac{1}{N} \sum_{i=1}^{N} SRF_i$. In addition to this scalar, the full SRF vector $[SRF_1, \ldots, SRF_N]$ can be used as a high-dimensional feature for supervised attacks.

3.2 Temporal Memorization: Generative Stability

The second perspective of memorization lies in motion patterns. However, measuring motion at the pixel level (e.g., optical flow) is unreliable (see Section 8.3), since T2V models learn stochastic motion textures rather than exact trajectories. Instead, the key lies in whether the model reproduces the same *scene-level dynamics* consistently across generations.

Signal ❷: Temporal Generative Stability (TGS). We propose TGS, which measures the stability of semantic scene evolution under repeated queries. For each generation, we first compute a frame-wise *consistency vector* that captures background stability. By sampling the same text prompt Q times, we then obtain Q such vectors and measure per-dimension standard deviations. Member videos exhibit lower instability, reflecting stronger memorization. As shown in Figure 2b, member videos exhibit consistently lower instability than non-members, establishing TGS as a robust probe of temporal memorization.

We follow the formulation of background consistency from [26]. For a generated video with CLIP features $\{\tilde{\mathbf{f}}_0, \ldots, \tilde{\mathbf{f}}_{N-1}\}$, the consistency score for frame $i > 0$ is:

$$C_i = \tfrac{1}{2} \left(\cos(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_{i-1}) + \cos(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_0) \right) \tag{2}$$

This yields a consistency vector $\mathcal{V} = [C_1, \ldots, C_{N-1}]$. Across Q generations, we compute the instability vector $\mathbf{s}_{instab} \in \mathbb{R}^{N-1}$, where each element is:

$$\mathbf{s}_{instab}[j] = \mathrm{StdDev}\left(\mathcal{V}_1[j], \ldots, \mathcal{V}_Q[j] \right) \tag{3}$$

The final TGS score is $S_{TGS} = \frac{1}{N-1} \sum_{j=1}^{N-1} \mathbf{s}_{instab}[j]$. Lower S_{TGS} (higher stability) implies stronger memorization.

Remark

SRF captures whether the model memorizes salient visual anchors, while TGS measures whether it memorizes stable temporal evolution. Together, they provide complementary sparse-temporal signals that serve as the foundation of our attack.

4 The Sparse-Temporal MIA Framework

Based on the insights from our initial explorations, we now formalize our sparse-temporal membership inference attack (MIA) framework against text-to-video (T2V) models. This section first defines the threat models and then presents the overall attack pipeline.

4.1 Threat Models

We consider a black-box adversary who aims to determine whether a given **target video** v_t was included in the training set of a target T2V model $\mathcal{M}$. The adversary can query $\mathcal{M}$ with text prompts and observe the generated videos, but has no access to $\mathcal{M}$'s internal parameters, architecture, or training data. Critically, we assume the adversary does *not* possess the ground-truth prompt paired with v_t during training, as real-world auditing scenarios (e.g., creators verifying unauthorized use of their content) typically provide access only to the video itself rather than the original video–text pair. Following established MIA paradigms [20], we formalize three progressively restrictive threat models:

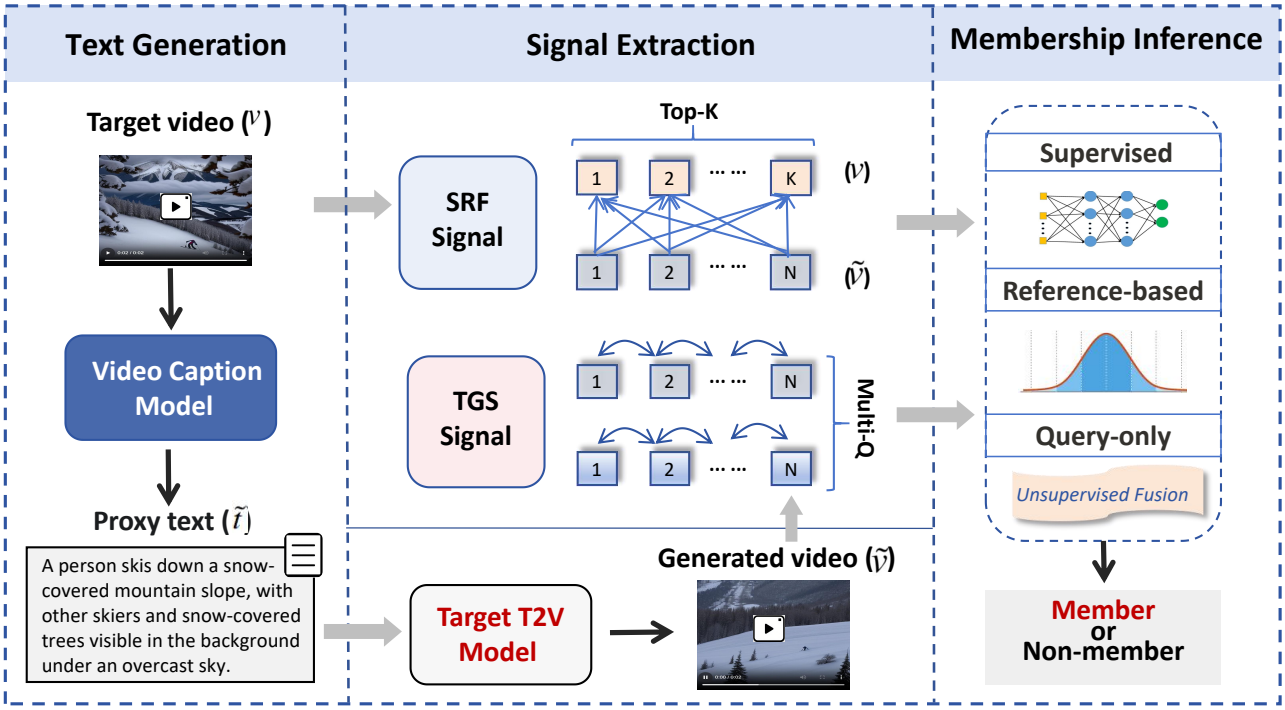


Figure 3: Overview of our sparse-temporal MIA framework. The attack begins with only a target video. A public video captioning tool generates a proxy text, which is then used to query the T2V model. The signal extraction stage computes SRF and TGS from the original and generated videos. Finally, the membership inference module, instantiated according to the threat model, outputs the final membership decision.

- **Supervised Inference.** This setting establishes a theoretical upper bound on the attack performance. The adversary is assumed to have a *shadow dataset* $\mathcal{D}_{shadow}$ containing samples explicitly labeled as members or non-members of the target model's training set.
- **Reference-based Inference.** In this more realistic model, the adversary no longer has access to labeled member samples but possesses a *reference set* $\mathcal{D}_{ref}$ of confirmed non-member samples.
- **Query-only Inference.** This is the most restrictive and practical model. The adversary operates under a *zero-knowledge* assumption, having neither a shadow nor a reference dataset. The inference must be made solely based on the query results of the target video itself.

4.2 Attack Framework Overview

Our attack pipeline, illustrated in Figure 3, is a modular framework consisting of three main steps: Text Generation, Signal Extraction, and Membership Inference.

- **Step ❶ – Text Generation.** The attack realistically begins with only a target video v. The adversary first feeds v into a publicly available video captioning model, $\mathcal{C}$, to generate a descriptive text, $\tilde{t} = \mathcal{C}(v)$. This proxy text serves as the input for querying the target T2V model. This step makes our attack highly practical as it removes the often unrealistic assumption of knowing the ground-truth prompt (see Section 8.5 for the impact of caption source and quality).
- **Step ❷ – Signal Extraction.** Using the generated prompt $\tilde{t}$, the adversary queries the target model to obtain a generation video $\tilde{v} = \mathcal{M}(\tilde{t})$. This stage then computes our two core signals based on the original video v and the generated video $\tilde{v}$: the Sparse Reconstruction Fidelity (SRF) and the Temporal Generative Stability (TGS).
- **Step ❸ – Membership Inference.** The extracted signals are then passed to a membership inference module, which outputs a final membership decision. The implementation of this module depends on the threat models. It can be instantiated as a trained classifier in the supervised scenario, a statistical anomaly scorer in the reference-based scenario, or an unsupervised fusion in the query-only scenario. The specific implementations for each instantiation are detailed in Section 5, Section 6, and Section 7.

5 Supervised Inference

5.1 Assumptions and Intuition

We begin with the **Supervised Inference** threat model, which provides a theoretical upper bound on potential information leakage. In this setting, the adversary is assumed to have a shadow dataset $\mathcal{D}_{\text{shadow}} = \{(v_i, \tilde{t}_i, y_i)\}_{i=1}^{N_s}$, where each sample is labeled as a member ($y_i = 1$) or non-member ($y_i = 0$) of the target model's training set. The key intuition is that a supervised classifier, trained on this labeled data, can fully exploit the high-dimensional sparse-temporal signals to learn discriminative boundaries between members and non-members.

