Black-Box Membership Inference Attacks for Video Training Data in Multimodal Large Language Models

Jinrui Wang[1] **Zhenfeng Gao**[2] **Wendan Wang**[1] **Huili Wang**[3] **Zichen Qin**[1]
Linjie Zhu[2] **Hongke Fu**[2] **Shangguang Wang**[1] **Tao Qi**[1] *
[1] Beijing University of Posts and Telecommunications
[2] Sangfor Technologies, Inc. [3] Tsinghua University

Abstract

The increasing use of video data in training multimodal large language models (MLLMs) raises significant concerns on privacy leakage and copyright violations, highlighting the need for detecting improperly used training videos through membership inference attacks (MIAs). Most existing video MIA methods assess model memorization of key semantic concepts within a video (e.g., the name of a well-known movie character). However, such concepts usually appear repeatedly throughout the training corpus, and memorization of them does not constitute reliable evidence that a specific video was used during training. Besides, while some methods mitigate this limitation by capturing relationships between frames, they require a model logit-accessible setting and are impractical in realistic black-box scenarios. To address these challenges, we propose a black-box MIA framework, named VideoMIA, that can provide reliable evidence of specific video data usage for training MLLMs. The key of our method is to leverage temporal dependencies across video frames to evaluate the model's memorization of sequential dynamics within the video data, which cannot be inferred solely from general world knowledge or individual image data. The results across ten MLLMs and four benchmarks demonstrate that our method consistently achieves superior performance over all baselines in black-box evaluation settings. Code is available in https://github.com/jinruiwang258/VideoMIA.

1 Introduction

Video data has become a critical resource for training multimodal large language models (MLLMs) (Liu et al., 2023; Achiam et al., 2023; Chen et al., 2024; Liu et al., 2024; Wang et al., 2024a,b; Zhang et al., 2025). However, its widespread use also raises significant concerns regarding the unauthorized utilization of privacy-sensitive and copyrighted content (Duan et al., 2024). In this context, membership inference attacks (MIAs) offer a principled methodology for identifying potential data inclusion of MLLMs (Hu et al., 2022; Golchin and Surdeanu, 2024; Oren et al., 2024). Existing MIA methods for MLLMs are primarily developed for image data, typically relying on the model confidence on target images. While these methods can, in principle, be extended to video data by applying them to individual frames, memorization of isolated frames cannot serve as reliable evidence that the model was trained on the entire video. This is because the model may have been exposed to only a subset of similar or extracted frames during training, rather than the full temporal sequence, leading to false indications of video-level membership.

Some recent studies have attempted to adapt MIAs to video data by examining the model memorization of key semantic concepts within videos.

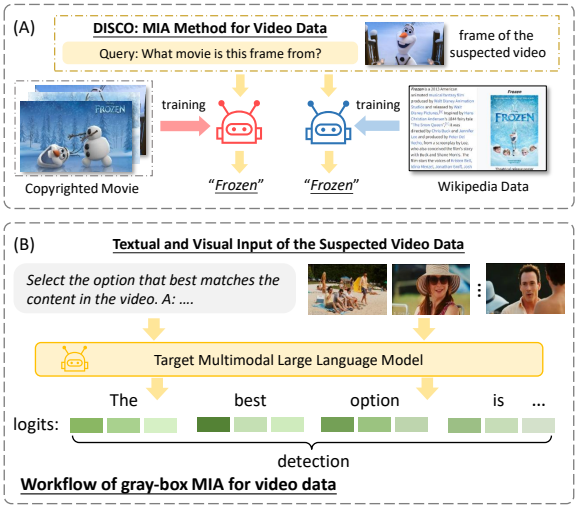

Figure 1: MIA methods for video data. (A) Methods that focus on semantic concepts from individual frames, leading to unreliable assessments. (B) Gray-box MIA methods, which assume feature access and is therefore inapplicable in black-box scenarios.

*Corresponding Author. (Email: taoqi.qt@gmail.com)

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 39235–39248
July 2-7, 2026

For instance, DIS-CO (Duarte et al., 2025) prompts the model to identify the title of a movie based on selected frames, classifying videos with correct model responses as membership samples. However, such semantic concepts usually appear broadly across the image or text corpora used to train MLLMs, thereby reducing the reliability of using concept-level memorization as an indicator of video-level training inclusion. As illustrated in Figure 1(A), a model can correctly recognize a movie title given a frame from the corresponding film, leading DIS-CO to confidently classify the video as a member of the training set. In reality, the model's correct response may originate from related information contained in its pre-training corpus (e.g., Wikipedia articles), which can lead to ambiguous or spurious attributions of model memorization that are unrelated to actual training exposure to the video. Besides, to address this limitation, Vid-SME (Li et al., 2025), a gray-box MIA for video data (Figure 1(B)), leverages Sharma-Mittal entropy (Esteban and Morales, 1995) to quantify the model's confidence across consecutive video frames. Nonetheless, most commercial MLLMs are fully closed-source, providing no access to the intermediate inference features these methods require, thereby rendering these gray-box methods impractical in real-world scenarios.

