MEMBERSHIP INFERENCE FOR GENERATIVE VIDEO MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

With growing interest in copyright, safety, and bias of generative video models, the need for robust membership inference (MI) for these models has never been greater. Yet, only one such method currently exists, designed for old generator architectures. What's more, it comes with problematic methodological design, since the members and non-members on which it was calibrated and tested came from two fundamentally distinct distributions. To allow for robust MI research in generative video models, we construct and open-source a framework and dataset of non-members complementing the VideoUFO dataset, used to train a large text-to-video (T2V) diffusion transformer. We then examine prominent MI methods from image generation and classification across the access spectrum and adapt them for T2V models. Additionally, we present our own gray-box method called The Reversal Gap (TRG) and find that an ensemble including multiple methods achieves the best performance.

1 INTRODUCTION

As text-to-video (T2V) generative models like Sora 2 (OpenAI, 2025) and Veo 3 (Google DeepMind, 2025) have reached levels of photorealism and versatility sufficient for use in movie production (Zhang et al., 2025b), advertising and marketing (Kapoor & Kumar, 2025), and journalism and documentary filmmaking (Kääpä & Sills-Jones, 2026), concerns about data privacy, copyright infringement, and data leaks in these models have emerged in parallel (Manduchi et al., 2024). These concerns are not hypothetical: Youtube CEO Neal Mohan has stated that training OpenAI's Sora on YouTube videos would violate platform terms (Bloomberg, 2024), and a large video corpus curated for the Open-Sora project was removed from Hugging Face following a DMCA request from a stock-footage provider (Furman, 2025). One avenue for auditing T2V models for these potential violations is membership inference (MI) (Shokri et al., 2017), where one attempts to determine whether a particular data point (in this case, a video) has been included in the model's training dataset.

The MI attack framings differ based on the amount of access given to the attacker. Black-box attacks assume access to the model as a closed API, with observability of the input and output spaces, but no access to model weights or intermediate state; membership must therefore be inferred from generated samples alone, typically by measuring the reconstruction fidelity between a generation and the candidate training sample as a proxy for memorization (Wu et al., 2022; Chen et al., 2020). Gray-box attacks assume some amount of additional information given to the attacker: the ability to invoke the denoiser at chosen timesteps, to observe intermediate latents along the reverse trajectory, or to evaluate the training objective without differentiating through it. This enables substantially sharper signals, such as comparing the model's step-wise denoising against the forward-process posterior (Duan et al., 2023), initializing the diffusion process at a proximal point and tracing the resulting trajectory (Kong et al., 2024), or exploiting conditional likelihood discrepancies between conditional and unconditional branches (Zhai et al., 2024). Finally, white-box attacks assume complete access to the model weights, enabling direct computation of the per-sample training loss (Matsumoto et al., 2023) and its gradients (Pang et al., 2023), which remain the strongest known signals on small- and medium-scale diffusion models.

While much progress has been made in MI for language models (Shi et al., 2024; Zhang et al., 2025a; Duan et al., 2024; Mattern et al., 2023) and image models (Wu et al., 2022; Duan et al., 2023; Kong et al., 2024; Pang et al., 2023; Li et al., 2024), the literature on MI for video models is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

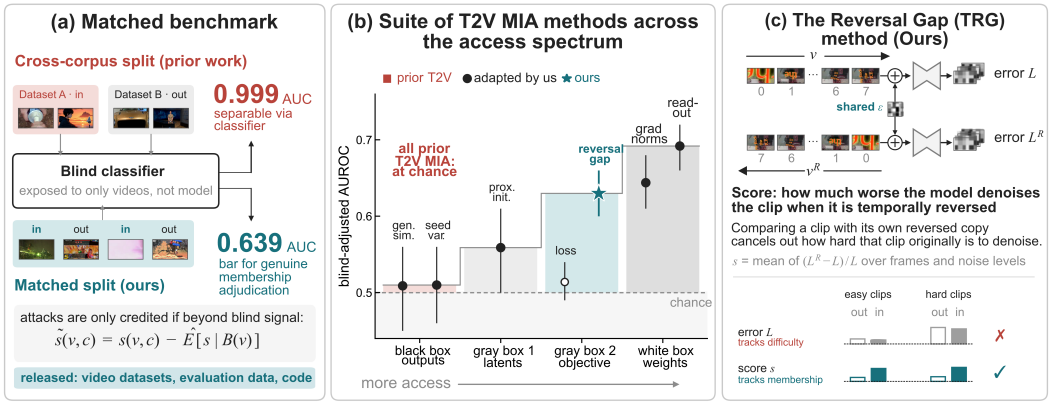

Figure 1: **Overview of the work.** (a) We introduce a matched non-member dataset for an existing member dataset used to train T2V models, constructed such that blind attacks are unable to separate the two. (b) We extend prominent MIA methods from other domains of computer vision and machine learning to T2V models across the access spectrum. (c) We introduce our own method, The Reversal Gap (TRG), and find that an ensemble including it is the best-performing approach.

nascent (Kapoor et al., 2025). The only prior MI method for T2V generators operates exclusively in the black-box regime, utilizing keyframe reconstruction similarity (Wang et al., 2026); the gray- and white-box regimes remain entirely unexplored for video. Moreover, the evaluation targets of this work are confined to earlier and smaller architectures, leaving the much larger Transformer-based architectures, which dominate contemporary T2V systems, unaudited. Adjacent work on video large language models exists (Li et al., 2026; Song et al., 2026), but it addresses a fundamentally different setting due to the difference in architecture and modality mapping.

We note that a further and more serious limitation cuts across this nascent literature:

> **Remark 1.1 (Blind Attack Susceptibility).** Recent work has shown that MI evaluations are frequently confounded, because members and non-members are drawn from fundamentally different distributions: ones that do not merely differ in their relationship to the scrutinized model, but in many other attributes. Blind attacks that never query the model at all (e.g., classification based on metadata or embeddings), can match or exceed state-of-the-art MI methods on published benchmarks (Das et al., 2025). This failure mode has been documented in the image (Dubiński et al., 2024), vision-language (Emelyanov et al., 2026), and audio (Dong et al., 2026) domains; it has not yet been tested for the video domain, where the only existing method evaluates on a cross-dataset split (Section 2.4).

To close this gap, we present the first systematic study of MI against T2V models across the full access regime spectrum, evaluated under distribution-matched conditions. To this end, we construct an evaluation protocol and artifacts, including model weights, data, and recipes, that allow one to conduct MIA work on recent T2V models. Additionally, we introduce our own T2V-specific gray-box MIA method called The Reversal Gap (TRG). Our contributions can be summarized as follows:

1. **First gray- and white-box MI study for T2V models.** We introduce the first MI attacks against T2V generators that exploit access beyond the output space (i.e., beyond black-box access), spanning the denoising loss surface, proximal initialization, per-block gradient norms, and internal representations. We evaluate these on MVDiT-VideoUFO, whose training corpus is known in its entirety.
2. **A distribution-matched evaluation protocol for video MIA.** We construct a split of 500 member/non-member pairs, where non-members are sampled from the identical distribution by reusing the same collection procedure, and report the blind-baseline performance. Whereas prior T2V MIA work suffered from a blind-baseline performance of 0.999, our split reaches 0.639 on the strongest of seven blind feature families. We consider AUROC after blind adjustment as the metric of primary interest. To verify the validity of our find-

ings further, we measure performance on models trained on completely distinct data, where even the best recorded method drops to AUROC of 0.573–0.622.

3. **A new gray-box MIA method tailored to T2V models.** To control for per-clip difficulty, which generally dominates the loss magnitude, we introduce a method that scores a candidate video clop against its own time-reversed copy and a shared noise sample. We demonstrate its superiority in the gray-box setting.

2 Related Work

2.1 Text-to-Video Generation

Contemporary T2V models descend from three general approaches: (1) adaptation of a pretrained image backbone through inserted temporal layers (Guo et al., 2023; Wang et al., 2023a; 2025b; Chen et al., 2023; Blattmann et al., 2023); (2) end-to-end training over a joint spatio-temporal representation, often pairing a 3D causal autoencoder with a diffusion transformer (Wang et al., 2025b; Chen et al., 2023; Blattmann et al., 2023; Yang et al., 2025; Wan et al., 2025; Wang & Yang, 2026; Polyak et al., 2024; Chen et al., 2025), the recipe of most current open-weight (Zheng et al., 2024; Hong et al., 2022) and, presumably, frontier models in the industry (OpenAI, 2025; Google DeepMind, 2025); and (3) post-training of an existing generator, through reward fine-tuning (Yuan et al., 2024; Prabhudesai et al., 2024) or consistency distillation (Wang et al., 2024c; 2023b; Salimans & Ho, 2022; Song et al., 2023). These distinctions are important to consider, as they have impact on the auditability of the resulting system: adapted architectures inherit spatial memorization from a frozen backbone, so reconstruction-based attacks may recover the image model's training set rather than the video corpus, and distillation shortens the reverse trajectory an attacker is able to interrogate. The limited prior work on MI for T2V has evaluated only on the first and third families.