5.2 Attack Implementation

The supervised attack leverages both SRF and TGS signals to train a classifier that can then infer the membership of target samples. The procedure consists of two stages: feature construction, and classifier training and inference (see Algorithm 1 in Appendix A).

Feature Construction. To capture the rich information, we represent each sample by its full underlying signal vectors rather than scalar scores (S_{SRF}, S_{TGS}) defined in Section 3. For each sample $(v_i, \tilde{t}_i)$ in the shadow dataset, we extract:

- **SRF vector:** $\mathbf{v}_{srf} = [SRF_1, \ldots, SRF_N] \in \mathbb{R}^N$, where each element is the Top-K fidelity of a generated frame (Equation 1).
- **TGS vector:** $\mathbf{s}_{instab} \in \mathbb{R}^{N-1}$, where each element is the standard deviation of temporal consistency score at frame index j across Q generations (Equation 3).

These vectors are then concatenated to form a single, high-dimensional feature vector $\mathbf{x} \in \mathbb{R}^{2N-1}$ for each sample. This representation preserves both spatial and temporal patterns essential for distinguishing members from non-members.

Classifier Training and Inference. Using the labeled feature vectors $\{\mathbf{x}_i, y_i\}_{i=1}^{N_s}$, the adversary trains a Multi-Layer Perceptron (MLP) to serve as the attack model, $\mathcal{A}_\theta$. The MLP is trained to minimize the loss between its predictions and the ground-truth labels, thereby learning the discriminative patterns between member and non-member samples. Once trained, $\mathcal{A}_\theta$ can process the feature vector $\mathbf{x}$ of any target video and outputs a membership probability $p = \mathcal{A}_\theta(\mathbf{x}) \in [0, 1]$, which is used to decide whether the video belongs to the model's training set.

5.3 Experimental Setting

Target T2V Models. To ensure coverage of the main T2V paradigms introduced in Section 2.1, we evaluate three representative open-source models:

- **AnimateDiff** [11]: a canonical T2I-adaptation model that freezes image backbones and inserts lightweight motion modules.
- **Mira** [29]: an end-to-end spatio-temporal transformer trained on the curated MiraData dataset. Mira (Mini-Sora) represents an initial foray into high-quality, long-duration video generation in the style of Sora.
- **InstructVideo** [25]: a reward-fine-tuned model aligned with human preference signals.

These models jointly cover the adaptation, end-to-end, and alignment paradigms of current T2V systems. All three models have publicly documented training corpora, allowing us to reliably determine membership for rigorous MIA evaluation.

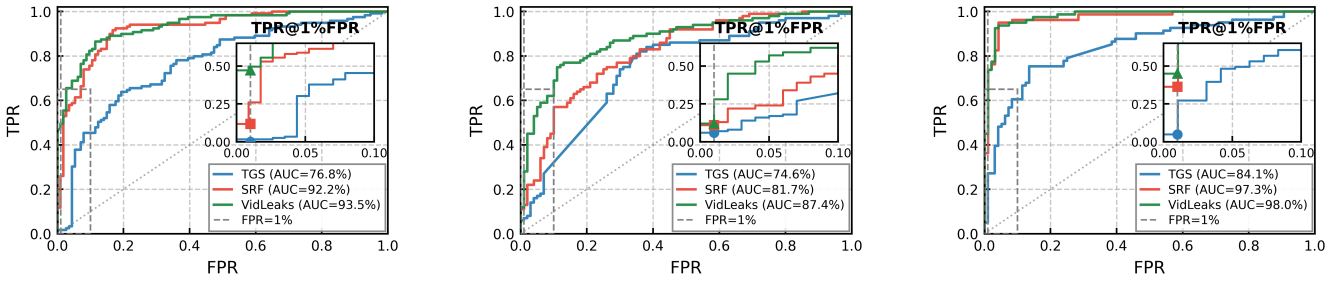
Datasets. Member samples are drawn directly from the publicly documented training sources of each target model: WebVid-10M [13] for AnimateDiff and InstructVideo, and MiraData [15] for Mira, ensuring unambiguous membership labels. Non-member samples are taken from Panda-70M [12], a high-quality dataset disjoint from all training sources, and are randomly sampled for evaluation. In the supervised scenario, the shadow dataset for each T2V model contains approximately 500 member and 500 non-member videos. All datasets are stratified and split 8:2 into training and testing sets.

Video Caption Model. Although ground-truth text prompts are available in the training datasets, we adopt a more realistic setting where target videos may not always come with captions. To query target models, we therefore employ a proxy captioner (Gemini Pro via Google AI Studio [56]) to generate descriptive prompts. Crucially, *VidLeaks* maintains strong performance across diverse caption sources and qualities (see Section 8.5), ensuring our pipeline's practicality when ground-truth or high-quality captions are unavailable.

Attack Model and Training. For the supervised setting, the attack model is a MLP as defined in Section 5.2, with two hidden layers using ReLU activations and dropout. We train the model with binary cross-entropy loss and early stopping based on validation AUC.

Evaluation Metrics. We evaluate attack performance using three standard metrics:

- **AUC (Area Under ROC Curve)**, a threshold-free measure of separability;
- **Balanced Accuracy**, which accounts for balanced performance across classes;
- **TPR@1%FPR**, the true positive rate at a 1% false positive rate.



(a) AnimateDiff (b) Mira (c) InstructVideo

Figure 4: ROC curves for the supervised attack on different T2V models.

Table 2: Performance of the supervised attack across different T2V models.

Target Model	Method	AUC(↑)	TPR@1%FPR(↑)	ACC(↑)
AnimateDiff	SRF	92.19%	26.89%	87.46%
	TGS	76.75%	2.52%	72.03%
	VidLeaks	**93.46%**	**49.58%**	**87.58%**
Mira	SRF	81.73%	11.00%	74.50%
	TGS	74.64%	6.00%	73.00%
	VidLeaks	**87.45%**	**12.00%**	**82.00%**
InstructVideo	SRF	97.29%	36.25%	94.44%
	TGS	84.07%	4.94%	80.88%
	VidLeaks	**98.04%**	**45.00%**	**94.77%**

We emphasize TPR@1%FPR, following the paradigm established by Carlini et al. [57] and adopted in recent studies [19], as it is particularly crucial for quantifying the practical risk of high-confidence membership inference attacks.

5.4 Experimental Results

Overall Performance. Table 2 reports the results of our supervised attack across three representative T2V models. By jointly leveraging SRF and TGS vectors, our method achieves consistently strong performance, establishing a theoretical upper bound on membership leakage. On AnimateDiff, the attack reaches 93.46% AUC and nearly 50% TPR@1%FPR, showing that almost half of the member videos can be identified with extremely high confidence at only 1% false positives. Even on Mira, which is trained end-to-end and thus considered harder to attack, our method achieves 87.45% AUC, significantly above chance. InstructVideo appears especially vulnerable, with 98.04% AUC and 45% TPR@1%FPR, underscoring that preference-aligned fine-tuning can substantially amplify memorization risks. These results highlight the generality of our attack and its ability to extract sensitive membership information across diverse T2V architectures.

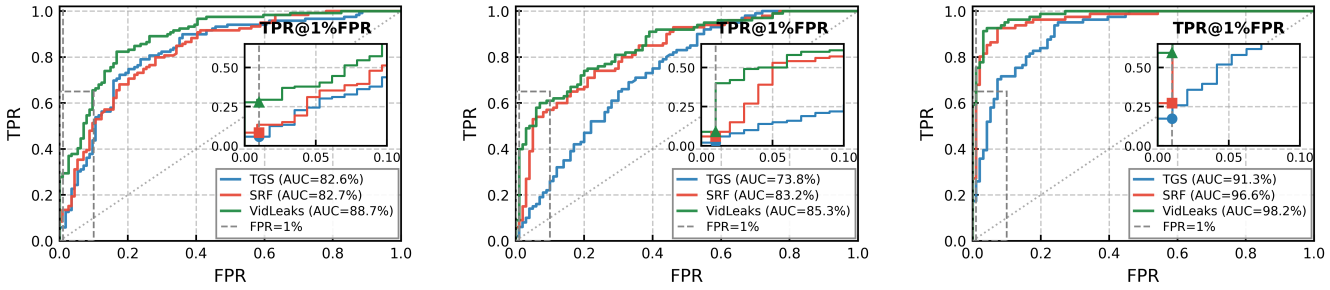
Contribution of Each Signal. The ablation results in Table 2 further validate our key insight: both SRF and TGS independently expose membership leakage. SRF achieves up to 97.29% AUC (InstructVideo), reflecting the model's tendency to memorize key visual anchors. TGS also provides non-trivial predictive power (e.g., 84.07% AUC on InstructVideo), confirming that temporal stability encodes valuable membership cues despite being a noisier signal. Crucially, fusing SRF and TGS consistently improves performance across all models, boosting AUC by 1–6% and substantially increasing TPR@1%FPR. This demonstrates that the two signals capture complementary aspects of memorization, sparse spatial fidelity and temporal stability, that a supervised classifier can effectively combine to form a robust decision boundary. The ROC curves in Figure 4 corroborate this, with the fused model's curve dominating in the low-FPR region, where practical attacks are most relevant.