In this paper, we propose a temporal-aware black-box membership inference framework (VideoMIA), that leverages the intrinsic sequential dependencies within videos to more reliably ascertain their inclusion in model training. The central idea is to construct temporal visual reasoning tasks that can be correctly solved only by memorizing inter-frame temporal relationships, rather than relying on general world knowledge or isolated frame cues. Based on this principle, we design two temporal reasoning tasks with increasing temporal complexity. Given several contextual frames, the model is first required to identify the most recent frame from multiple candidates. Besides, the model is provided with preceding frames and asked to reorder a set of shuffled subsequent frames according to their original temporal order. Videos that yield consistently high success rates on these tasks are then identified as potential training members of the target MLLM. Moreover, our method forms a generalizable framework that can be extended with additional temporal reasoning tasks over suspected video data, further strengthening the effectiveness of membership inference attacks.

Prior work relies on a single film-based benchmark for video-level membership inference attacks. However, membership detection on this dataset can be achieved through concept-level memorization alone, which limits its ability to support comprehensive and reliable evaluation. To address this limitation, we construct three new benchmark datasets. The first two are derived from open-source video LLMs with clearly documented training corpora. The third, YoutubeTection, is built from YouTube video data and enables broader assessment of MIAs against closed-source multimodal LLMs. Based on the four benchmark datasets, we evaluate VideoMIA and other baselines on ten MLLMs. Results show that existing video MIA methods degrade to near-random performance under our YouTubeTection, whereas our method achieves significant and consistent improvements in detection accuracy across different models. Further analysis also validates the robustness of the proposed framework. Our work makes the following contributions:

(1) To the best of our knowledge, we are the first to reveal the vulnerability of existing image-based MI attacks when applied to video data, and to exploit frame temporal dependencies for reliable detection.

(2) We propose a temporal-aware black-box MI framework capable of identifying training videos solely based on model textual outputs.

(3) We release three new benchmarks and demonstrate through extensive experiments that our method consistently outperforms strong baselines across diverse settings.

2 Related Work

2.1 Image Membership Inference Attacks

Membership inference attacks (Shokri et al., 2017; Long et al., 2018; Song et al., 2019; Song and Mittal, 2021; Carlini et al., 2022; Li et al., 2024b,b,a; Zarifzadeh et al., 2024; Kokhlikyan et al., 2024; Wang et al., 2025; Li et al., 2025) aim to determine whether a data instance was used to train a model. Recent works on image MIAs for MLLMs mostly focus on assessing the model's overconfidence on specific data to determine whether it was included in the training set (Kokhlikyan et al., 2024; Jayaraman et al., 2024; Pinto et al., 2024; Li et al., 2024b; Duarte et al., 2025; Li et al., 2025; Hu et al., 2025). For example, previous research (Li et al., 2024b) introduced the MaxRényi-K% metric, which enables image membership inference by analyzing the output logits correspond-

ing to the model's image-specific segments. However, directly applying these image-based methods to videos is insufficient, as they can only reveal whether individual frames have been seen during training, rather than the entire video.

2.2 Video Membership Inference Attacks

Compared with studies on image data, membership inference attacks on videos remain an underexplored challenge. Most existing approaches are built upon techniques originally developed for text or image MIAs, relying on memory of coarse-grained concepts to perform membership inference (Li et al., 2024b; Duarte et al., 2025; Li et al., 2025). For example, DIS-CO (Duarte et al., 2025) detects copyrighted movie content by querying the models with specific frames from targeted videos and analyzing their free-form text responses. Nevertheless, the frame-static methods are inherently unreliable, as models may answer the question correctly by relying on general knowledge acquired from other textual or visual training corpora. As a result, relying on coarse-grained concepts alone provides an unreliable assessment of a video's inclusion in the training set. Vid-SME (Li et al., 2025) derives membership scores for video MIA by computing the Sharma-Mittal entropy (Esteban and Morales, 1995) of the model's output logits. However, a key challenge of these grey-box approaches is their reliance on access to token probability distributions, restricting their applicability to black-box models such as Gemini (Team et al., 2024). In this paper, we leverage fine-grained sequential dependencies within videos to perform reliable membership inference through model-generated text, which is practical in real-world black-box settings.

3 Method

3.1 Problem Setting

A video frame sequence is denoted as $F_{1:T}$, where T represents the number of frames. Given an MLLM, the goal of the attacker is to determine whether a video was used during the model training. We assume a black-box scenario in which the adversary can query the MLLM using the video frames $F_{1:T}$ and instruction context X_{ins}, and is only allowed to access the model-generated text.