2.2 Large-scale Video Datasets

Early large-scale video datasets scraped video clips from the internet and applied limited annotations; WebVid-10M (Bain et al., 2021), for example, used original alt-text from the source websites and carried watermarks, which the generators ended up learning (Ju et al., 2024; Nan et al., 2025), but has since been withdrawn. Subsequent efforts improved annotation and video quality in different ways: Panda-70M (Chen et al., 2024b) segmented long-form footage captioned by an ensemble of models; InternVid (Wang et al., 2024b) scaled to hundreds of millions of clips with language-model captions; MiraData (Ju et al., 2024) targeted long-duration video with dense, structured captions; and OpenVid-1M, VidGen-1M, and Koala-36M (Nan et al., 2025; Tan et al., 2024; Wang et al., 2025a) filtered aggressively on aesthetics, clarity, and temporal consistency. Lastly, instead of performing indiscriminate web scraping, VideoUFO (Wang & Yang, 2026) derived its topics from user prompts (Wang & Yang, 2024). Critically, these datasets differ in source platform, segmentation policy, resolution, watermarking, captioner identity, topical composition, and other properties (Chen et al., 2024b; Ju et al., 2024; Wang et al., 2025a), and so a member/non-member split drawn from two of them encodes all of these differences alongside the one of interest, not to mention possible overlaps.

2.3 Memorization in T2V Models

That generative models reproduce training data is well established for images, both through direct extraction (Carlini et al., 2023; Somepalli et al., 2023a) and through characterization of when replication occurs, with training-set duplication and highly specific conditioning text as principal drivers (Somepalli et al., 2023b); training data deduplication is the standard mitigation mechanism (Lee et al., 2022). This has since been also shown for video models: Rahman et al. (2025) document replication using a video-adapted copy-detection metric, and Chen et al. (2024a) separate content memorization from motion memorization, propose metrics for each, extract training videos from open-source generators, and find that video models also replicate their image pretraining corpora. These studies characterize the training datasets in aggregate, reporting how often generations exceed a similarity threshold (Rahman et al., 2025); an auditor (e.g., someone alleging copyright infringement or a privacy violation), holding just a single clip, instead needs a calibrated decision about that individual clip.

2.4 MEMBERSHIP INFERENCE

MI asks whether a data point appeared in a model's training set (Shokri et al., 2017; Yeom et al., 2018). Prior literature shows that this signal is largely captured by the point-level loss (Matsumoto et al., 2023). However, the decisive issue has often proved to be calibration rather than choice of scrutinized statistic. A single global threshold often conflates hard non-members with easy members, so scores are instead calibrated per example against reference models that did not train on the point in question (Carlini et al., 2022; Zarifzadeh et al., 2023). Below, we review this literature under each of the access regimes introduced in Section 1, and then turn to the state of MI for video.

White-box Attacks. Given weights, the denoising loss and gradient features have been shown as highly discriminative, often reaching near-perfect attack success on small- and medium-scale models (Pang et al., 2023). These numbers are fragile to details of the attack setup and preprocessing, however (Dubiński et al., 2024).

Gray-box Attacks. The SecMI method applies a single reverse step and measures how far the sample deviates from the data point in question (Duan et al., 2023). PIA substitutes the model's own noise prediction at $t = 0$ for the unknown ground-truth noise (Kong et al., 2024). Further variants exploit the discrepancy between conditional and unconditional predictions (Zhai et al., 2024) or replace pixel-wise comparison with structural comparison (Li et al., 2024). The analog for language models is the token-level likelihood (Shi et al., 2024), though several widely reported attacks perform close to chance once evaluated carefully at pretraining scale (Zhang et al., 2025a). While generally architecture-agnostic (Liu et al., 2026; Qi et al., 2026), these methods are often confounded by properties of the data itself: reconstruction error not only tracks membership of the data point but also how much high-frequency content an image contains, so detailed members can be misclassified as held out, and less detailed non-members as members (Lian et al., 2025; Rao et al., 2025).

Black-box Attacks. When the only observable signal is a generated sample without internal signals in the model, membership is generally scored by reconstruction fidelity, i.e., how closely the model can be induced to reproduce the candidate (Chen et al., 2020; Wu et al., 2022; Pang & Wang, 2023; Bohacek & Farid, 2025).

Blind Attacks. The blind attack, which mostly points to shortcomings of existing MIA methods, never queries the target model at all. Das et al. (2025) show that across eight published benchmarks, blind classifiers match or exceed SOTA MIA methods, meaning these benchmarks largely measure the difference between the member and non-member distributions. The same has since been confirmed for vision-language and audio (Emelyanov et al., 2026; Dong et al., 2026), yielding various responses (Dubiński et al., 2025; Maini et al., 2024).

Video-specific MIA. VidLeaks (Wang et al., 2026) is the only MIA method targeting T2V models. It is black-box, and fuses two signals: reconstruction fidelity over extracted keyframes and the variance of scene consistency across generations from the same prompt under different seeds. It is evaluated on AnimateDiff (Guo et al., 2023), Mira Ju et al. (2024), and InstructVideo (Yuan et al., 2024), with members drawn from WebVid-10M (Bain et al., 2021) and MiraData (Ju et al., 2024) and non-members from Panda-70M (Chen et al., 2024b). The confounding cross-distribution setup of this work matches the critique of Das et al. (2025). In fact, the authors themselves report a blind classifier reaching 68.39-82.05% AUC. The remaining work in video-specific MIA (Li et al., 2026; Song et al., 2026) targets models for video understanding, not video generation.

3 THREAT MODEL AND PROBLEM SETUP

A target T2V model g_θ is trained on a dataset $\mathcal{D}_{\text{train}} = \{(v_i, c_i)\}$ of captioned video clips. Given a candidate clip (v, c) and access level to g_θ, the adversary outputs a score $s(v, c) \in \mathbb{R}$ that is arbitrarily larger for $(v, c) \in \mathcal{D}_{\text{train}}$ than for $(v, c) \notin \mathcal{D}_{\text{train}}$.

We assume that the adversary has the candidate clip and a caption (not necessarily the caption from $\mathcal{D}_{\text{train}}$). Furthermore, we assume that the parameters of the target's preprocessing pipeline are known

to them. However, we do not assume any access to $\mathcal{D}_{\text{train}}$, and generally nor to a shadow model. We do evaluate a few methods where a shadow model is invoked, but state that additional assumption.

Definition 3.1 (Black box, BB). The adversary submits prompts and receives generated videos but does not have access to any internal state.

Definition 3.2 (Gray box level 1, GB1). The adversary may invoke the denoiser at arbitrary timesteps and observe intermediate latents, but may not evaluate the training objective.

Definition 3.3 (Gray box level 2, GB2). The adversary may evaluate the training objective $\mathcal{L}(v, c; \theta)$ at any chosen (t, ϵ), but without differentiating through it.

Definition 3.4 (White box, WB). The adversary has full access to the model weights, activations, and gradients.

4 Target Model and Data

We choose a Multi-modal Video Diffusion Transformer (MVDiT) (Nan et al., 2025) trained on VideoUFO (Wang & Yang, 2026) as our target g_θ. We find this to be the best option satisfying two criteria: (1) large model utilizing a modern T2V architecture, beyond mere adaptation of image models (Section 2.1) and (2) complete knowledge of $\mathcal{D}_{\text{train}}$.

Since MVDiT-VideoUFO was trained over the entirety of the VideoUFO dataset, no held-out members exist out of the box. To avoid susceptibility to blind attacks (Remark 1.1), we purposefully refrain from reaching for a separate dataset to source the non-members.[1] We are, hence, forced to construct the non-member set $\mathcal{D}_{\text{out}}$ ourselves.

We do so by utilizing the identical scraping pipeline used to create the original VideoUFO. In particular, we ensure that $\mathcal{D}_{\text{out}}$ is drawn from the same distribution as $\mathcal{D}_{\text{train}}$ while $\mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{out}} = \emptyset$. The resulting dataset contains $18{,}387$ clips across $1{,}286$ topics. We describe the collection procedure in Appendix C.1 and the resulting dataset in Appendix C.4.

5 Evaluation Protocol

We report AUROC with 95% bootstrap intervals resampled over matched pairs and the blind-adjusted AUROC. Any stochastic quantity is averaged over $K = 8$ noise draws and W temporal windows, chosen in a convergence experiment (Appendix D.3).

To avoid susceptibility to blind attacks (Remark 1.1), we adopt the following policy: before any attack is reported, we first measure what a blind classifier would achieve on the same configuration. An attack is only credited with detecting membership to the extent that it exceeds what is obtainable without the model at all, which we define as the blind baseline.