6 Reference-based Inference

6.1 Assumptions and Intuition

We relax the strong assumption of a labeled shadow dataset and consider a more realistic threat model: **Reference-Based Inference**. This scenario assumes an adversary who, while lacking access to verified member samples, possesses a confirmed *non-member reference set*, $\mathcal{D}_{\text{ref}} = \{(v_i, \tilde{t}_i)\}_{i=1}^{N_r}$. This setting is practical and realistic, since such non-member data can often be readily obtained (e.g., videos created after the target T2V model's knowledge cutoff date), whereas a labeled member set is typically inaccessible.

The key intuition of this attack is to use a reference set to establish a statistical baseline for non-member behavior. Membership is then inferred by measuring how much a target sample's signals deviate from this baseline. A significant deviation from the non-member distribution serves as strong evidence for membership. We formalize this as a statistical anomaly detection problem, where non-members form the distribution of normal data, and members are treated as out-of-distribution anomalies.



(a) AnimateDiff (b) Mira (c) InstructVideo

Figure 5: ROC curves for the reference-based attack on different T2V models.

6.2 Attack Implementation

The reference-based attack replaces the supervised classifier with a statistical anomaly scoring mechanism. The procedure consists of three stages: calibration on the reference set, anomaly scoring for the target sample, and signal fusion and inference (see Algorithm 2 in Appendix A).

Calibration on the Reference Set. The first step is to establish a statistical baseline for non-member behavior. The adversary computes the scalar SRF score and TGS score for all samples in the non-member reference set $\mathcal{D}_{\text{ref}}$. From these two sets of scores, they compute the mean (μ_{srf}, μ_{tgs}) and standard deviation ($\sigma_{srf}, \sigma_{tgs}$), characterizing the "normal" range for each signal.

Anomaly Scoring. For a target video, the adversary extracts its scalar S_{SRF} and S_{TGS} scores. These raw scores are then transformed into normalized anomaly scores using the statistics derived from the reference set. We employ the Z-score for this normalization, which measures how many standard deviations a score is from the non-member mean.

For the S_{SRF} score, where higher values indicate membership, the anomaly score $\mathcal{A}_{SRF}$ is its Z-score:

$$\mathcal{A}_{SRF} = \frac{S_{SRF} - \mu_{srf}}{\sigma_{srf}} \tag{4}$$

Conversely, for the S_{TGS} score, where lower values (higher stability) indicate membership, the anomaly score $\mathcal{A}_{TGS}$ is its negative Z-score to align the directionality:

$$\mathcal{A}_{TGS} = -\frac{S_{TGS} - \mu_{tgs}}{\sigma_{tgs}} \tag{5}$$

For both anomaly scores, a larger positive value signifies a greater deviation from non-member behavior and thus a higher likelihood of membership.

Signal Fusion and Inference. As SRF and TGS signals capture complementary aspects of memorization (Section 3), we combine them into a single robust membership score via linear fusion:

$$\mathcal{S}_{final} = w_{srf} \cdot \mathcal{A}_{SRF} + w_{tgs} \cdot \mathcal{A}_{TGS} \tag{6}$$

where w_{srf} and w_{tgs} are weighting parameters. This final score $\mathcal{S}_{final}$ is then used to infer membership, typically by comparing it against a decision threshold.

Table 3: Performance of the reference-based attack across different T2V models.

Target Model	Method	AUC(↑)	TPR@1%FPR(↑)	ACC(↑)
AnimateDiff	SRF	82.70%	8.40%	75.88%
	TGS	82.63%	5.88%	77.65%
	VidLeaks	**88.68%**	**27.73%**	**82.84%**
Mira	SRF	83.15%	6.00%	75.50%
	TGS	73.75%	2.00%	68.50%
	VidLeaks	**85.34%**	**9.00%**	**77.00%**
InstructVideo	SRF	96.62%	27.16%	92.65%
	TGS	91.31%	17.28%	85.03%
	VidLeaks	**98.17%**	**59.26%**	**94.21%**

6.3 Experimental Setting

We evaluate the reference-based attack on the same target models, datasets, and proxy captioning setup as in Section 5, and report results with the same metrics (AUC, Balanced Accuracy, and TPR@1%FPR). The key difference lies in the attack configuration: instead of a labeled shadow dataset, the adversary is given only a disjoint non-member reference set. For each model, 80% of the non-member pool is used to construct $\mathcal{D}_{ref}$ (for calibrating the SRF and TGS score distributions), while the remaining 20% is reserved for evaluation alongside the member samples. For signal fusion (Equation 6), we adopt a conservative scheme with approximately equal weights ($w_{srf} \approx w_{tgs}$), avoiding hyperparameter tuning and providing a clean baseline for assessing the effectiveness of our fusion strategy.

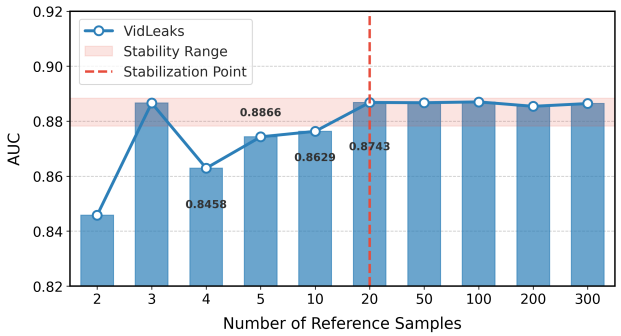


Figure 6: Impact of reference set size on attack performance.

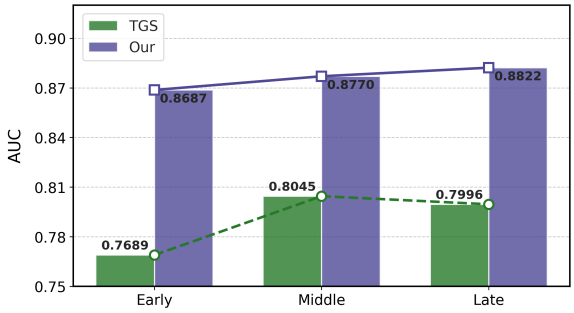


Figure 7: Analysis of TGS across temporal dimensions.

6.4 Experimental Results

Table 3 reports the results of the reference-based attack. Despite the absence of labeled member data, our fused method remains highly effective. On AnimateDiff, it achieves 88.68% AUC and 27.73% TPR@1%FPR, showing that a non-member reference baseline alone suffices to reveal strong membership leakage. Similar patterns hold on Mira (85.34% AUC) and InstructVideo (98.17% AUC), with the latter reaching nearly 60% TPR@1%FPR. While overall performance is naturally below the supervised upper bound, these results confirm that significant privacy risks persist even under this more realistic threat model.

Both SRF and TGS signals provide predictive power individually, with SRF generally performing better, but the fused approach consistently delivers the strongest results—raising AUC by up to 2–6% and substantially improving TPR@1%FPR. The ROC curves in Figure 5 further validate this trend, showing that even with only a non-member baseline, the fused score maintains a clear advantage over single-signal methods, particularly in the critical low-FPR region. This demonstrates that our *VidLeaks*, which jointly exploits SRF and TGS, provides strong evidence of memorization even when inference relies solely on non-member statistics.

Impact of Reference Set Size. We further examine how the size of the non-member reference set affects the attack performance. Figure 6 shows the AUC on AnimateDiff when varying the reference set from 2 to 300 samples. The attack already achieves an AUC of 88.66% with only 3 samples, and performance quickly stabilizes once the reference set reaches around 20 samples, maintaining AUC above 88.5% thereafter.

These results demonstrate the data efficiency of our approach: only a handful of non-member videos are sufficient to construct a stable baseline for anomaly scoring. Consequently, the reference-based attack remains highly practical, as an adversary does not need access to a large dataset to reliably expose membership leakage.

Temporal Dynamics of TGS Signal. To validate our TGS design, we analyze how its discriminative power evolves across the video timeline. For a generated video with N frames, the $(N-1)$-dimensional instability vector (Equation 3) is uniformly divided into three segments—Early, Middle, and Late—and segment-wise TGS scores are computed for evaluation.