3.2 Motivation for Overall Framework

The core idea underlying existing Video MIA methods is to examine the model memorization of semantic concepts within videos. However, as discussed in the introduction, these approaches face significant challenges as semantic concepts usually appear broadly across the image or text corpora used to train MLLMs, thereby reducing the reliability of using concept-level memorization as an indicator of video-level training inclusion. For current MLLMs applied to video understanding, including LLaVA-NeXT-Video (Zhang et al., 2025), the training tasks are designed to understand the temporal dynamics of input video frames. Specifically, the model takes densely sampled video frames as input, forming a sequence of frame embeddings $v_{1:T}$ that preserve the continuous and temporal evolution of scenes and events. Then, the model generates fine-grained natural language outputs (e.g., descriptions or answers) for the video clip, describing the sequence of events over time or providing answers to temporal reasoning questions. The training objective is to maximize the likelihood of each target text token conditioned on both the video embeddings and preceding text tokens:

$$\mathcal{L} = -\sum_{m=1}^{M} \log p(x_m \mid v_{1:T}, x_{<m}), \qquad (1)$$

where x_m denotes a text token. This objective encourages the model to leverage inter-frame information for video comprehension.

Building on the video understanding training paradigms of target models, we exploit the intrinsic sequential dependencies within videos to perform membership inference on video data. To this end, we propose a framework named VideoMIA that incorporates temporal reasoning tasks (Figure 2). The first task is Nearest-Frame Identification (FrameIdentify), which asks the model to identify the earliest frame among multiple candidates. The second is Temporal Frame Ranking (TempRank), which requires the model to reorder shuffled frames into their original temporal sequence. These tasks represent different levels of temporal reasoning difficulty, with TempRank being more challenging than FrameIdentify. This design is motivated by the observation that MLLMs acquire varying memory depths during training. For example, pretraining results in deeper memory, whereas instruction fine-tuning yields shallower memory. Our method establishes a generalizable framework that can be extended to incorporate additional temporal reasoning tasks. Considering membership inference, videos with the highest temporal reason-

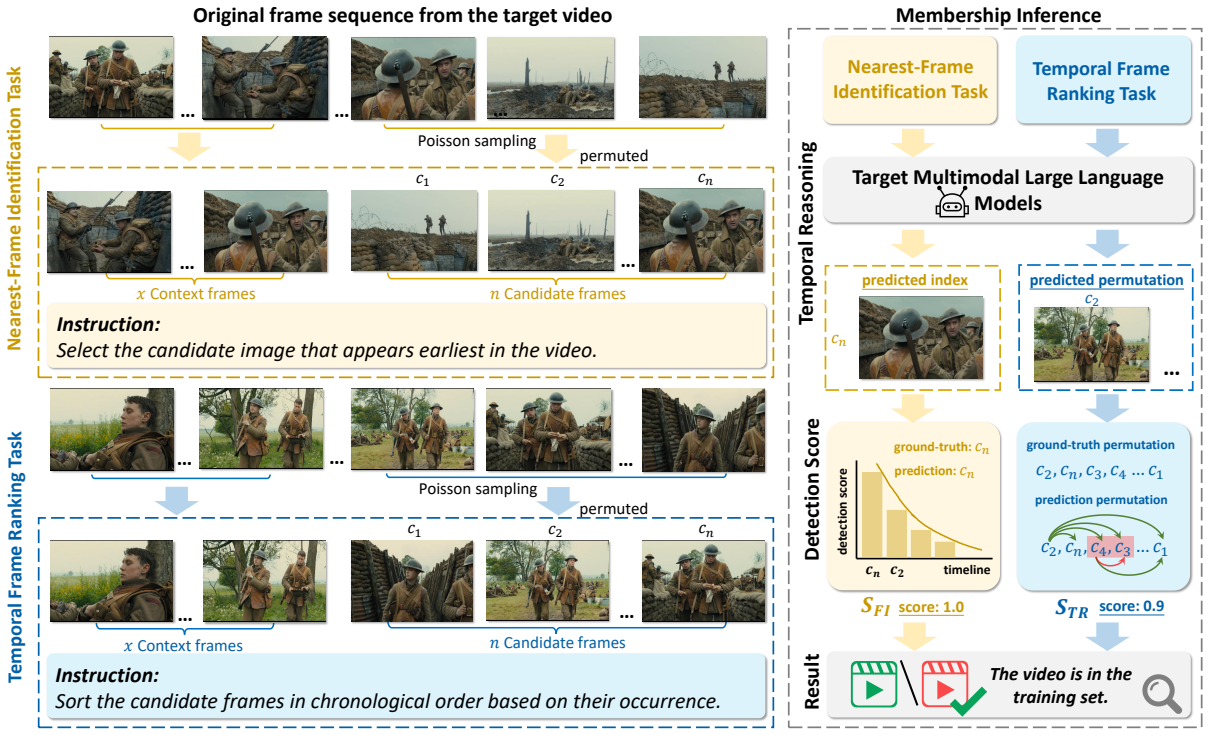

Figure 2: VideoMIA is the first membership inference framework for detecting training videos of black-box MLLMs. It leverages temporal-reasoning tasks across video frames to extract fine-grained and reliable membership signals.

ing success rates are regarded as potential training members of the target MLLM. Through the framework, we can evaluate a model's memorization at the video level, enabling robust black-box MIA.