Furthermore, to make the reported metrics more informative with regards to the genuine MI ability, we perform a blind adjustment:

Definition 5.1. A blind-adjusted AUROC is obtained by regressing the attack score on the full blind feature block $B(v)$ with cross-fitted ridge regression and scoring the residual,

$$\tilde{s}(v, c) \;=\; s(v, c) \;-\; \hat{\mathbb{E}}\big[\, s \mid B(v) \,\big]. \tag{1}$$

[1] To illustrate the dangers of such an approach, consider the following configuration, which may seem, at first glance, reasonable. Pairing VideoUFO members with Panda-70M as non-members (Chen et al., 2024b) would allow for a blind classifier (specifically, an ensemble over 15 low-level video statistics) reach AUC of 0.996. This would diminish any additional findings. Normalizing the encoding of Panda-70M to match VideoUFO along codec, resolution, frame rate, and bitrate reduces this to 0.767, which is still rather undesirable.

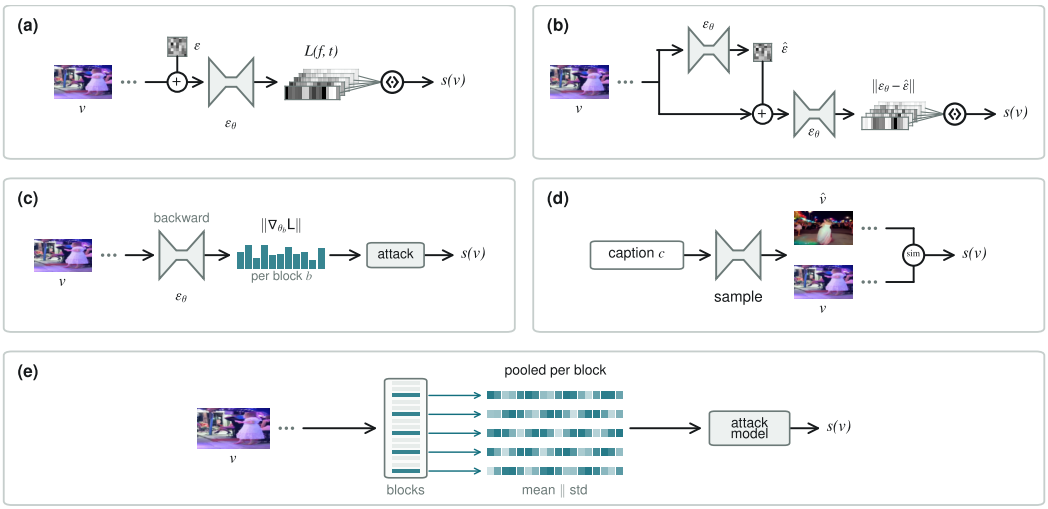

Figure 2: **Existing T2V-Adapted MIA Methods.** (a) Denoising loss, adapted from Matsumoto et al. (2023). (b) Proximal initialization, adapted from Kong et al. (2024). (c) Per-block gradient norms, adapted from Pang et al. (2023). (d) Generation similarity, adapted from Wang et al. (2026). (e) Representation read-out, adapted from Nasr et al. (2019).

> A score significantly above 0.5 means the attack carries membership information that no non-querying adversary could obtain.

6 METHODS

We next describe MIA methods studied in this work, which we adapted from MIA work in other modalities and domains. We then turn to our own method, which we outline in detail.

6.1 EXISTING METHODS ADAPTED FOR T2V

Since no gray- or white-box MIA methods exist for T2V generators, we adapt foundational methods from the language and image modeling literature, choosing at least one standard baseline or state-of-the-art method per access level. These methods, shown in Figure 2, were constructed for architectures where a single scalar loss is available. However, in our case, the T2V denoiser produces a set of losses at different frames and timesteps, and so we need to apply a temporal reduction to process these losses in a unified way and help surface the underlying signal in the attack. We hence average the losses for all methods.

Denoising loss. Adapted from Matsumoto et al. (2023) and shown in Figure 2(a), this attack scores a clip by the negative of its training objective, since members sit closer to a local optimum and should hence be denoised more accurately. We evaluate $\mathcal{L}(v, c; \theta)$ at a fixed grid of timesteps, using a seeded noise draw shared across all candidates to remove sampling variance between them. This is the weakest assumption available in the GB2 regime, and we treat it as the reference point for every later refinement.

Proximal initialization. Adapted from Kong et al. (2024) and shown in Figure 2(b), this attack replaces the unknown ground-truth noise with the model's own prediction at $t \approx 0$. The clip is re-noised with that estimate, and membership is scored by the distance between the model's subsequent noise prediction and the estimate it was given. The substitution removes the sampling variance that dominates single-draw loss estimates, and so the statistic is considerably cheaper than a timestep sweep at comparable stability.

Per-block gradient norms. Adapted from Pang et al. (2023) and shown in Figure 2(c), this attack differentiates the training objective with respect to the parameters, since a fitted example should produce a smaller update. We record the gradient norm separately for each transformer block rather than only in aggregate, to preserve the depth structure that a white-box attack model can exploit (Nasr et al., 2019), and fit the attack model on the resulting per-block vector.

Generation similarity and seed variance. Adapted from Wang et al. (2026) and shown in Figure 2(d), this is the only prior MIA method for T2V generators. It samples from the target using the candidate's caption and fuses two black-box signals: the maximum similarity between generated and candidate frames, and the variance of scene consistency across generations from the same prompt under different seeds. We reproduce both signals from the released code, sampling five generations per caption. We additionally report the z-scored fusion of the two, since the published fusion adds quantities differing by an order of magnitude in scale and is in practice dominated by the similarity term.

Representation read-out. Adapted from Nasr et al. (2019) and shown in Figure 2(e), this attack fits an attack model on the target's intermediate activations rather than on any quantity derived from its output. For a noised clip we record the token-pooled mean and standard deviation of every transformer block and concatenate over blocks, timesteps, and noise draws. The concatenation over timesteps has no analogue in the original formulation, since a classifier has no such axis, and we fit a cross-fitted linear model on the leading principal components rather than the learned per-component encoders of the original, as the latter offers no gain at our sample size.

6.2 Our Method: The Reversal Gap (TRG)

Every method above measures the model's response to a candidate in absolute terms and asks whether that response looks like a member's. This is what makes them vulnerable to the difficulty confound, since the denoising loss of a clip depends on its high-frequency and motion content at least as much as on its membership, and no amount of matching removes a property of the individual clip. Our method instead scores a clip against a perturbed version of itself, so that whatever makes the clip hard to denoise appears on both sides of the comparison and cancels.

Reversal gap (GB2). We score a clip against itself rather than against a threshold. Let $L(f,t)$ be the per-frame, per-timestep denoising loss of a clip v under caption c, and let $L^R(f,t)$ be the same quantity computed on the time-reversed clip v^R using the *same* noise draw, as shown in Figure 3. The membership score is the normalized difference,

$$s(v,c) \;=\; \big\langle \big(L^R(f,t) - L(f,t)\big) \,/\, L(f,t) \big\rangle_{f,t}, \tag{2}$$

where $\langle\cdot\rangle_{f,t}$ denotes the mean over frames and timesteps. Since both terms share the clip, the caption, and the noise, any per-clip difficulty term that scales the loss cancels, which is precisely what the absolute-loss attacks cannot do. Two properties of the perturbation matter. It must preserve the data manifold, since a time-reversed video remains a plausible video whereas a frame-shuffled one does not, and it must share the noise draw, since with independent draws the variance of the difference exceeds the effect. We further average the score over K noise draws and W temporal windows, sampled as the training pipeline sampled them, to marginalize the two nuisance variables that training itself randomized.

7 Results

We first examine the performance of every method of Section 6 on the distribution-matched split. Table 1 reports the AUROC and the blind-adjusted AUROC of each method, together with the blind baseline and the number of queries per candidate.

Main results. The methods reading an absolute loss are near chance, with the denoising loss at 0.514, proximal initialization at 0.485, and the scalar gradient norm at 0.523. The two black-box signals of Wang et al. (2026) reach 0.473 and 0.514, and neither is lifted above chance by the blind adjustment. This shows that these signals do not transfer to our matched split, not that the

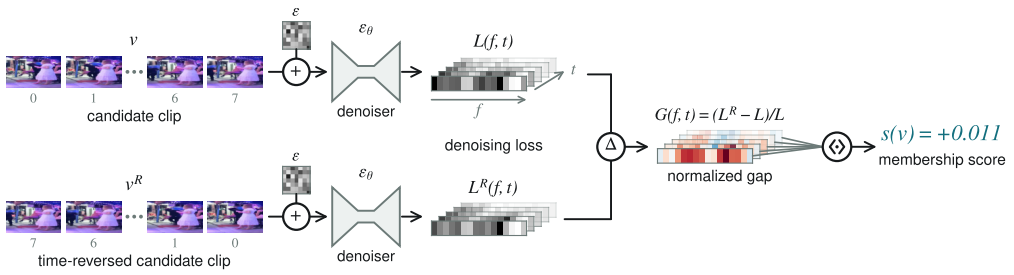

Figure 3: **Our Method: The Reversal Gap.** This GB2 MIA method utilizes a single, shared noise ϵ for both the v and v^R passes through the model. The measured gap is over both frames and timesteps, averaged down to a single score.