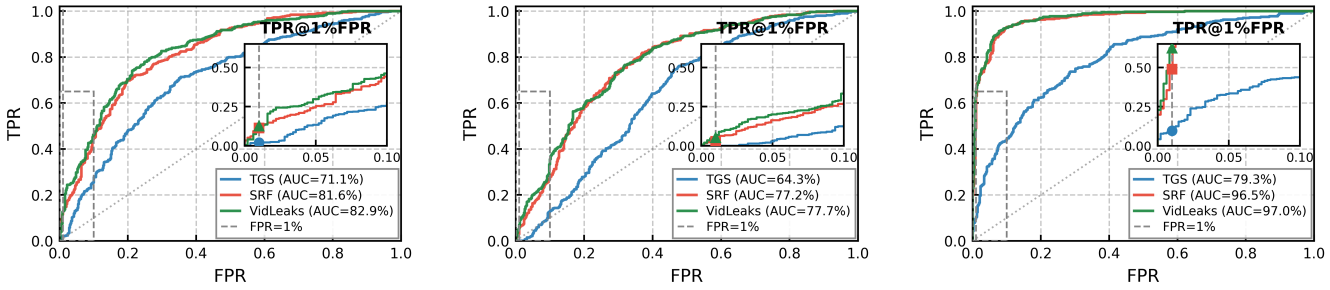
As shown in Figure 7, the membership signal tends to be weaker in early frames and becomes progressively stronger in the middle and late stages. This is consistent with our earlier visualization (Figure 2b), where the instability gap between members and non-members widens as generation unfolds. These results empirically justify our per-dimension variance formulation. A naive "mean-then-std" alternative collapses temporal information, obscuring the stronger signals present in later frames. Direct comparison confirms this: our per-dimension design achieves an AUC of 82.63%, significantly outperforming the 80.86% of "mean-then-std". By preserving variance at each temporal step, TGS better captures the instability dynamics that expose membership leakage.

7 Query-only Inference

7.1 Intuition and Assumptions

We consider the most restrictive yet practical threat model: **Query-Only Inference**. In this setting, the adversary operates under the zero-knowledge assumption, possessing neither a shadow dataset for training nor a reference set for calibration. Inference must be made solely from the signals derived for the target video, without any external calibration. This scenario simulates a real-world attacker with only black-box query access to the T2V model and the target video.

The key intuition is that our sparse-temporal signals are intrinsically discriminative. Although the global signal distributions are unknown, member samples are expected to exhibit both high reconstruction fidelity and high generative stability—a combination statistically rare among non-members. The attack therefore relies on unsupervised scoring and fusion of the two signals to infer membership.



(a) AnimateDiff (b) Mira (c) InstructVideo

Figure 8: ROC curves for the query-only attack on different T2V models.

7.2 Attack Implementation

The query-only attack removes the need for calibration and relies solely on intrinsic scoring and unsupervised fusion of signals. The procedure consists of two stages: intrinsic signal scoring, and unsupervised fusion and inference (see Algorithm 3 in Appendix A).

Intrinsic Signal Scoring. Without a reference distribution, we define intrinsic scores based on the inherent properties of the raw signals, following a unified "higher is more member-like" convention:

- **SRF scoring:** S_{SRF} already aligns with this convention, as higher values indicate membership. We thus use its raw value as the intrinsic score:

$$\mathcal{S}_{SRF} = S_{SRF} \tag{7}$$

- **TGS scoring:** Since S_{TGS} measures instability, lower values indicate membership. To align the directionality, we transform it into a stability score by inverting its value:

$$\mathcal{S}_{TGS} = 1 - S_{TGS} \tag{8}$$

Unsupervised Fusion and Inference. With the two intrinsic scores directionally aligned, we fuse them into a single membership score. Lacking data to learn optimal weights, we employ a pre-defined linear combination based on general heuristics about the relative importance of each signal:

$$\mathcal{S}_{final} = w_{SRF} \cdot \mathcal{S}_{SRF} + w_{TGS} \cdot \mathcal{S}_{TGS} \tag{9}$$

where w_{SRF} and w_{TGS} are pre-defined weights. This final score enables the adversary to rank candidate videos by membership likelihood. We evaluate $\mathcal{S}_{final}$ primarily with threshold-free metrics such as AUC, and report balanced accuracy without data-dependent calibration.

Table 4: Performance of the query-only attack on different T2V models.

Target Model	Method	AUC(↑)	TPR@1%FPR(↑)	ACC(↑)
AnimateDiff	SRF	81.61%	11.40%	75.00%
	TGS	71.07%	2.00%	68.00%
	VidLeaks	**82.92%**	**12.60%**	**76.60%**
Mira	SRF	77.18%	2.86%	72.45%
	TGS	64.27%	1.00%	64.18%
	VidLeaks	**77.66%**	**5.10%**	72.55%
InstructVideo	SRF	96.54%	48.88%	91.52%
	TGS	79.32%	9.68%	72.55%
	VidLeaks	**97.01%**	**62.28%**	**91.80%**

7.3 Experimental Setting

We again adopt the same models, datasets, and proxy captioning setup as in previous experiments, and report results with the same evaluation metrics (AUC, Balanced Accuracy, and TPR@1%FPR). In the query-only scenario, however, the adversary possesses neither a shadow dataset nor a non-member reference set. Inference must be made solely from the signals derived for the target video. Consequently, all member and non-member samples are directly used for evaluation. Accordingly, we employ the intrinsic scoring and unsupervised fusion mechanism defined in Section 7.2 using pre-defined weights ($w_{SRF} \approx w_{TGS}$). No external calibration or classifier training is required.

7.4 Experimental Results

Overall Performance. Table 4 shows that our attack remains highly effective even in the most restrictive *query-only* setting. Without access to any reference distribution or labeled data, our method still achieves an AUC of 82.92% on AnimateDiff and 77.66% on Mira, while reaching a striking 97.01% on InstructVideo. These results demonstrate that the sparse-temporal artifacts of memorization are sufficiently strong to be detected directly from raw signals, with no external calibration. The ROC curves in Figure 8 further confirm

Table 5: Comparison of TGS signal against alternative temporal signals across three attack scenarios.

Signal Type	Supervised			Reference-based			Query-only		
	AUC(↑)	TPR@1%FPR(↑)	ACC(↑)	AUC(↑)	TPR@1%FPR(↑)	ACC(↑)	AUC(↑)	TPR@1%FPR(↑)	ACC(↑)
Subject Consistency	65.22%	0.84%	62.12%	74.19%	4.20%	68.13%	63.17%	1.20%	59.40%
Temporal Jitter	68.59%	1.84%	65.38%	67.42%	5.04%	67.26%	60.33%	1.00%	59.80%
TGS	**76.75%**	**2.52%**	**72.03%**	**82.63%**	**5.88%**	**77.65%**	**71.07%**	**2.00%**	**68.00%**

that our fused model consistently dominates the single-signal baselines, particularly in the low-FPR region where practical risks are most relevant.

Intrinsic Signal Power. The query-only results also provide insight into the relative strength of our signals. As shown in Table 4, SRF consistently outperforms TGS when used alone, indicating that sparse reconstruction fidelity serves as a more absolute, calibration-free indicator of memorization. By contrast, temporal stability (TGS) is more sensitive to the lack of reference normalization, which explains its weaker standalone performance in this setting. Nevertheless, fusing SRF and TGS yields the strongest overall performance across all models, confirming that their complementary nature persists even under zero-knowledge assumptions. This highlights the robustness of our *VidLeaks* and underscores the severe privacy risks posed by T2V models.

8 Ablation and Robustness Analysis

8.1 Effectiveness of the Top-K Strategy

Our Sparse Reconstruction Fidelity (SRF) is designed as a matched filter over *key anchors*, averaging similarities only over the Top-K most similar keyframes to combat content sparsity. We evaluate how K affects performance by comparing $K \in \{1,3,5\}$ against two averaging baselines: *All-key* (average over all keyframes, i.e., no Top-K filtering) and *All-frame* (average over all frames). Figure 9 reports AUC for both SRF-only and our fused attack (*VidLeaks*) across the three threat models.

The results reveal two consistent trends: (1) Performance is consistently stable for $K \in \{1,3,5\}$ in all scenarios and for both SRF and *VidLeaks*, indicating that SRF is insensitive to the precise choice of a small K so long as it focuses on the most relevant anchors; and (2) Removing the Top-K filter causes a clear drop: both *All-key* and *All-frame* substantially underperform Top-K variants across settings. This validates our core hypothesis from Section 3.1: averaging over many non-memorized (generalized) frames dilutes the sparse membership signal carried by a few memorized key anchors.

In summary, the Top-K mechanism is essential for isolating spatial memorization; a small K (e.g., $K=3$) provides a robust choice that consistently benefits SRF and, by extension, *VidLeaks* across all threat models.

8.2 Effectiveness of the Multi-Query Strategy

Our Temporal Generative Stability (TGS) signal relies on variance across multiple generations to capture determinism in temporal dynamics. To validate this design, we analyze the effect of the number of queries (Q) on attack performance. Figure 10 reports AUC scores for both the TGS-only attack and our fused method as Q increases from 2 to 5.