3.3 Nearest-Frame Identification Task

To assess a model's behavioral differences on member and non-member videos, we propose the Nearest-Frame Identification task, in which the model is asked to identify the earliest frame among multiple candidates. To ensure stable evaluation, we use N test instances per video. For notational simplicity, we focus on a single instance in the following description. Specifically, the input test sample is formulated as $f \oplus X_{\text{ins}}$,and $f = \{F_{p:p+x-1}, C\}$ contains x consecutive context frames $F_{p:p+x-1}$ that start from a uniformly sampled position p, together with n candidate frames $C = \{c_1, \dots c_n\}$. The number of context frames x is chosen to balance the difficulty of the task. Too few context frames render the task overly difficult, whereas too many make it overly easy. In both cases, the discriminative performance of MIA is reduced, as the model's responses to member and non-member videos may become indistinguishable. The candidate set C consists of n consecutive frames sampled after a short random gap following the context frames. To simulate independently arriving frames, we employ Poisson sampling to collect candidate frames, and their order is subsequently randomly permuted. X_{ins} is the instruction that prompts the MLLM to identify the earliest appearing frame among the candidates. Given each test sample $f \oplus X_{\text{ins}}$, the model predicts the index of the earliest frame among the candidates, denoted as $\hat{j}$. For membership inference, we design a stepwise, exponentially decaying scoring scheme that assigns higher scores to predictions closer to the genuine earliest frame. Specifically, we define the detection score of FrameIdentify as S_{FI}, where $S_{\text{FI}} = 2^{-(j-1)}$ for $j < n$, and $S_{\text{FI}} = 0$ for $j = n$, with j is the ground-truth rank position corresponding to the predicted index $\hat{j}$. The S_{FI} score reflects the model's memorization of temporal dependencies within the video, serving as evidence for MIA.

3.4 Temporal Frame Ranking Task

Beyond FrameIdentify, we introduce a more challenging temporal task, Temporal Frame Ranking, which requires shuffled frames to be reordered to match their original temporal sequence. The Temporal Frame Ranking Task adopts the same sampling and construction procedure for frame input as the Nearest-Frame Identification Task. Here, X_{ins} instructs the MLLM to sort the candidate frames C in chronological order. Formally, we define the

time-ordered permutation over C as π. The target model is required to predict the permutation π for each test sample. To capture the detection score for membership inference, let $\hat{\pi}$ and π denote the predicted and ground-truth time-ordered permutations over the candidate frames, respectively. The detection score of TempRank S_{TR}, is computed as the pairwise AUC (Hand and Till, 2001) between the two ranking orders: $S_{\mathrm{TR}} = \mathrm{AUC}(\hat{\pi}, \pi)$, where $S_{\mathrm{TR}} \in [0, 1]$. The higher values of S_{TR} indicate stronger consistency between the predicted and ground-truth temporal orders, thereby reflecting a greater degree of the target model's memorization of inter-frame relationships.

3.5 Membership Inference Attack

To ensure stable results, as mentioned earlier, we perform N sampling rounds per video to construct test samples. This yields N individual S_{FI} scores from FrameIdentify, which are then averaged to obtain the detection score for each video. Similarly, the detection score of TempRank is obtained by averaging its N computed S_{TR} scores. We linearly combine the average detection scores from FrameIdentify and TempRank to obtain a unified detection score of VideoMIA. A higher score reflects stronger video-specific memorization of the MLLM, providing evidence of data membership. VideoMIA provides a generalizable framework capable of incorporating additional temporal reasoning tasks to enhance membership inference performance. Moreover, our proposed temporal visual reasoning tasks can serve as modular components and be integrated with existing image-based MIA methods, such as MaxkRényi (Li et al., 2024b). The integration is achieved by combining the detection scores from VideoMIA with those from image-based MIA methods.

4 Experiment

4.1 VideoMIA Benchmark Construction

Currently, the target video benchmarks available for conducting MIA on MLLMs are limited, with MovieTection (Duarte et al., 2025) being the only one. MovieTection is a question-answering dataset based on movie frames, where the task is to identify the movie corresponding to a given frame. However, MovieTection is relatively unreliable, despite being used in some studies. This dataset primarily focuses on semantic concepts in videos that are largely present in the image or text pretraining corpora of target models. As a result, membership inference attacks on MovieTection may yield unreliable assessments. Therefore, we construct three benchmarks specifically designed for video MIA. The first two are constructed for open-source models, and the third targets closed-source models. For the benchmarks focusing on open-source models, namely VideoInstruct/Video-MME and LLaMA-178 K/Video-MME, construction is based on publicly available datasets. Specifically, members are sampled from the open training datasets VideoInstruct-100K (Maaz et al., 2024) and LLaVA-Video-178K (Zhang et al., 2025), while non-members are drawn from the open-source evaluation dataset Video-MME. The detailed information about these member/non-member sets is in the Appendix B.