Table 1: **MI Against MVDiT-VideoUFO.** These attacks are performed on our distribution-matched dataset (Section 4). Blind-adj. AUROC, as defined in Definition 5.1, is treated as the primary metric for comparison. Brackets give 95% bootstrap intervals over matched pairs. Queries are denoiser evaluations per candidate; † additionally requires a backward pass. The joint model is a single cross-fitted classifier on the concatenated features of the representation read-out, the per-block gradients, and the reversal gap.

Method	Access	AUROC	Blind-adj. AUROC	Queries
Blind baseline (best family)	—	0.639 [0.61, 0.67]	—	0
Generation similarity (Wang et al., 2026)	BB	0.473 [0.42, 0.52]	0.509 [0.45, 0.56]	300
Seed variance (Wang et al., 2026)	BB	0.514 [0.46, 0.57]	0.510 [0.46, 0.56]	300
Proximal initialization (Kong et al., 2024)	GB1	0.485 [0.44, 0.53]	0.559 [0.50, 0.61]	10
Denoising loss (Matsumoto et al., 2023)	GB2	0.514 [0.49, 0.54]	—	3
Per-block gradient norms (Pang et al., 2023)	WB	0.646 [0.61, 0.68]	0.644 [0.61, 0.68]	3†
Representation read-out (Nasr et al., 2019)	WB	**0.717 [0.68, 0.75]**	0.692 [0.66, 0.72]	6
The reversal gap (TRG, Ours)	GB2	0.608 [0.58, 0.64]	0.630 [0.60, 0.66]	32
All signals, joint model	WB	0.702 [0.67, 0.73]	**0.697 [0.66, 0.73]**	38†

published numbers are incorrect on their original targets, which we have not reproduced. At gray-box access, the reversal gap reaches 0.608, or 0.630 after blind adjustment, and is the only method above chance at this access level. At white-box access, the per-block gradient norms reach 0.646 and the representation read-out reaches 0.717, the highest raw AUROC in Table 1. The joint model reaches 0.702, or 0.697 after blind adjustment, which is the highest blind-adjusted AUROC but within the interval of the read-out alone. This means that the three signals it combines are largely redundant.

7.1 Is the Evaluation Clean?

We next turn to the blind baseline. Table 2 reports the AUROC of seven blind feature families on our matched split and on the cross-corpus construction used by prior work. On the cross-corpus construction, CLIP features alone separate members from non-members at 0.999. On our matched split, the strongest family is the 12-dimensional radial FFT profile at 0.639, and the three families used in the matching cost reach 0.531, 0.540, and 0.610. The spectral profile was not used in the matching, and so it is the least biased estimate of the residual separability, which is hence spectral rather than semantic.

7.2 Further Analyses

We report further analyses in Appendix D and summarize them here. Comparing each member against its matched non-member shows that the denoising loss largely tracks how hard a clip is to denoise, and that the reversal gap removes most of this dependence (Appendix D.1). The signal of the read-out is weakest in the first transformer blocks and peaks mid-network, at block 13 with an

Table 2: **Blind Baselines by Feature Family.** Each family is a classifier that never queries the target model, trained on the same 500 matched pairs and cross-fitted with pairs held together across folds. The rightmost column gives the same family on the cross-corpus construction used by prior work.

Feature family	Matched	Cross-corpus
Spectral (radial FFT)	**0.639**	—
CLIP-bigG	0.610	0.999
VBench-style quality	0.579	0.840
Temporal differences	0.551	—
DINOv2-large	0.540	0.807
Encoding statistics	0.531	0.955
Color histogram	0.497	—

Table 3: **Reference-Model Verification.** The representation read-out applied to three models of identical architecture: the target, a reference model trained on Panda-70M, and a randomly initialized model.

Weights	AUROC	Blind-adj. AUROC
Target	**0.717**	**0.692 [0.66, 0.72]**
Random initialization	0.573	0.566 [0.53, 0.60]
Reference (Panda-70M)	0.622	0.587 [0.55, 0.62]

AUROC of 0.731 (Appendix D.3). A single timestep and a single group of mid-network blocks each match the full read-out, which reduces its cost from 6 queries to 2, and its performance saturates by 400 pairs. Neither a more expressive attack model nor conditioning on the exact training caption improves the attacks, and calibrating against the reference model lifts the denoising loss but not the reversal gap (Appendix D.2). No measured clip property predicts which members are identified (Appendix D.4). Finally, the read-out survives re-encoding and downscaling of the candidate, with a blind-adjusted AUROC of at least 0.611 under every condition (Appendix D.5).

7.3 Controls

We next turn to controls for the read-out. Table 3 reports the read-out applied, with identical taps, pooling, and classifier, to the target, to a randomly initialized model, and to a reference model of identical architecture trained on Panda-70M. The target reaches 0.717, the reference model 0.622, and the randomly initialized model 0.573, or 0.692, 0.587, and 0.566 after blind adjustment. This shows that the read-out measures what the target learned from its training data, with a margin of 0.105 blind-adjusted AUROC over a model that never saw the members.

8 Discussion

Prior to this work, the landscape of MIA for T2V models was practically nonexistent, and the only existing method was flawed in its evaluation design, since its members and non-members were drawn from fundamentally different distributions. We believe that this work lays the groundwork for more transparent and productive research in this space, providing a distribution-matched evaluation protocol, a non-member dataset, and adaptations of existing methods across the access spectrum. We also introduce a method that exploits the temporal structure of video diffusion models, achieves the best performance in the gray-box setting, and, as part of an ensemble with the adapted methods, achieves the best performance overall.

8.1 Limitations

Our experiments are conducted on a single model, since no other modern T2V model discloses the dataset it was trained on, which in many ways restricted any additional comparisons. Moreover, this model saw each member only 20 times among 1.09M clips and memorizes weakly, and so our absolute numbers may understate what is achievable against models trained on smaller or more repetitive data.

AI USE STATEMENT

For some experiments in this work, we used generative AI tools to aid with code implementation, data cleaning, and feedback on experimental design. We have not used generative AI tools for the following tasks, which are not applicable to this work: data generation, development of theoretical models or conceptual frameworks, formulation of mathematical claims, provision of critical ingredients for proving mathematical claims, assistance in the writing of proofs, refinement of hypotheses, translation assistance, and support with qualitative or thematic data analysis. Additionally, we used generative AI tools for help with formatting of LaTeX tables and figures and revisions/suggestions for paper writing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The authors have considered the ethical impacts of this work and found that the ability to conduct MIA against T2V models is beneficial for individuals seeking to protect their intellectual property rights, as well as for other auditing and security purposes. That said, we also see potential risks. The same methods could be used to determine whether a video depicting a specific individual was used to train a model, which reveals information about that individual, and our non-member set itself consists of publicly available videos of real people. We hence release the weights and the dataset under gated access, with each request reviewed by the research team, while the code is fully open-sourced.

REPRODUCIBILITY STATEMENT

The code, data, and weights for the project are available at `https://anonymous.4open.science/r/iclr-t2v-mia-CC9E/`. The code is fully open-sourced, whereas the weights and the dataset are released under gated access, as described in the Ethics Statement.

ACKNOWLEDGMENTS

Anonymized

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew 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*, pp. 1728–1738, 2021.

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

Bloomberg. YouTube says using videos to train OpenAI's Sora breaks rules. Bloomberg News video, April 2024. URL `https://www.bloomberg.com/news/videos/2024-04-04/youtube-says-using-videos-to-train-openai-s-sora-breaks-rules`.

Matyas Bohacek and Hany Farid. GenAI Confessions: Black-box membership inference for generative image models. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 321–330. IEEE, 2025.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Chen Chen, Enhuai Liu, Daochang Liu, Mubarak Shah, and Chang Xu. Investigating memorization in video diffusion models. *arXiv preprint arXiv:2410.21669*, 2024a.

Dingfan Chen, Ning Yu, Yang Zhang, and Mario 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*, pp. 343–362, 2020.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan 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*, pp. 13320–13331, 2024b.

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. In *2025 IEEE Security and Privacy Workshops (SPW)*, pp. 118–125. IEEE, 2025.

Jia-Kai Dong, Yu-Xiang Lin, and Hung-Yi Lee. Membership inference attacks against large audio language models. *arXiv preprint arXiv:2603.28378*, 2026.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pp. 8717–8730. PMLR, 2023.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*, 2024.

Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemyslaw Rokita, Tomasz Trzciński, and Paweł Morawiecki. Towards more realistic membership inference attacks on large diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4860–4869, 2024.

Jan Dubiński, Antoni Kowalczuk, Franziska Boenisch, and Adam Dziedzic. CDI: Copyrighted data identification in diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18674–18684, 2025.