The results show a consistent and substantial performance gain with more queries. For instance, in the supervised setting, TGS improves from 66.3% (Q=2) to 76.8% (Q=5), while our fused method rises from 89.7% to 93.5%. Similar monotonic improvements are observed in the reference-based and query-only scenarios, with performance stabilizing when $Q \geq 4$. This trend confirms our hypothesis: a single generation cannot reliably reflect temporal stability, while aggregating across multiple runs yields a robust estimate of generative variance.

Overall, the Multi-Query design is essential for extracting strong temporal memorization signals, making TGS both effective and stable across different inference settings.

8.3 Comparison with Alternative Temporal Signals

To validate our design for Temporal Generative Stability (TGS), we compare it with two intuitive alternatives for capturing temporal dynamics: (1) **Subject Consistency**, a semantic-level metric that measures whether a subject's appearance remains consistent across frames using DINO features [58]; and (2) **Temporal Jitter**, a low-level motion metric that quantifies inter-frame pixel changes via RAFT optical flow [28]. For fairness, both alternatives are adapted into instability signals using the same multi-query formulation as TGS.

The results in Table 5 clearly show that: (1) *Low-level motion (Temporal Jitter) fails as a membership signal*, with AUC dropping to only 60.33% in the query-only scenario, indicating that pixel-level dynamics are too stochastic to separate members from non-members. (2) *Subject Consistency has moderate discriminative power*, but remains substantially weaker than TGS across all settings. In the critical reference-based scenario, for example, it achieves only 74.19% AUC

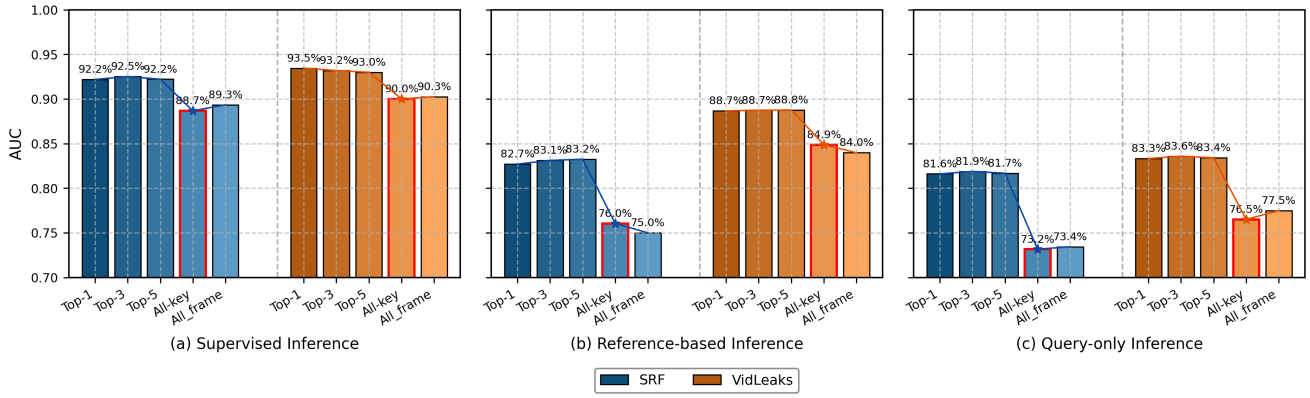


Figure 9: Impact of the Top-K strategy on SRF signal across three attack scenarios.

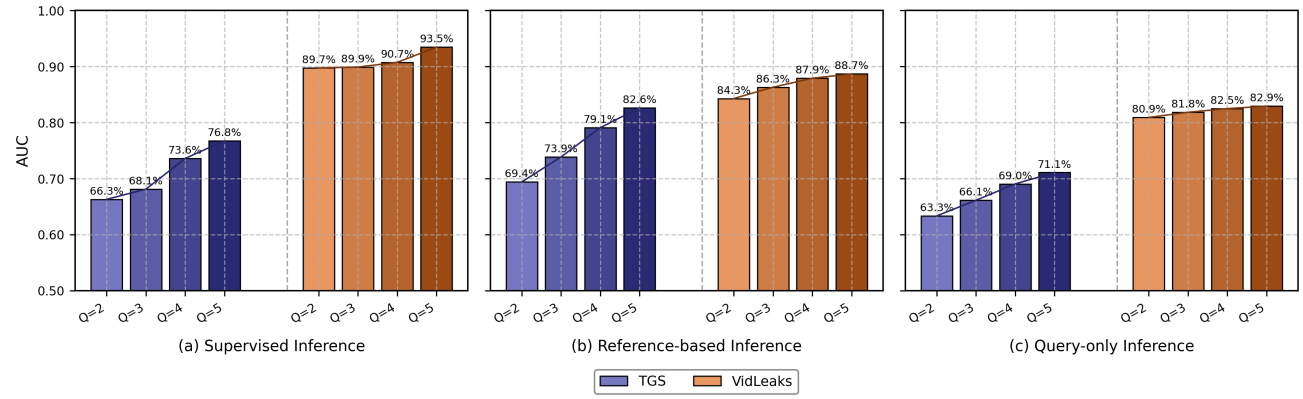


Figure 10: Impact of query count (Q) on TGS signal across three attack scenarios.

compared to 82.63% for TGS. (3) *TGS consistently outperforms both alternatives*, achieving the best results across all metrics and scenarios.

These findings empirically validate our core intuition: focusing solely on a subject introduces noise from its legitimate motion, while low-level pixel dynamics are inherently unstable. By contrast, TGS leverages the stability of the holistic scene, anchored by relatively static elements, to provide a robust and highly discriminative temporal signal for membership inference.

8.4 Comparison with Static Similarity Baselines

To assess the applicability of static similarity metrics in the video-based MIA setting, we compare *VidLeaks* against two intuitive baselines that are derived from image-based MIAs: (1) *Frame-wise CLIP Similarity*: Following [47], we compute CLIP similarity for each generated frame against all frames of the target video and average the results. This represents the most straightforward extension of image-based MIA to videos. (2) *Video-level Similarity*: We compute cosine similarity between VideoCLIP [59] embeddings of the target and generated videos, capturing holistic video semantics.

Table 6: Comparison with static similarity baselines across three attack scenarios.

Method	Supervised	Reference-based	Query-only
Frame-wise	89.3%	75.0%	73.4%
Video-level	57.2%	53.0%	47.5%
VidLeaks	**93.5%**	**88.7%**	**82.9%**

As shown in Table 6, both baselines substantially underperform our *VidLeaks* across all threat scenarios. Frame-wise similarity achieves only 73.4% AUC in the query-only setting (vs. 82.9% for *VidLeaks*), while video-level similarity yields near-random performance (AUC 47.5-57.2%). These results show that static similarity metrics fail to handle the sparsity and temporal dynamics of video memorization. Our SRF and TGS, however, specifically target and capture these sparse-temporal leakage signals from T2V models.

Table 7: Impact of caption source and quality on attack performance across three attack scenarios.

Prompt Type	Supervised			Reference-based			Query-only		
	AUC(↑)	TPR@1%FPR(↑)	ACC(↑)	AUC(↑)	TPR@1%FPR(↑)	ACC(↑)	AUC(↑)	TPR@1%FPR(↑)	ACC(↑)
Ground-Truth	89.06%	32.77%	83.24%	86.99%	10.84%	79.81%	80.67%	8.60%	74.70%
Doubao	98.06%	62.57%	93.39%	90.84%	34.45%	84.59%	85.25%	31.00%	77.50%
30% Word Dropout	90.34%	34.35%	84.12%	88.55%	32.77%	81.62%	82.53%	12.80%	74.70%
Gemini Pro (default)	93.46%	49.58%	87.58%	88.68%	27.73%	82.84%	82.92%	12.60%	76.60%

8.5 Impact of Caption Source and Quality

Our attack pipeline assumes a practical setting where the adversary only has access to the target video and thus must rely on a public captioning tool to generate a **proxy caption**. A natural question is how such proxy captions compare with the **ground-truth captions** used in training, and whether the attack is robust to variations in caption quality. To examine this, we evaluate our attack across all three inference scenarios using multiple caption sources: (1) ground-truth captions from the original datasets, (2) proxy captions generated by Gemini Pro [56], (3) proxy captions generated by an alternative model (Doubao [60]), and (4) degraded captions obtained by applying 30% random word dropout to Gemini Pro's outputs.

As shown in Table 7, proxy captions not only avoid degrading performance but in fact *often outperform* ground-truth captions. For example, in the supervised setting, AUC increases from 89.06% with ground-truth captions to 93.46% with Gemini Pro, and further to 98.06% with Doubao. Similar improvements appear under the reference-based and query-only scenarios. This counter-intuitive result may arise from two factors: (1) proxy captions produced by modern captioning models are typically more descriptive and semantically detailed, which forces the T2V model to reconstruct fine-grained spatial and temporal anchors, thereby amplifying SRF and TGS differences between members and non-members; and (2) in contrast, ground-truth captions from datasets such as WebVid-10M are often short or weakly descriptive. Although they were used during training, their limited expressiveness may fail to fully activate the model's memorized patterns.