To evaluate closed-source models and probe the limitations of existing concept memorization methods, we construct YouTubeTection. YouTubeTection is built from YouTube videos with explicitly specified publication dates and excludes commonly occurring concepts from existing image or text training datasets, providing a more realistic assessment of video MIA. To enable YouTubeTection to be used for evaluating current mainstream multimodal LLMs, we construct clean and suspected subsets according to the models' knowledge cutoff dates. The specific knowledge cutoff dates for multimodal models are provided in the Appendix C. Specifically, videos released in January 2025 or later are considered clean, as they fall outside the knowledge cutoff dates of the models. Popular videos released in 2022 or earlier are treated as suspected data, as they are more likely to have been included in the training datasets of these models. The details of YouTubeTection, including resolution and sampling rate, are provided in the Appendix B. Currently, YouTubeTection comprises frames from 800 videos, ranging from 5 to 40 minutes in length, with plans for future expansion.

4.2 Experimental Setup

Datasets and Models. To comprehensively evaluate attack performance, we apply VideoMIA to 10 target models, including commercial closed-source models. Unless otherwise stated, all experiments adopt a default number ($n = 4$) for candidate frames. We evaluate our method on four open-source models: LLaVA-NeXT-Video-7B (Zhang et al., 2025), LLaVA-NeXT-Video-34B-hf, MiniCPM-V-200K-Video-Finetune (Yin

Method		LLaVA-7B		LLaVA-34B		MiniCPM-Finetune		LLaVA-Qwen2	
		AUC	TPR@5%	AUC	TPR@5%	AUC	TPR@5%	AUC	TPR@5%
Perplexity		0.5176	0.0280	0.5477	0.0480	0.5144	0.0024	0.5067	0.0078
Max_Prob_Gap		0.5052	0.0386	0.5536	0.0287	0.5473	0.0496	0.5550	0.0489
DIS-CO		0.5039	0.0060	0.5105	0.0330	0.5085	0.0270	0.5129	0.0280
Caption		0.5007	0.0040	0.5062	0.0320	0.5009	0.0370	0.5075	0.0250
MCQA		0.5019	0.0040	0.5067	0.0345	0.5124	0.0240	0.5062	0.0360
Min-K%	$K = 0$	0.4943	0.0282	0.5543	0.0068	0.5092	0.0164	0.4486	0.0147
	$K = 10$	0.4951	0.0524	0.5564	0.0343	0.5134	0.0164	0.4558	0.0068
	$K = 100$	0.5014	0.0624	0.4687	0.0339	0.5092	0.0147	0.4569	0.0076
ModRényi	$\alpha = 0.5$	0.5183	0.0468	0.5601	0.0329	0.5392	0.0248	0.5408	0.0048
	$\alpha = 1.0$	0.5164	0.0400	0.5564	0.0387	0.5490	0.0346	0.5018	0.0378
	$\alpha = 2.0$	0.5183	0.0368	0.5612	0.0469	0.5223	0.0567	0.5590	0.0468
MaxRényi ($\alpha = 0.5$)	Max_0%	0.4362	0.0342	0.5034	0.0100	0.5127	0.0447	0.5039	0.0080
	Max_10%	0.5063	0.0500	0.5167	0.0524	0.4352	0.0524	0.4879	0.0137
	Max_100%	0.5063	0.0528	0.5209	0.0457	0.5236	0.0487	0.5462	0.0524
Vid-SME	Min_0%	0.5092	0.0485	0.5283	0.1610	0.4907	0.0635	0.5509	0.0774
	Min_5%	0.5447	0.1068	0.5550	0.1148	0.5234	0.1870	0.5794	0.0971
	Min_30%	0.5496	0.1047	0.5444	0.0884	0.5550	0.1096	0.5725	0.1088
Ours	VideoMIA	0.5608	0.1800	0.5784	0.2000	0.5908	**0.2400**	0.5903	0.1877
	VideoMIA*	**0.5772**	**0.2040**	**0.5875**	**0.2200**	**0.5952**	0.1674	**0.6077**	**0.2430**

Table 1: Performance comparison of different baselines across four open-source models. We report the AUC and the true positive rate under a false positive rate constrained to at most 5% (TPR@5%). Our method yields statistically significant improvements over the strongest baselines, with significance levels of $p \leq 0.001$, across all settings.

et al., 2026), and LLaVA-Video-7B-Qwen2. For each model, we use VideoInstruct/Video-MME or LLaVA-178K/Video-MME as the benchmark, where members are drawn from the model's training set and non-members from its evaluation set. Furthermore, considering the closed-source models, we focus on four models: GPT-4o (Hurst et al., 2024), GPT-4o-Mini, Claude-3.7, Gemini-2.0-flash. Beyond these MLLMs, we also include Qwen2-VL-7B (Wang et al., 2024a) and Qwen2.5-VL-32B, which provide publicly released weights but were trained on datasets that are not publicly available. Notably, different models have distinct knowledge cutoff dates. Accordingly, we filter clean and suspected subsets from MovieTection and YouTube-Tection based on each target model's knowledge cut-off date. As mentioned before, MovieTection exhibits certain biases, which can allow some methods to achieve high membership detection accuracy without truly performing membership inference.