Anton Emelyanov, Sergei Kudriashov, and Alena Fenogenova. FiMMIA: scaling semantic perturbation-based membership inference across modalities. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 139–153, 2026.

Josh Furman. Report: Copyright infringement, June 2025. URL `https://huggingface.co/datasets/hpcai-tech/open-sora-pexels-45k/discussions/2`.

Google DeepMind. Veo 3 tech report, May 2025. URL `https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf`. Accessed: July 20, 2026.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. MiraData: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

Pietari Kääpä and Dafydd Sills-Jones (eds.). *Documentary in the Age of AI: Creativity, Power, Technology*, volume 5 of *Documentary Film Cultures*. Peter Lang, Oxford, 2026. ISBN 978-1-80374-888-7. doi: 10.3726/b22530. URL https://www.peterlang.com/document/1510340.

Anuj Kapoor and Madhav Kumar. Frontiers: Generative AI and personalized video advertisements. *Marketing Science*, 44(4):733–747, 2025.

Shashank Kapoor, Sanjay Surendranath Girija, Lakshit Arora, Dipen Pradhan, Ankit Shetgaonkar, and Aman Raj. Adversarial attacks in multimodal systems: A practitioner's survey. In *2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1643–1650. IEEE, 2025.

Fei Kong, Jinhao Duan, Heng Tao Shen, Xiaoshuang Shi, Xiaofeng Zhu, Kaidi Xu, et al. An efficient membership inference attack for the diffusion model by proximal initialization. In *International Conference on Learning Representations*, volume 2024, pp. 57876–57896, 2024.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, 2022.

Qi Li, Runpeng Yu, and Xinchao Wang. Vid-SME: Membership inference attacks against large video understanding models. *Advances in Neural Information Processing Systems*, 38:111572–111596, 2026.

Qiao Li, Xiaomeng Fu, Xi Wang, Jin Liu, Xingyu Gao, Jiao Dai, and Jizhong Han. Unveiling structural memorization: Structural membership inference attack for text-to-image diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10554–10562, 2024.

Puwei Lian, Yujun Cai, and Songze Li. Unveiling impact of frequency components on membership inference attacks for diffusion models. *arXiv e-prints*, pp. arXiv–2505, 2025.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Yuxuan Liu, Peihong Zhang, Rui Sang, Zhixin Li, Yizhou Tan, Yiqiang Cai, and Shengchen Li. Membership inference attack against music diffusion models via generative manifold perturbation. *arXiv preprint arXiv:2602.01645*, 2026.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. LLM dataset inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092, 2024.

Laura Manduchi, Clara Meister, Kushagra Pandey, Robert Bamler, Ryan Cotterell, Sina Däubener, Sophie Fellenz, Asja Fischer, Thomas Gärtner, Matthias Kirchler, et al. On the challenges and opportunities in generative AI. *arXiv preprint arXiv:2403.00025*, 2024.

Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops (SPW)*, pp. 77–83. IEEE, 2023.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11330–11343, 2023.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. In *International Conference on Learning Representations*, volume 2025, pp. 1045–1064, 2025.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 739–753. IEEE, 2019.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PmLR, 2021.

OpenAI. Sora 2 system card, September 2025. URL `https://cdn.openai.com/pdf/50d5973c-c4ff-4c2d-986f-c72b5d0ff069/sora_2_system_card.pdf`. Accessed: July 20, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Yan Pang and Tianhao Wang. Black-box membership inference attacks against fine-tuned diffusion models. *arXiv preprint arXiv:2312.08207*, 2023.

Yan Pang, Tianhao Wang, Xuhui Kang, Mengdi Huai, and Yang Zhang. White-box membership inference attacks against diffusion models. *arXiv preprint arXiv:2308.06405*, 2023.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie Gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

Tao Qi, Huili Wang, Yuanhong Huang, Wendan Wang, Lianchao Zhao, Jinrui Wang, Zichen Qin, Shangguang Wang, and Yongfeng Huang. Black-box membership inference attacks on the pretraining data of image-generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 62–71, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Aimon Rahman, Malsha V Perera, and Vishal M Patel. Frame by familiar frame: Understanding replication in video diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2766–2776. IEEE, 2025.

Mingxing Rao, Bowen Qu, and Daniel Moyer. Score-based membership inference on diffusion models. *arXiv preprint arXiv:2509.25003*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, volume 2024, pp. 51826–51843, 2024.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6048–6058, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023b.

Wei Song, Yuxin Cao, Ziqi Ding, Yi Liu, Gelei Deng, and Yuekang Li. Membership inference attacks against video large language models. *arXiv preprint arXiv:2604.27002*, 2026.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. VidGen-1M: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. ModelScope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

Li Wang, Wenyu Chen, Ning Yu, Zheng Li, and Shanqing Guo. VidLeaks: Membership inference attacks against Text-to-Video models. *arXiv preprint arXiv:2601.11210*, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36M: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8428–8437, 2025a.

Wenhao Wang and Yi Yang. VidProM: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024.

Wenhao Wang and Yi Yang. VideoUFO: A million-scale user-focused dataset for Text-to-Video generation. *Advances in Neural Information Processing Systems*, 38, 2026.

Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. VideoLCM: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023b.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025b.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024b.

Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. LiFT: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024c.

Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models. *arXiv preprint arXiv:2210.00968*, 2022.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. InstructVideo: Instructing video diffusion models with human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6463–6474, 2024.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. *arXiv preprint arXiv:2312.03262*, 2023.

Shengfang Zhai, Huanran Chen, Yinpeng Dong, Jiajun Li, Qingni Shen, Yansong Gao, Hang Su, and Yang Liu. Membership inference on text-to-image diffusion models via conditional likelihood discrepancy. *Advances in Neural Information Processing Systems*, 37:74122–74146, 2024.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Li. Min-K%++: Improved baseline for pre-training data detection from large language models. In *International Conference on Learning Representations*, volume 2025, pp. 64845–64862, 2025a.

Ruihan Zhang, Borou Yu, Jiajian Min, Yetong Xin, Zheng Wei, Juncheng Nemo Shi, Mingzhen Huang, Xianghao Kong, Nix Liu Xin, Shanshan Jiang, et al. Generative AI for film creation: A survey of recent advances. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6267–6279, 2025b.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

A Evaluation Protocol

One of the contributions of this work is a protocol under which modern diffusion-based T2V architectures can be studied for robust MIA. As a part of this protocol, we provide the target model, MVDiT-VideoUFO, described in full in Appendix B, the non-member set, VideoUFO-NonMember, described in Appendix C, and the blind feature set described below. All components are available at `https://anonymous.4open.science/r/iclr-t2v-mia-CC9E/`, with the code fully open-sourced and the weights and the dataset released under gated access.

A.1 Blind Features

The following feature categories are utilized for the blind baseline analysis. These are all based on the video clip content alone, without access to the model.

- **Learned semantic embeddings**: CLIP-bigG and DINOv2-large, pooled over frames.
- **Handcrafted encoding statistics**: resolution, aspect, duration, frame rate, file size, sharpness, inter-frame motion, and color moments.
- **Perceptual quality**: VBench-style dimensions, subject and background consistency, temporal flickering, motion smoothness, imaging quality, aesthetic proxies.
- **Spectral**: a radial FFT energy profile, which captures compression and resampling history.
- **Temporal**: first- and second-difference statistics.
- **Text**: caption TF-IDF and a sentence encoder.

B MVDiT-VideoUFO

MVDiT-VideoUFO is a Multi-modal Video Diffusion Transformer (MVDiT) (Nan et al., 2025) trained by Wang & Yang (2026) on the entirety of VideoUFO. The weights are available at `https://anonymous.4open.science/r/iclr-t2v-mia-CC9E/`. Below, we describe its architecture, training recipe, and qualitative examples.

B.1 Architecture

MVDiT follows the diffusion transformer design, operating on the latents of a pretrained variational autoencoder (Rombach et al., 2022) applied frame by frame, and is conditioned on text embeddings from a T5-XXL encoder (Raffel et al., 2020) truncated to 120 tokens. We use the XL/2 variant, which has 28 transformer blocks of hidden width 1,152 and 1.07B parameters in total. Each 16-frame clip at 256×256 is encoded to a 4-channel latent of 32×32 per frame and divided into 2×2 spatial patches, yielding 256 spatial tokens per frame. Each block comprises spatial self-attention, temporal self-attention, and cross-attention to the text. Unlike designs adapted from image models, each block also carries a dedicated feed-forward branch for the text tokens, so that the text representation is updated alongside the video.

B.2 Training Recipe

The model is trained for 20 epochs over VideoUFO, which amounts to 72,500 optimization steps, and we use the exponential-moving-average weights of the final checkpoint throughout. Each training sample is a 16-frame clip taken at a frame interval of 3 from a randomly positioned temporal window, flipped horizontally with probability 0.5, and resized and center-cropped to 256×256. The diffusion process follows the improved DDPM formulation (Nichol & Dhariwal, 2021) with 1,000 training timesteps. These randomized augmentations matter for MIA, since the model never sees a member clip in exactly the same form twice, and so we average our attack scores over temporal windows and noise draws (Section 5).