Importantly, even under substantial caption degradation (30% random word dropout), the attack maintains strong performance (e.g., 90.34% AUC in the supervised setting), demonstrating robustness to imperfect caption quality. These findings reinforce the practicality of our video-only pipeline and show that advanced external captioners can, perhaps unexpectedly, yield stronger membership inference attacks than the ground-truth captions themselves.

8.6 Dataset Distribution Shift Analysis

In our evaluation, member samples originate from WebVid-10M (AnimateDiff, InstructVideo) or MiraData (Mira), while all non-member samples are drawn from Panda-70M. Although these datasets share broad semantic domains, they differ in collection pipelines, visual quality, and content composition, which may induce dataset-level distributional shifts. It is therefore important to assess whether the signals exploited by our attack reflect genuine model memorization or merely these distributional differences.

Table 8: Comparison between a blind dataset-based classifier and *VidLeaks* under the supervised setting.

Target Model	Method	AUC	TPR@1%FPR	ACC
AnimateDiff	Blind Classifier	68.39%	3.36%	64.10%
	VidLeaks	**93.46%**	**49.58%**	**87.58%**
Mira	Blind Classifier	82.05%	7.00%	76.00%
	VidLeaks	**87.45%**	**12.00%**	**82.00%**
InstructVideo	Blind Classifier	75.27%	2.38%	64.29%
	VidLeaks	**98.04%**	**45.00%**	**94.77%**

To isolate the effect of dataset-level distributional shift, we employ a *blind classifier*: an MLP trained solely on VideoCLIP embeddings extracted from raw videos (100 epochs, learning rate 1×10^{-5}), without using any generated outputs or temporal signals from the target T2V models. This reflects an upper bound on the separability attributable purely to dataset statistics. As shown in Table 8, the blind classifier reveals varying degrees of distribution shift across models. AnimateDiff shows the weakest shift (AUC 68.39%), InstructVideo exhibits a moderate shift (AUC 75.27%), likely due to its animal-focused fine-tuning, and Mira demonstrates the strongest shift (AUC 82.05%), indicating a more pronounced mismatch between MiraData and Panda-70M. Notably, despite these shifts, the blind classifier's discriminative power remains limited, particularly in the low-FPR regime.

Crucially, this distributional shift pattern does not align with the performance of *VidLeaks*. Our attack achieves substantially higher AUC on InstructVideo and AnimateDiff (up to 98.04%) despite their smaller dataset shifts, while its performance on Mira (AUC 87.45%) remains lower even though the distributional shift is greatest. Moreover, the blind classifier yields only weak high-confidence inference (TPR@1%FPR 2.38%–7.00%), far below the supervised performance of *VidLeaks* (TPR@1%FPR 12.00%–49.58%). These results, com-

Table 9: Performance under API-Level perturbation defenses across three attack scenarios.

Defense Setting	Supervised			Reference-based			Query-only		
	AUC(↑)	TPR@1%FPR(↑)	ACC(↑)	AUC(↑)	TPR@1%FPR(↑)	ACC(↑)	AUC(↑)	TPR@1%FPR(↑)	ACC(↑)
cfg_scale ± 0.5	93.48%	52.94%	87.48%	89.97%	19.33%	81.84%	83.96%	9.80%	76.90%
inference_steps ± 1	92.25%	57.14%	84.51%	89.82%	21.01%	82.50%	82.53%	8.40%	75.70%
VidLeaks	93.46%	49.58%	87.58%	88.68%	27.73%	82.84%	82.92%	12.60%	76.60%

bined with the distinct SRF and TGS patterns of member videos (Figure 2), demonstrate that dataset-level distributional bias alone cannot account for our results. Instead, *VidLeaks* exposes genuine sparse-temporal memorization effects inherent to the generative behavior of T2V models.

9 Discussion

Our study demonstrates that state-of-the-art T2V models are highly vulnerable to membership inference attacks through their sparse-temporal memorization. In this section, we discuss the broader implications of our findings, potential countermeasures, and the limitations of our study.

Summary of Key Findings. Our work provides three major insights into how T2V models memorize. First, we show that memorization is inherently dual-faceted: models tend to preserve *salient visual anchors* (captured by SRF) and *stable temporal dynamics* (captured by TGS). Second, we establish that these two signals are complementary rather than redundant—their fusion consistently amplifies attack power across architectures, datasets, and threat models. Third, and most critically, we reveal that these memorization artifacts are so strong that they can be reliably exploited even in the strictest **query-only** setting—without access to member data, non-member baselines, or model internals. This suggests that the membership leakage is not merely a statistical anomaly but a fundamental vulnerability of current T2V models, posing a severe and practical threat.

Countermeasures and Defenses. Mitigating our sparse-temporal MIA is a non-trivial challenge, as the exploited signals are inherently tied to the generative objectives of T2V models. Accordingly, effective mitigation is likely to require interventions that reduce memorization during training and data construction. Below, we outline two potential defense directions. (1) **Training-Time Defenses**. The fundamental protections are integrated during model optimization. For example, Differential Privacy (DP) [61] offers strong theoretical guarantees but remains impractical for large-scale generative models due to prohibitive costs. Alternatively, targeted regularization strategies could be designed to penalize over-deterministic temporal generation or excessively faithful reconstruction of training samples. (2) **Data Pre-processing**. Since memorization often stems from unique training examples, thorough filtering and de-duplication of pre-training corpora [62] may help reduce leakage risk. However, this comes at the potential cost of diminished data diversity and downstream performance.

Given that training-time modifications and large-scale data curation may be costly to deploy, we further evaluate a lightweight **API-level defense** that can be readily applied to existing systems. Specifically, we consider simple parameter randomization, where each query introduces small random jitter to key sampling parameters (*cfg_scale* ± 0.5 and *inference_steps* ± 1). As shown in Table 9, *VidLeaks* remains highly robust across all threat models: in supervised settings, AUC varies by less than 1.5%, and in the query-only setting performance remains above 82.5%. These results indicate that such lightweight perturbations do not meaningfully suppress the sparse-temporal signals that drive the attack, underscoring the limitations of simple API-level defenses.

Overall, our analysis suggests that mitigating membership leakage in modern T2V systems will likely require a combination of deeper training-time interventions, careful data curation, and system-level safeguards. Designing defenses that effectively suppress sparse-temporal memorization while preserving generation quality and usability remains an important and open challenge for future research.

Limitations and Future Work. Despite its strengths, our study has several limitations that suggest directions for future research. First, our evaluation focuses on three representative T2V models, which cover the major architectural paradigms but do not exhaust the rapidly evolving landscape of video generation. Extending the analysis to newer or larger-scale systems will be critical for validating the generality of our findings. Second, although we demonstrate that proxy captions generated by a public captioner are not only practical but often more effective than ground-truth captions (Section 8.5), our exploration of this dependency is not exhaustive. A deeper investigation into how different captioning strategies or language models influence membership inference remains an important open question. Finally, while we have outlined potential defenses, their concrete design, implementation, and rigorous evaluation are non-trivial. Developing privacy-preserving

training and inference mechanisms that specifically mitigate the sparse-temporal leakage channels we identify represents an important and challenging direction for future work.

10 Conclusion

In this paper, we present the first systematic study of membership inference attacks against modern text-to-video models. We identify two fundamental challenges unique to this domain—the sparsity of content memorization and the dynamics of temporal memorization—and propose the sparse-temporal MIA framework (*VidLeaks*) to address them. Our framework introduces two complementary signals: Sparse Reconstruction Fidelity (SRF) for detecting salient visual anchors and Temporal Generative Stability (TGS) for capturing stable temporal evolution.

Through extensive experiments across three diverse T2V models and under multiple threat models, we have demonstrated that these systems are highly vulnerable to membership inference. Notably, our attack remains effective even in the strict zero-knowledge, query-only setting, revealing a severe and practical privacy risk. Our study provides the first concrete evidence that T2V models leak significant membership information through both sparse fidelity and temporal dynamics, underscoring the urgent need for targeted defenses to safeguard data privacy and content ownership in the era of generative video.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant (No. 62372268, No. 62502276), Key R&D Program of Shandong Province, China (No. 2024CXGC010114, No. 2025CXPT085), and the Postdoctoral Innovation Program of Shandong Province, China (No. SDCX-ZG-202503030).

Ethical Considerations

This work presents the first systematic analysis of membership inference attacks against text-to-video (T2V) models. As security research that exposes system vulnerabilities, we fully acknowledge its dual-use potential and have carefully considered its ethical implications.