Baselines. We adopt several prior MIAs as baselines, covering both image-based and video-based methods. For image-based MIAs, we include the perplexity (Yeom et al., 2018), and the Min-K% method (Shi et al., 2024). We also consider R'enyi-based approaches, including Max_Prob_Gap, MaxR'enyi-K%, and its modified variant ModRényi (Li et al., 2024b). These Rényi-based methods require access to generated token logits, making them inapplicable to fully closed-source models, except for GPT-4o, which provides top-5 token probability distributions. For video-based MIAs, we include DIS-CO (Duarte et al., 2025) and Vid-SME (Li et al., 2025). In addition to the original VideoMIA, we also consider VideoMIA*, a setting that integrates VideoMIA with the image-MIA, MaxRényi-K%.

Evaluation metric. Following the previous works, we evaluate different MIA methods by their AUC scores and TPR@5%FPR.

4.3 Main Results

We conduct a comparative analysis of different methods across four open-source target MLLMs. As shown in Table 1, the experimental results lead to two main findings. First, across all evaluated models, our proposed VideoMIA consistently outperforms existing MIA approaches in the black-box setting, with significance levels of $p \leq 0.001$. In particular, even when compared with recent baseline methods such as Vid-SME, our approach maintains clear and stable advantages in both AUC and TPR@5% metrics, demonstrating superior and robust effectiveness across different target models. These results validate that a model's memorization of temporal dependencies provides a reliable and discriminative signal for determining whether a video was included in the training data. Sec-

Method	Black-box	Video-level	MovieTection					
			GPT-4o	GPT-4o-Mini	Claude	Gemini	Qwen-7B	Qwen-32B
DIS-CO	✓	✗	0.6328	**0.6460**	**0.6326**	**0.6370**	0.6350	0.6005
Captions	✓	✗	0.5880	0.6260	0.5898	0.5980	0.5978	0.5880
MCQA	✓	✗	0.5560	0.5770	0.6094	0.5425	0.5894	0.5700
Max_Prob_Gap	✗	✗	0.5160	0.5074	-	-	0.6091	0.5778
MaxRényi_0%	✗	✗	0.4892	0.5020	-	-	0.6247	0.5550
Vid-SME	✗	✗	-	-	-	-	0.5194	0.5283
VideoMIA	✓	✓	**0.6468**	0.6071	0.6254	0.5962	0.6068	0.6504
VideoMIA*	✓	✓	0.6264	0.6129	-	-	**0.6712**	**0.6656**
Method	Black-box	Video-level	YouTubeTection					
			GPT-4o	GPT-4o-Mini	Claude	Gemini	Qwen-7B	Qwen-32B
DIS-CO	✓	✗	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
Captions	✓	✗	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
MCQA	✓	✗	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
Max_Prob_Gap	✗	✗	0.5009	0.5218	-	-	0.5373	0.5447
MaxRényi_0%	✗	✗	0.4718	0.5046	-	-	0.5384	0.5060
Vid-SME	✗	✗	-	-	-	-	0.4969	0.5109
VideoMIA	✓	✓	**0.6109**	0.5938	**0.6076**	**0.5722**	**0.5996**	0.6430
VideoMIA*	✓	✓	0.6092	**0.6088**	-	-	**0.6122**	**0.6712**

Table 2: Performance comparison (AUC) across four closed-source and two open-source models on MovieTection and YouTubeTection.

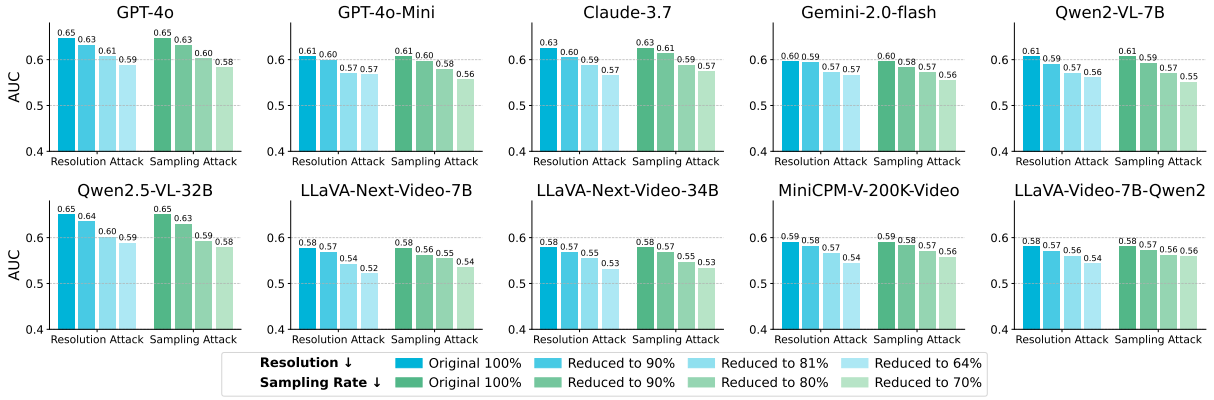

Figure 3: Robustness study. We examine the robustness of VideoMIA under data-processing strategies commonly used in real-world model training, including resolution degradation and frame downsampling, and evaluate their impact on the overall detection performance.

ond, we observe that combining VideoMIA with image-level MIA further improves membership inference performance, achieving the highest AUC and TPR@5% on nearly all benchmarks. This highlights the strong generalization capability of our approach, demonstrating that it can be effectively integrated with other MIA methods to further enhance overall attack performance.