B.3 Qualitative Examples

Eight clips drawn at random (seed 0) from the member set with the training captions, demonstrating model generations on text seen during training. Frames are sampled at uniform intervals across the interior of each clip.

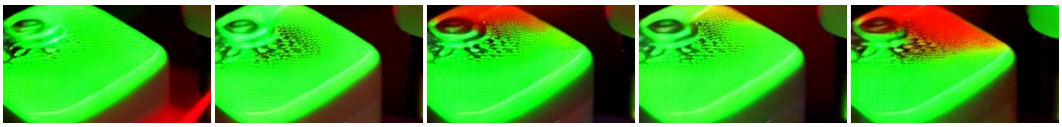

Figure B.3.1: A green 3D-printed object is being scanned by a red laser beam inside a machine, highlighting its intricate details.



Figure B.3.2: A person rides a large blue inflatable ring around a poolside area with tiled flooring and lush greenery in the background.



Figure B.3.3: Confucius stands before a black background with his quote about signs and symbols displayed beside him.



Figure B.3.4: A man with dreadlocks dances energetically on a stage, surrounded by a crowd under bright lights.

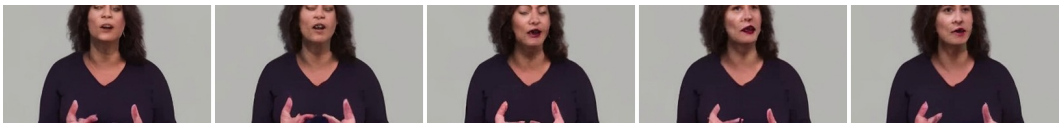

Figure B.3.5: A woman stands against a plain background, discussing digital rights and artistic expression online as part of a presentation titled ”Magna Carta: My Digital Rights.”

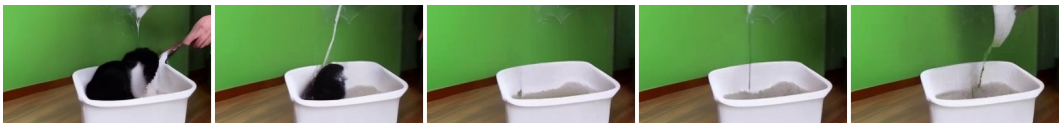

Figure B.3.6: A person is pouring cat litter into a white litter box on a wooden floor with a green wall featuring a panda design.



Figure B.3.7: Shrek stands outside a building, shouting angrily as a dog runs past him, with a picnic scene featuring a woman and a man lying on a blanket in the background.



Figure B.3.8: A World Health Day logo featuring a globe with a heartbeat line and a stethoscope is displayed on a screen, promoting awareness and health initiatives globally.

C VIDEOUFO-NONMEMBER

We release VideoUFO-NonMember as the non-member counterpart to VideoUFO, intended for any T2V model trained on the VideoUFO dataset. Below, we outline the collection procedure, statistics, and examples of this dataset.

C.1 COLLECTION PIPELINE

Following Wang & Yang (2026), the collection pipeline comprises five steps: (1) clustering topics from the VidProM prompt dataset (Wang & Yang, 2024); (2) extracting permissible videos under the Creative Commons license from YouTube (using the official API); (3) segmenting extracted videos into clips; (4) captioning each clip; and (5) verifying that the clips correspond to desired topics. Wang & Yang (2026) have not released the official code, but did provide enough detail to enable us to reconstruct it. Step 1 did not have to be reconstructed since we were able to use topics from the released VideoUFO dataset directly.

C.1.1 TOPICS

We use the $1{,}289$ topics from the original VideoUFO metadata as search queries. We view the topic distribution as one of the critical attributes along which our non-member set needs to match the member set.

C.1.2 RETRIEVAL

We search for videos on YouTube with the Creative Commons license filter, mirroring the licensing constraint of VideoUFO. Unlike Wang & Yang (2026), we use the public search interface rather than the official YouTube Data API, namely the YouTube results page with the license and upload-date filters applied, which we query and download from using `yt-dlp`.

C.1.3 SEGMENTATION

While the VideoUFO documentation does not describe its segmentation policy, we were able to deduce it from the released dataset: clip durations have a floor at exactly $2.00\,\mathrm{s}$, a median of $5.23\,\mathrm{s}$, and a long right tail, with approximately 3.2 clips retained per source video. We reproduce this with scene detection at a $2.00\,\mathrm{s}$ minimum and retain at most two clips per source video, which is below the rate of VideoUFO.

C.1.4 ENCODING

The clips in VideoUFO are transcoded (median 640×360, MPEG-4, $\approx$600 kbps). Each clip gathered by our pipeline is hence assigned an empirical VideoUFO clip profile (namely, width, height, frame rate, and bitrate) and encoded to it directly from the retrieved video using scale-to-cover followed by center cropping with FFmpeg. Member clips are not re-encoded, and so they remain bit-identical to the data the model was trained on, unless an undocumented step was taken during pre-processing by the authors of the model.

C.1.5 CAPTIONS

We caption both the members and the non-members using the identical vision-language model (VLM), namely Qwen2.5-VL-7B-Instruct (Bai et al., 2025). This mirrors the proxy-captioner protocol of Wang et al. (2026), adopted to better simulate a real-world scenario. VideoUFO itself uses the captioning model of Panda-70M (Chen et al., 2024b) for its brief captions and Qwen2-VL-7B (Wang et al., 2024a) within the pipeline of Open-Sora-Plan (Lin et al., 2024) for its detailed captions. We re-caption both sets with a single model rather than reproducing these pipelines for the non-members only, so that no difference between the captions of the two sets can arise from the captioner. This matters, since with each set keeping its own captioner, a TF-IDF classifier on caption text alone separates the two sets at an AUROC of 0.990.

C.2 Guaranteeing Non-Membership

We apply two filters to guarantee non-membership. First, we gather the $345{,}437$ distinct YouTube identifiers underlying VideoUFO and exclude any retrieved video whose identifier appears among them. Second, we retain only videos uploaded after June 1, 2025, safely after the release of VideoUFO. The second filter is not redundant, since identifier exclusion alone admits videos that the VideoUFO pipeline encountered and rejected, which are systematically worse than those it accepted. Searching with the license filter alone, 10.2% of retrieved videos fall within the identifier set of VideoUFO, whereas adding the date filter reduces this to 0%.

C.3 Pairing

Each non-member clip is matched to a member clip, minimizing a cost that combines standardized low-level statistics (e.g., duration, resolution, motion, high-frequency energy, and color) with distances in the DINOv2 (Oquab et al., 2023) and CLIP (Radford et al., 2021) embedding spaces, subject to calipers on the standardized differences. Candidates are restricted to blocks of similar content, obtained by k-means clustering in the CLIP embedding space, since blocks defined by topic proved too small to admit good matches. Pairs are assigned globally by the Hungarian algorithm (Kuhn, 1955), and the poorest-matching fraction is discarded. The 500 evaluation pairs are drawn at random from the resulting $4{,}222$.

C.4 Collection Statistics

Table 4: **Collection Statistics of VideoUFO-NonMember.**

Source videos	10,007
Topics covered	1,286 of 1,289
Clips	18,387
Matched pairs	4,222
Evaluation pairs	500
Overlap with VideoUFO identifiers	0

Figure C.4.1 compares the non-member set against VideoUFO along clip duration, orientation, file size, and topic coverage.

C.5 Validation

To validate that our non-member set comes from a reasonably similar distribution and does not allow for blind attacks, we perform three tests.

Blind classifiers. We first train classifiers that do not query the target model but instead only look at the properties of the candidate clips, as reported in Table 5. Since a family used to construct the split no longer provides an honest estimate, we also report a family held out of the matching.

Table 5: **Blind Classifiers on VideoUFO-NonMember.** AUROC on the 500 evaluation pairs. † marks a family used during matching, whose value is hence optimistically biased.