Stakeholders and Potential Impact. Our work primarily involves three stakeholder groups: (1) *T2V Model Developers and Companies*. Our research reveals previously unexplored privacy risks in their systems, enabling a clearer understanding of when and why MIAs succeed. This knowledge can support the development of defenses that reduce unintended memorization and improve privacy guarantees. (2) *Content Creators and Data Subjects*. The methodology clarifies the technical feasibility and limitations of using MIA to assess whether specific content may have been used during training. While this does not establish a definitive or legally robust auditing mechanism, it helps delineate the evidentiary strength and constraints of such approaches. Notably, improvements in model defenses informed by this line of research may simultaneously reduce the effectiveness of membership inference for external auditing, reflecting an inherent tension between privacy protection and auditability. (3) *The Research and Policy Community*. We provide crucial, empirical insights into video-level privacy risks, informing the development of ethical guidelines and technical standards for generative video technologies. A concurrent risk is that the publication could inform malicious actors.

Mitigations and Responsible Disclosure. We use only publicly released datasets under their respective licenses and do not attempt to identify individuals or recover video content. The attack is designed solely to reveal training data membership and is evaluated under progressively restrictive threat models. We provide methodological detail for scientific validation but intentionally avoid releasing turnkey attack tools, consistent with responsible disclosure principles.

Justification and Risk-Benefit Balance. Our work follows the preventive security research paradigm: proactively identifying privacy vulnerabilities in T2V models allows defenses to be developed before potential exploitation. Although the methodology could in principle be misused for membership inference, the risk is limited and further mitigated by our responsible disclosure practices. By contrast, concealing such vulnerabilities would leave data subjects exposed without impeding malicious discovery. Transparent disclosure provides clear benefits—strengthening accountability, informing risk assessment, and motivating privacy-preserving design. Therefore, responsible publication with ethical reflection serves the net interests of the security community and society.

Open Science

We follow open science principles to promote transparency and reproducibility. We provide access to the codebase implementing our attack pipeline, along with documentation describing how to set up the environment, obtain the publicly available datasets used in our experiments, and reproduce the main results reported in the paper. Because the raw video data are large and subject to license restrictions, we do not redistribute them directly; instead, we include metadata and scripts for retrieving the videos from their original public sources. All materials used in this paper — including code, configuration files, and representative result files — are available in a public repository: https://zenodo.org/records/17972831.

References

[1] OpenAI, "Sora," https://openai.com/index/sora/, 2024.

[2] Kwai, "Kling," https://kling.kuaishou.com, 2024, accessed: Aug. 2025.

[3] Luma, "Luma dream machine," https://lumalabs.ai/dream-machine, 2024.

[4] Runway, "Gen-3," https://runwayml.com/blog/introducing-gen-3-alpha, 2024.

[5] K. Sun, K. Huang, X. Liu, Y. Wu, Z. Xu, Z. Li, and X. Liu, "T2v-compbench: A comprehensive benchmark for compositional text-to-video generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8406–8416.

[6] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang *et al.*, "Lavie: High-quality video generation with cascaded latent diffusion models," *International Journal of Computer Vision*, vol. 133, no. 5, pp. 3059–3078, 2025.

[7] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[8] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.

[9] W. Fan, C. Si, J. Song, Z. Yang, Y. He, L. Zhuo, Z. Huang, Z. Dong, J. He, D. Pan *et al.*, "Vchitect-2.0: Parallel transformer for scaling up video diffusion models," *arXiv preprint arXiv:2501.08453*, 2025.

[10] X. Wu, S. Huang, G. Wang, J. Xiong, and F. Wei, "Boosting text-to-video generative model with mllms feedback," *Advances in Neural Information Processing Systems*, vol. 37, pp. 139 444–139 469, 2024.

[11] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.

[12] T.-S. Chen, A. Siarohin, W. Menapace, E. Deyneka, H.-w. Chao, B. E. Jeon, Y. Fang, H.-Y. Lee, J. Ren, M.-H. Yang *et al.*, "Panda-70m: Captioning 70m videos with multiple cross-modality teachers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 320–13 331.

[13] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1728–1738.

[14] W. Wang and Y. Yang, "Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 65 618–65 642, 2024.

[15] X. Ju, Y. Gao, Z. Zhang, Z. Yuan, X. Wang, A. Zeng, Y. Xiong, Q. Xu, and Y. Shan, "Miradata: A large-scale video dataset with long durations and structured captions," *Advances in Neural Information Processing Systems*, vol. 37, pp. 48 955–48 970, 2024.

[16] Y. Miao, Y. Zhu, L. Yu, J. Zhu, X.-S. Gao, and Y. Dong, "T2vsafetybench: Evaluating the safety of text-to-video generative models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 63 858–63 872, 2024.

[17] Z. Li and Y. Zhang, "Membership leakage in label-only exposures," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 880–895.

[18] H. Li, Z. Li, S. Wu, Y. Ye, M. Zhang, D. Feng, and Y. Zhang, "Enhanced {Label-Only} membership inference attacks with fewer queries," in *34th USENIX Security Symposium (USENIX Security 25)*, 2025, pp. 5465–5483.

[19] Y. He, B. Li, L. Liu, Z. Ba, W. Dong, Y. Li, Z. Qin, K. Ren, and C. Chen, "Towards label-only membership inference attack against pre-trained large language models," in *USENIX Security*, 2025.

[20] Y. Hu, Z. Li, Z. Liu, Y. Zhang, Z. Qin, K. Ren, and C. Chen, "Membership inference attacks against vision-language models," *arXiv preprint arXiv:2501.18624*, 2025.

[21] W. Menapace, A. Siarohin, I. Skorokhodov, E. Deyneka, T.-S. Chen, A. Kag, Y. Fang, A. Stoliar, E. Ricci, J. Ren *et al.*, "Snap video: Scaled spatiotemporal transformers for text-to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7038–7048.

[22] W. Feng, J. Li, M. Saxon, T.-j. Fu, W. Chen, and W. Y. Wang, "Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation," *arXiv preprint arXiv:2406.08656*, 2024.

[23] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, G. Liu, A. Raj *et al.*, "Lumiere: A space-time diffusion model for video generation," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.

[24] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng, "Magicvideo: Efficient video generation with latent diffusion models," *arXiv preprint arXiv:2211.11018*, 2022.

[25] H. Yuan, S. Zhang, X. Wang, Y. Wei, T. Feng, Y. Pan, Y. Zhang, Z. Liu, S. Albanie, and D. Ni, "Instructvideo: Instructing video diffusion models with human feedback," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6463–6474.

[26] Z. Huang, F. Zhang, X. Xu, Y. He, J. Yu, Z. Dong, Q. Ma, N. Chanpaisit, C. Si, Y. Jiang *et al.*, "Vbench++: Comprehensive and versatile benchmark suite for video generative models," *arXiv preprint arXiv:2411.13503*, 2024.

[27] M. Liao, Q. Ye, W. Zuo, F. Wan, T. Wang, Y. Zhao, J. Wang, X. Zhang *et al.*, "Evaluation of text-to-video generation models: A dynamics perspective," *Advances in Neural Information Processing Systems*, vol. 37, pp. 109 790–109 816, 2024.

[28] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*. Springer, 2020, pp. 402–419.

[29] Z. Zhang, Z. Yuan, X. Ju, Y. Gao, X. Wang, C. Yuan, and Y. Shan, "Mira: A mini-step towards sora-like long video generation," https://github.com/mira-space/Mira, 2024, available on GitHub.

[30] Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You, "Open-sora: Democratizing efficient video production for all," *URL https://github.com/hpcaitech/Open-Sora*, 2024.

[31] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 563–22 575.

[32] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.

[33] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[35] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan, "Videofusion: Decomposed diffusion models for high-quality video generation," *arXiv preprint arXiv:2303.08320*, 2023.

[36] R. Sun, Y. Zhang, T. Shah, J. Sun, S. Zhang, W. Li, H. Duan, B. Wei, and R. Ranjan, "From sora what we can see: A survey of text-to-video generation," *arXiv preprint arXiv:2405.10674*, 2024.

[37] D. J. Zhang, J. Z. Wu, J.-W. Liu, R. Zhao, L. Ran, Y. Gu, D. Gao, and M. Z. Shou, "Show-1: Marrying pixel and latent diffusion models for text-to-video generation," *International Journal of Computer Vision*, vol. 133, no. 4, pp. 1879–1893, 2025.

[38] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 7623–7633.

[39] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang *et al.*, "Hunyuanvideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.

[40] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[41] H. Wu, C. Chen, L. Liao, J. Hou, W. Sun, Q. Yan, and W. Lin, "Discovqa: Temporal distortion-content transformers for video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4840–4854, 2023.

[42] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[43] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[44] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[45] V. T. Truong, L. B. Dang, and L. B. Le, "Attacks and defenses for generative diffusion models: A comprehensive survey," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–44, 2025.