4.4 Results on Closed-Source Models

The results of closed-source models are shown in Table 2. For the baseline method DIS-CO, which is concept-level and performs MIA by measuring model confidence in answering video titles, its performance on YouTubeTection essentially collapses, yielding near-random AUC scores around 0.5. This highlights the poor generalization of this method. On the challenging benchmark of YouTubeTection, our method achieves competitive AUC scores across all models, with our VideoMIA performing the best. In contrast, the image-level baselines (e.g., MaxProbGap and MaxRényi) show limited discriminative capability, underscoring the importance of temporal dynamics in detecting video membership. These results demonstrate that our temporal-aware approach remains effective against frontier closed-source models such as GPT-4o, showing the practicality of our method in real-world scenarios.

4.5 Robustness Study

To evaluate the robustness of our approach, we introduce perturbations to simulate common data-

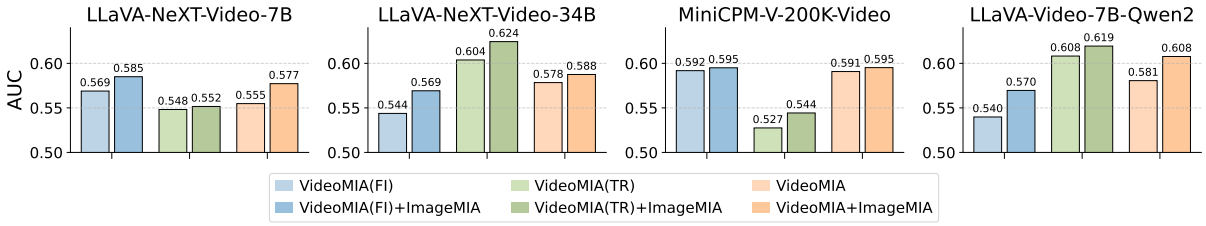

Figure 4: Ablation study of VideoMIA. We evaluate the effectiveness of the FrameIdentify task and the TempRank task, under conditions with and without the enhancement of the image-level MIA method (i.e., MaxRényi).

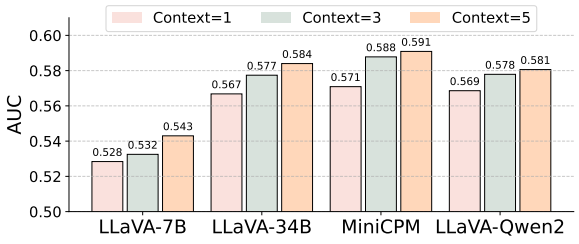

Figure 5: Impact of the frame context quantity x.

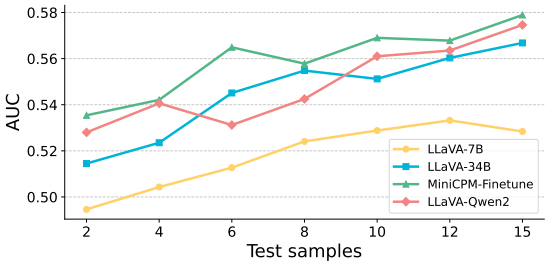

Figure 6: Impact of the probe trial quantity N.

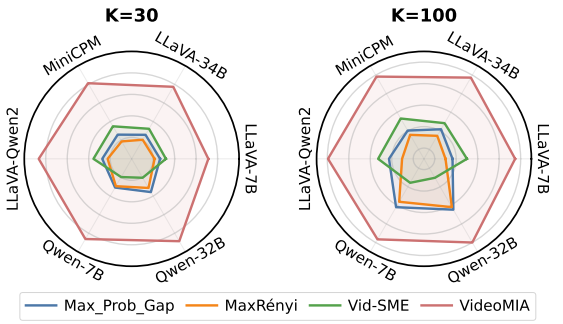

Figure 7: Set-level MIA performance. For Qwen-7B and Qwen-32B, we use the MovieTection Benchmark.

processing strategies used in real-world model training, including resolution degradation and frame downsampling. The robustness study on four open-source models is conducted using the VideoInstruct/Video-MME and LLaVA-178K/Video-MME benchmarks, while experiments on four closed-source models and two Qwen models are performed on Movitection. Details of the original video resolutions and frame sampling rates for these benchmarks are included in the Appendix B. As illustrated in Figure 3, reducing the input resolution (from 100% to 64%) leads to only marginal performance degradation for most models, demonstrating the robustness of VideoMIA against resolution degradation. Meanwhile, under reduced sampling rates (from 100% to 70%), VideoMIA remains largely stable, demonstrating strong robustness to frame downsampling.

4.6 Ablation Study

We conduct ablation experiments on four open-source MLLMs. The results are summarized in Figure 4. First, we observe that both of our temporal probing tasks, FrameIdentify and TempRank, positively contribute to video membership inference. Second, for most models, combining the FrameIdentify and TempRank tasks further improves AUC performance. Furthermore, combining VideoMIA and ImageMIA (MaxRényi) yields the best performance across all evaluated models. These results highlight that VideoMIA is a generalizable framework capable of incorporating multiple temporal reasoning tasks and image-based methods to further enhance membership inference.