Blind family	AUROC
Encoding statistics†	0.531
DINOv2-large†	0.540
CLIP-bigG†	0.610
Spectral profile (held out)	0.639
VideoUFO vs. raw Panda-70M	0.996
VideoUFO vs. encoding-matched Panda-70M	0.767

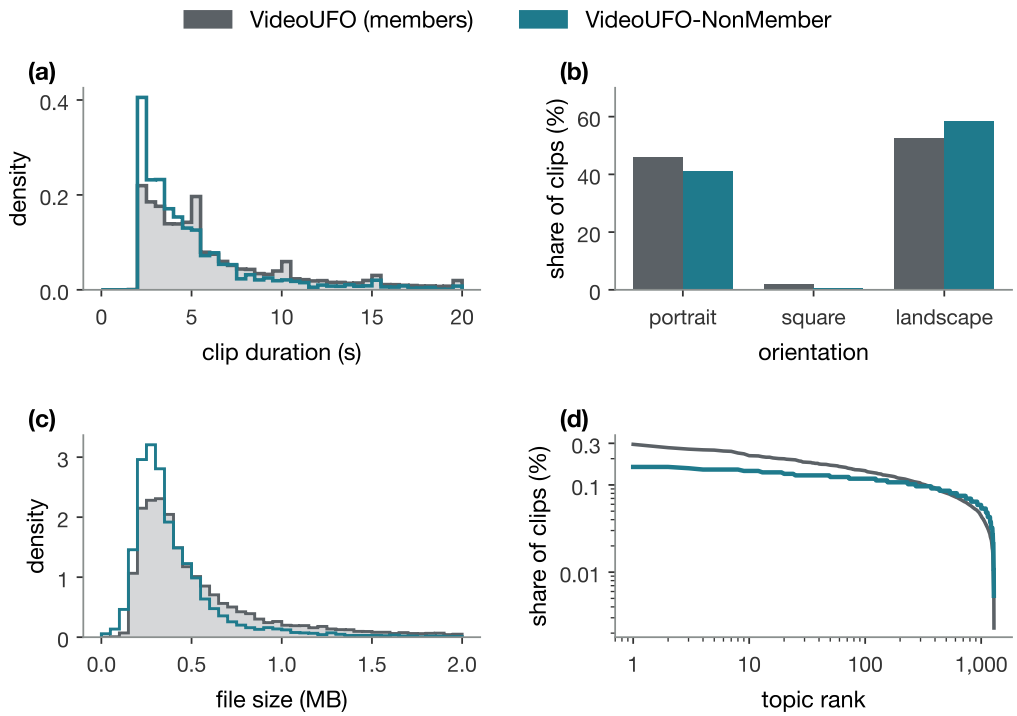

Figure C.4.1: **VideoUFO-NonMember Against VideoUFO.** (a) Clip duration. (b) Share of portrait, square, and landscape clips. (c) File size. (d) Share of clips per topic, ranked from the most to the least common topic. Panels (a) and (d) use the full VideoUFO metadata, whereas panels (b) and (c) use the pool of 21,940 member clips from which the pairs are matched.

Cross-split transfer. Pairing cancels differences between matched pairs but not a constant offset shared by every collected clip, which within-split cross-validation cannot see. Training a blind classifier on one independently collected split and applying it to another yields an AUROC of 0.54–0.62, and so a small pipeline signature survives per-split matching.

Reference-model control. We train a model of identical architecture on Panda-70M (Chen et al., 2024b), excluding any clip whose source video appears in VideoUFO. On the 500 evaluation pairs its loss correlates with that of the target at 0.978, yet separates members from non-members at 0.513 with the plain loss and 0.502 with the reversal gap. Since both models score identical clips through an identical pipeline, this excludes our collection, encoding, and captioning as the source of any signal measured on the target.

C.6 Limitations

Our retrieval uses the public search interface rather than the official API, and the topic-verification and quality-filtering stages of VideoUFO are approximated rather than replicated, since neither is documented. Our sources are also limited to at most 480p, whereas VideoUFO evidently downscaled from higher-resolution originals, which leaves a residual in high-frequency energy that the calipers only partially remove. Finally, matching against a set of feature families does not guarantee that a further family is equalized, as the held-out spectral profile at 0.639 shows.

C.7 Qualitative Examples

Eight clips drawn at random (seed 0) from the non-member set, with the caption produced by the collection pipeline beneath each. Frames are sampled at uniform intervals across the interior of each clip.



Figure C.7.1: The bride and groom stand together in front of a red barn backdrop adorned with flowers and lanterns, holding hands and smiling at the camera.



Figure C.7.2: A large ant struggles with spider silk on a mossy log, its mandibles gripping the strands tightly.



Figure C.7.3: A woman in a suit walks through a dimly lit hallway, surrounded by people, as the ”Morning Show” season 4 logo appears on screen.



Figure C.7.4: Pepe the Frog characters play volleyball in a virtual arena with a scoreboard showing Red team leading Blue.



Figure C.7.5: A man walks down a city street past vintage cars and storefronts, heading towards the camera.



Figure C.7.6: A technician from the Ministry of Agriculture uses a computer displaying weather data to participate in a training on PICSA (Participatory Integrated Climate Services for Agriculture).



Figure C.7.7: Stargazer stands amidst dense foliage, gazing upward at the sky, surrounded by lush greenery.



Figure C.7.8: A soldier in camouflage gear and helmet is crouching on a rooftop, scanning the area with a rifle, under a clear sky.

D Additional Results

This appendix reports the analyses summarized in Section 7.2.

D.1 Denoising Difficulty

Comparing each member against the non-member it was matched to, the member has the lower denoising loss in 57.8% [53.6, 62.2] of pairs. This shows that membership does shift the loss. However, this comparison is not an attack, since it requires the pairing, which an adversary does not have. Figure D.1.1 shows the same comparison for the loss and for the reversal gap.

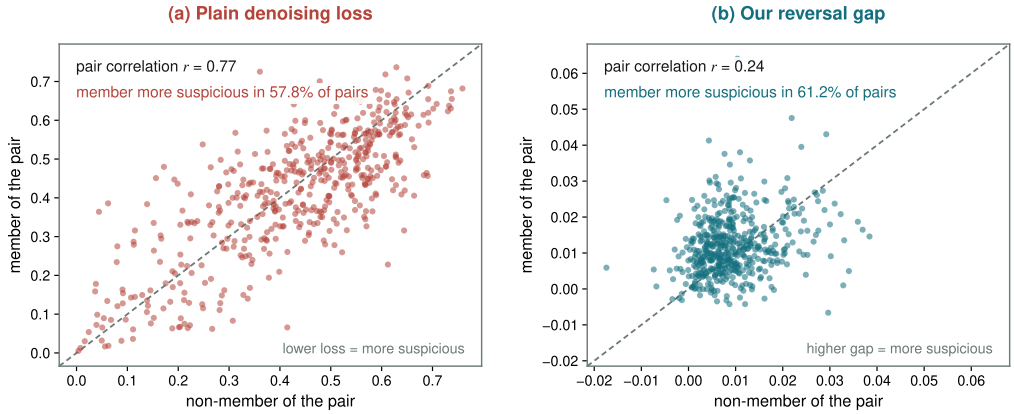

Figure D.1.1: **Denoising Difficulty Reliance.** Each dot is a matched pair, with the non-member on the horizontal axis and the member on the vertical. (a) The plain denoising loss is strongly correlated within pairs, since it largely reflects how hard a clip is to denoise. (b) The reversal gap removes most of this dependence.

D.2 Reference-Model Calibration

We calibrate the attacks against the reference model of Table 3. The losses of the target and of the reference model correlate at 0.978 on our split. Subtracting the loss of the reference model lifts the denoising loss from 0.533 to 0.586 but leaves the reversal gap essentially unchanged, at 0.608 before and 0.609 after. This shows that the reversal gap removes the per-clip difficulty that reference calibration removes, without requiring a second model trained on disjoint data.

D.3 Where Is the Signal Concentrated?

Depth. We next examine where in the network the signal of the read-out lies. Figure D.3.1 shows the AUROC of the read-out when restricted to a single transformer block at a single probed timestep. The first block reaches 0.617, the signal rises through the first ten blocks, peaks at block 13 with 0.731 at $t = 15$, and remains between 0.67 and 0.72 in the later blocks. The timestep $t = 150$ is below the other two in nearly every block. This shows that membership is a property of the representation the network builds rather than of its input.

Redundancy across timesteps and blocks. Restricted to a single timestep, the read-out reaches 0.724 at $t = 15$, 0.717 at $t = 50$, and 0.698 at $t = 150$, against 0.716 with all three. Restricted to groups of seven consecutive blocks, it reaches 0.649 for blocks 0–6, 0.720 for blocks 7–13, 0.727 for blocks 14–20, and 0.699 for blocks 21–27, against 0.716 with all 28. This means that a single timestep and a single group of mid-network blocks suffice, which reduces the cost of the read-out from 6 queries to 2.

Timesteps and frames. For the reversal gap, separation decreases with the timestep, from 0.660 at $t = 50$ to 0.574 at $t = 450$. A denser probe at $t \in \{1, 5, 15, 35, 75\}$ peaks around $t \approx 50$–75

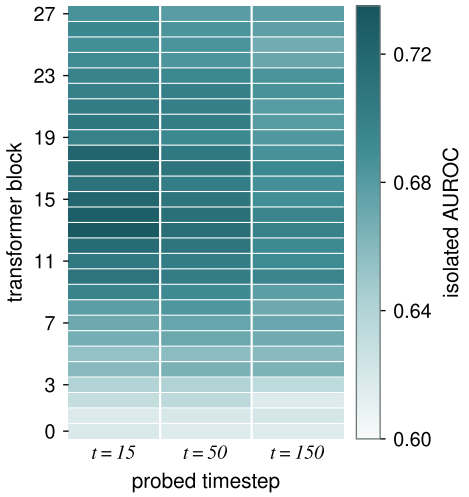

Figure D.3.1: **Signal Strength Localization.** AUROC of the representation read-out using a single transformer block at a single timestep. The signal peaks mid-network and is consistent across timesteps.

and falls to 0.610 at $t = 1$. Across the sixteen frames, single-frame AUROC ranges from 0.52 to 0.63, and averaging all frames at 0.668 outperforms selecting the top-K frames at 0.644, the bottom-K frames at 0.606, and the first frame alone at 0.532. This means that subsampling frames, as keyframe-based black-box attacks do, discards signal.