[46] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "Gan-leaks: A taxonomy of membership inference attacks against generative models," in *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, 2020, pp. 343–362.

[47] Y. Wu, N. Yu, Z. Li, M. Backes, and Y. Zhang, "Membership inference attacks against text-to-image generation models," *arXiv preprint arXiv:2210.00968*, 2022.

[48] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM transactions on intelligent systems and technology*, vol. 15, no. 3, pp. 1–45, 2024.

[49] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023.

[50] Y. Tang, J. Bi, S. Xu, L. Song, S. Liang, T. Wang, D. Zhang, J. An, J. Lin, R. Zhu *et al.*, "Video understanding with large language models: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[51] L. Momeni, M. Caron, A. Nagrani, A. Zisserman, and C. Schmid, "Verbs in action: Improving verb understanding in video-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 579–15 591.

[52] Z. Li, Y. Wu, Y. Chen, F. Tonin, E. Abad Rocamora, and V. Cevher, "Membership inference attacks against large vision-language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 98 645–98 674, 2024.

[53] P. Maini, H. Jia, N. Papernot, and A. Dziedzic, "Llm dataset inference: Did you train on my dataset?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 124 069–124 092, 2024.

[54] F. Bellard and F. contributors, "Ffmpeg: A multimedia framework," https://ffmpeg.org/, 2000, open-source software project.

[55] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[56] Google, "Google AI Studio," https://aistudio.google.com/, 2025.

[57] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE symposium on security and privacy (SP)*. IEEE, 2022, pp. 1897–1914.

[58] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[59] Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, Z. Wang, Y. Shi *et al.*, "Internvideo2: Scaling foundation models for multimodal video understanding," in *European Conference on Computer Vision*. Springer, 2024, pp. 396–416.

[60] ByteDance, "Volcano Engine: Doubao Large Model Suite," https://www.volcengine.com/, 2025.

[61] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[62] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini, "Deduplicating training data makes language models better," *arXiv preprint arXiv:2107.06499*, 2021.

A Attack Algorithms

This section provides the detailed pseudo-code for the attack methodologies presented in Section 5, Section 6, and Section 7.

Algorithm 1 Supervised Inference

Input: Target T2V model $\mathcal{M}$, captioning model $\mathcal{C}$, labeled shadow dataset $\mathcal{D}_{\text{shadow}}$, target video v.

procedure TRAIN ATTACK MODEL($\mathcal{D}_{\text{shadow}}, \mathcal{M}$)
  Initialize feature list $X \leftarrow \emptyset$, label list $Y \leftarrow \emptyset$
  for each $(v_i, \tilde{t}_i, y_i) \in \mathcal{D}_{\text{shadow}}$ **do**
    $\mathbf{v}_{srf} \leftarrow$ Compute SRF vector $(v_i, \tilde{t}_i, \mathcal{M})$
    $\mathbf{s}_{instab} \leftarrow$ Compute TGS vector $(v_i, \tilde{t}_i, \mathcal{M})$
    $\mathbf{x}_i \leftarrow$ Concatenate $(\mathbf{v}_{srf}, \mathbf{s}_{instab})$
    Append $\mathbf{x}_i$ to X and y_i to Y
  end for
  Split (X, Y) into a training set (X_{train}, Y_{train}) and a validation set.
  Train a classifier $\mathcal{A}_\theta$ on (X_{train}, Y_{train}).
  return Trained model $\mathcal{A}_\theta$
end procedure

procedure INFER MEMBERSHIP($v, \mathcal{M}, \mathcal{A}_\theta$)
  $\tilde{t} \leftarrow \mathcal{C}(v)$ Generate prompt
  $\mathbf{v}_{srf} \leftarrow$ Compute SRF vector $(v, \tilde{t}, \mathcal{M})$
  $\mathbf{s}_{instab} \leftarrow$ Compute TGS vector $(v, \tilde{t}, \mathcal{M})$
  $\mathbf{x} \leftarrow$ Concatenate $(\mathbf{v}_{srf}, \mathbf{s}_{instab})$
  $p \leftarrow \mathcal{A}_\theta(\mathbf{x})$
  return Membership probability p
end procedure

$\mathcal{A}_\theta \leftarrow$ Train Attack Model($\mathcal{D}_{\text{shadow}}, \mathcal{M}$)
$p \leftarrow$ Infer Membership($v, \mathcal{M}, \mathcal{A}_\theta$)
Output: p

B Additional Robustness Analysis

Table 10: Robustness of *VidLeaks* under video quality perturbations across three attack scenarios.

Quality Setting	Supervised	Reference-based	Query-only
H.264 Compression	91.70%	85.47%	81.50%
Resolution Reduction	92.52%	88.63%	82.75%
VidLeaks	**93.46%**	**88.68%**	**82.92%**

To further assess the robustness of *VidLeaks*, we evaluate its performance under common video quality perturbations, including (1) H.264 video compression and (2) spatial resolution reduction (e.g., from 1280×720 to 480×270). As shown in Table 10, *VidLeaks* maintains strong AUC performance under compression and resolution reduction across all threat scenarios, indicating that *VidLeaks* exploits intrinsic sparse-temporal memorization rather than low-level visual artifacts.

Algorithm 2 Reference-based Inference

Input: Target T2V model $\mathcal{M}$, captioning model $\mathcal{C}$, non-member reference set $\mathcal{D}_{\text{ref}}$, target video v.

procedure CALIBRATE STATISTICS($\mathcal{D}_{\text{ref}}, \mathcal{M}$)
  Initialize score lists $L_{SRF} \leftarrow \emptyset$, $L_{TGS} \leftarrow \emptyset$
  for each $(v_i, \tilde{t}_i) \in \mathcal{D}_{ref}$ **do**
    $S_{SRF,i} \leftarrow$ Compute SRF Score $(v_i, \tilde{t}_i, \mathcal{M})$
    $S_{TGS,i} \leftarrow$ Compute TGS Score $(v_i, \tilde{t}_i, \mathcal{M})$
    Append $S_{SRF,i}$ to L_{SRF} and $S_{TGS,i}$ to L_{TGS}
  end for
  $\mu_{srf}, \sigma_{srf} \leftarrow \text{Mean}(L_{SRF}), \text{StdDev}(L_{SRF})$
  $\mu_{tgs}, \sigma_{tgs} \leftarrow \text{Mean}(L_{TGS}), \text{StdDev}(L_{TGS})$
  return Statistics $(\mu_{srf}, \sigma_{srf}, \mu_{\text{tgs}}, \sigma_{\text{tgs}})$
end procedure

procedure INFER MEMBERSHIP($v, \mathcal{M}$, stats)
  $\tilde{t} \leftarrow \mathcal{C}(v)$ Generate prompt
  $S_{SRF} \leftarrow$ Compute SRF Score $(v, \tilde{t}, \mathcal{M})$
  $S_{TGS} \leftarrow$ Compute TGS Score $(v, \tilde{t}, \mathcal{M})$
  $\mathcal{A}_{SRF} \leftarrow (S_{\text{SRF}} - \mu_{\text{srf}})/\sigma_{\text{srf}}$
  $\mathcal{A}_{TGS} \leftarrow -(S_{\text{TGS}} - \mu_{tgs})/\sigma_{tgs}$
  $\mathcal{S}_{final} \leftarrow w_{srf} \cdot \mathcal{A}_{SRF} + w_{tgs} \cdot \mathcal{A}_{TGS}$
  return Membership score $\mathcal{S}_{final}$
end procedure

stats $\leftarrow$ Calibrate Statistics($\mathcal{D}_{\text{ref}}, \mathcal{M}$)
$S \leftarrow$ Infer Membership($v, \mathcal{M}$, stats)
Output: S

Algorithm 3 Query-only Inference

Input: Target T2V model $\mathcal{M}$, captioning model $\mathcal{C}$, target video v.

procedure INFER MEMBERSHIP($v, \mathcal{M}$)
  $\tilde{t} \leftarrow \mathcal{C}(v)$
  $S_{SRF} \leftarrow$ Compute SRF Score $(v, \tilde{t}, \mathcal{M})$
  $S_{TGS} \leftarrow$ Compute TGS Score $(v, \tilde{t}, \mathcal{M})$
  Compute intrinsic scores:
  $\mathcal{S}_{SRF} \leftarrow S_{SRF}$
  $\mathcal{S}_{TGS} \leftarrow 1 - S_{TGS}$
  Fuse scores:
  $\mathcal{S}_{final} \leftarrow w_{SRF} \cdot \mathcal{S}_{SRF} + w_{TGS} \cdot \mathcal{S}_{TGS}$
  return Membership score $\mathcal{S}_{\text{final}}$
end procedure

$S \leftarrow$ Infer Membership($v, \mathcal{M}$)
Output: S