Estimation. Averaging the reversal gap over noise draws lifts its AUROC from 0.628 at $K = 1$ to 0.675 at $K = 8$, with an extrapolated ceiling of 0.684, and adding a second temporal window adds a further 0.029. A variance decomposition bounds this statistic at 0.728 regardless of further averaging. Averaging over horizontal flips, which mirrors the training augmentation, lowers the AUROC by 0.010 and is not used.

Attack model. On identical features, a cross-fitted logistic model reaches 0.717, gradient boosting 0.715, and a two-layer MLP 0.678. The AUROC over the number of principal components is 0.654, 0.671, 0.703, 0.715, 0.711, and 0.703 for 16 through 256 components, and recording all 28 blocks rather than 8 changes the result by 0.002. This shows that a more expressive attack model does not improve the read-out.

Sample size. Subsampling the matched pairs, the read-out reaches 0.610 with 100 pairs, 0.663 with 200, 0.688 with 300, 0.708 with 400, and 0.710 with 500. This shows that the performance saturates by 400 pairs, and so the 500 pairs of our split are sufficient.

Conditioning. Every member can be conditioned on its original VideoUFO caption, which permits a paired test of the caption. Replacing the proxy caption with the original changes the member loss by $+0.00096$ and the non-member loss by $+0.00095$, and a statistic built on this change reaches 0.495. This shows that knowing the exact training caption does not help the attack, in contrast to the image-model finding that specific conditioning drives replication (Somepalli et al., 2023b).

Negative results. The following probes reach chance: input-space curvature under random perturbation at 0.488; weight-space sharpness under random parameter perturbation at $\sigma \in \{0.01, 0.02, 0.05\}$ at 0.497; masked-region reconstruction at 0.535 on its own and 0.494 once calibrated against a base model; and x_0-prediction consistency across noise draws and timesteps at 0.49–0.52. Frame shuffling, repetition, static replacement, brightness, and blur perturbations all underperform time reversal. The only exception is the mutual alignment of per-clip gradients across noise draws, which reaches 0.566 but adds nothing in combination.

D.4 Which Members Are Identified?

We next examine whether properties of a member predict whether it is identified. Table 6 reports the AUROC of the read-out within subgroups of members, and Figure D.4.1 shows the rank correlation between each of nine clip properties and the per-pair margin of the read-out. Every interval in Figure D.4.1(a) includes zero. A cross-fitted model over all nine properties predicts which pairs the read-out gets right at an AUROC of 0.555, against a permutation null of 0.497 ± 0.037 ($p = 0.062$). Figure D.4.1(b) shows that our split can detect a difference of 0.12 in the rate at which the member outscores its pair between the top and bottom half of a property, and every observed difference is below it. This means that no measured property predicts which members are identified, while smaller effects cannot be ruled out. Appendix E shows the clips on which the read-out is most confidently wrong.

Table 6: **Attack Performance by Member Subgroup.** AUROC of the representation read-out restricted to the pairs whose member falls in each subgroup, with 95% bootstrap intervals over pairs. Topic density counts the training clips sharing a member's topic. Since every member was seen for the same number of epochs, it measures how much similar content surrounded a clip rather than how often the clip itself was seen.

Member subgroup	Pairs	AUROC
Densest topics (top 10%)	50	0.684 [0.57, 0.79]
Sparsest topics (bottom 10%)	50	0.759 [0.66, 0.86]
Several clips per source	453	0.721 [0.69, 0.76]
One clip per source	47	0.679 [0.57, 0.77]

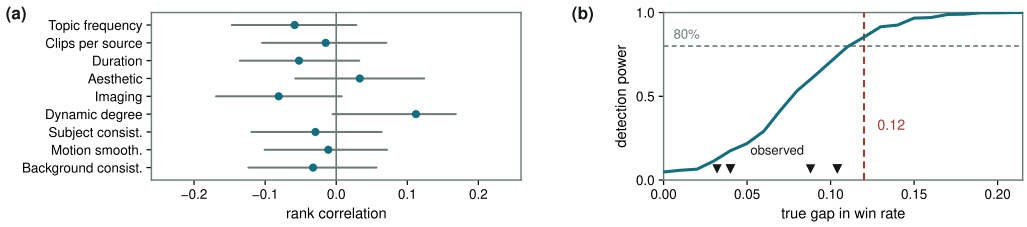

Figure D.4.1: **Blind Verification.** (a) Rank correlation between each clip property and the per-pair margin of the read-out, with 95% bootstrap intervals. (b) Power to detect a difference between the top and bottom half of a property, with the observed differences marked along the axis.

D.5 Robustness to Post-Processing

Finally, we examine whether the read-out survives the candidate being post-processed before it reaches the adversary. Table 7 reports the read-out after re-encoding every candidate at CRF 28 and at CRF 38, and after a half-resolution round trip. Re-encoding at CRF 28 lowers the blind-adjusted AUROC from 0.692 to 0.665, the half-resolution round trip to 0.665, and re-encoding at CRF 38 to 0.611. Every interval excludes 0.5. This shows that the read-out does not require the original file, and that re-encoding is not an effective defense against it.

Table 7: **Robustness to Post-Processing.** The read-out after every candidate, member and non-member alike, is re-encoded or downscaled and upscaled before it is scored. The blind block for the adjustment is computed on the original clips, and so the adjusted column is slightly optimistic.

Condition	AUROC	Blind-adj. AUROC
Original	**0.717**	**0.692 [0.66, 0.72]**
Re-encoded, CRF 28	0.697	0.665 [0.63, 0.70]
Re-encoded, CRF 38	0.647	0.611 [0.58, 0.64]
Half-res. round trip	0.693	0.665 [0.63, 0.70]

E Failure Cases

We investigate which videos are misclassified by the strongest attack of Table 1, the representation read-out, focusing on the cases where it is most confidently wrong. To this end, we rank all $1{,}000$ clips by attack score and select the four lowest-ranked members and the four highest-ranked non-members. Frames are sampled at uniform intervals across each clip, and the caption beneath each is the one used for conditioning. The explanations we provide are suspected rather than causally verified.

E.1 Missed Members

The following examples are members that the attack misses. Three of the four are not landscape videos but portrait or square videos surrounded by black padding. We suspect that this dilutes the read-out, since the pooled block statistics include the padded regions, which are particularly easy to denoise regardless of membership. These clips are also slow-moving and nearly static, which leaves the attack little clip-specific content to work with. The skydiving clip in Figure E.1.3 is an exception to both, and we do not have a strong hypothesis for why it was missed.



Figure E.1.1: Revanth Kanakam speaks confidently into a microphone at Revanth Kanakam Speaking Academy, promoting communication skills and life skills for great success.



Figure E.1.2: The dragon's head is shown in close-up, its eyes glowing and mouth open, revealing sharp teeth, against a dark background.



Figure E.1.3: Two men in skydiving gear prepare for a jump on a clear day in an open desert field.

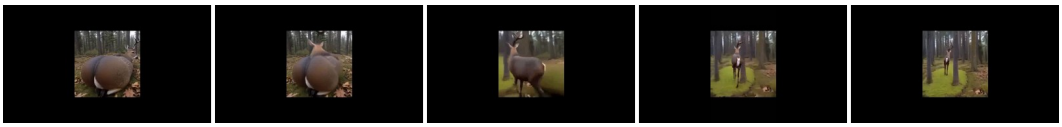

Figure E.1.4: A deer stands in a forest, its back to the camera, then turns and walks away through the trees.

E.2 Flagged Non-Members

The following examples are non-members that the attack flags as members. Two of the four are computer-generated, and we suspect that they are simply easy for the model to denoise given their clean backgrounds and flat color. This connects to the difficulty confound discussed in Section 6.2, which the read-out does not cancel since it does not form a within-clip contrast. The product clip is also unusually long at $412\,\mathrm{s}$, which is atypical of both sets.



Figure E.2.1: A tattooed hand holds a ”Mr. Fog Aura” vape product box against a plain white background, pointing to details on the packaging.



Figure E.2.2: A dancer in a blue shirt and jeans leads a group through dynamic movements across a dance studio floor, with other dancers practicing in the background.



Figure E.2.3: A figure descends through a dark, conical shaft illuminated by a bright light at the top.



Figure E.2.4: The golden robot character is encased in ice as another robot approaches, set against a plain white background.