4.7 Hyperparameter Analysis

We analyze the impact of the number of test samples N per video and the number of context frames x on our method's performance, as shown in Figures 5 and 6. For the analysis of the number of context frames x, we fix the number of test samples per video N to 15. Conversely, for the analysis of N, we fix x to 3. Across the four MLLMs, we observe that VideoMIA exhibits consistent performance gains with longer context lengths. Meanwhile, as the number of trials increases, performance gradually improves and then saturates, suggesting that an appropriate number of trials is sufficient to achieve stable and reliable performance.

Method	LLaVA-7B		LLaVA-34B		MiniCPM		LLaVA-Qwen2		Avg AUC
	AUC	TPR@5%	AUC	TPR@5%	AUC	TPR@5%	AUC	TPR@5%	
VideoMIA	0.5772	0.2040	0.5875	0.2200	0.5952	0.1674	0.6077	0.2430	0.5918
w/ linear decay	0.5780	0.2120	0.5887	0.2230	0.5964	0.2040	0.6083	0.2477	0.5929
w/ Kendall's τ	0.5677	0.1750	0.5947	0.2277	0.5950	0.1608	0.6108	0.2550	0.5921

Table 3: Performance of VideoMIA with alternative detection score design.

4.8 Membership Inference on Sets

Beyond the proof-of-concept at the sample level, we emphasize that our method can also serve as the backbone of a deployable tool for video training data in LLMs. In real-world scenarios, a data owner typically seeks to determine whether an unauthorized set of videos (rather than a single sample) has been used for training. We additionally provide set-level MIA experimental results in Figure 7. In set-level detection, we aggregate features across multiple samples by computing the average membership score over a set of K videos. This statistical aggregation effectively amplifies the membership signal, strengthening the distinction between member and non-member distributions. Concretely, when aggregating over a set of 100 videos, the separability becomes near-perfect. This demonstrates that set-level auditing, arguably the more practical use case, can achieve highly reliable detection performance.

4.9 Sensitivity Analysis on Weight

For the main results, we use equal weights $(0.5/0.5)$ to combine the detection scores from FrameIdentify and TempRank. We conduct a sensitivity analysis by varying the combination weights between the two detection scores from $0.3/0.7$ to $0.7/0.3$. The additional experimental results are reported in Figure 8. Across the range of weight configurations, the performance remains stable, with only minor fluctuations in AUC. This indicates that our method is not sensitive to the specific choice of combination weights, and that the detected membership signal is robust rather than dependent on a carefully tuned weighting scheme.

4.10 Alternative Detection Scoring

Our primary goal in this work is to introduce a general framework that leverages temporal memorization as a signal for video-level membership inference under a strict black-box setting. To validate this idea in a clean and controlled manner, we intentionally avoid over-optimizing the design of auxiliary detection scoring. In this setting, we consider alternative detection scoring designs for VideoMIA, conducting experiments that replace exponential decay with linear decay, and substitute pairwise AUC with Kendall's τ as the ranking-based detection metric. The corresponding results are reported in Table 3, and they remain consistent with our main findings. Overall, these results demonstrate that our proposed framework is effective across different score designs, confirming the robustness of our approach. Moreover, some of the refinements lead to modest positive improvements, highlighting the potential for further optimization and the flexibility of our method.

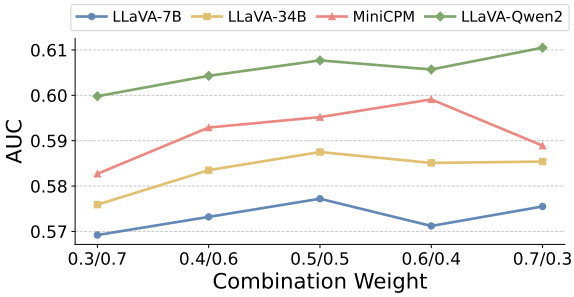

Figure 8: Sensitivity analysis on combination weight.

5 Conclusion

In this work, we introduce VideoMIA, the first black-box membership inference framework designed to identify video-level training data in multimodal LLMs. VideoMIA constructs temporal reasoning tasks to evaluate model memorization on fine-grained sequential dependencies across video frames, which cannot be solved without training exposure, thereby substantially improving the reliability of frame-based and concept-based MIAs. Comprehensive experiments on four benchmarks and ten MLLMs, including fully closed-source systems, show that VideoMIA consistently surpasses strong baseline methods. Moreover, our newly constructed benchmark dataset enhances the robustness and fidelity of evaluation for video-oriented MIA techniques, which will be released publicly to further support research in this field.

Limitations

The rapid proliferation of short-form video content has made it an increasingly important data source for training multimodal language models, while simultaneously raising growing concerns regarding unauthorized data usage and the need for effective auditing mechanisms. Although our proposed method is technically applicable to short-form video data, we do not include empirical evaluations in this setting due to the lack of publicly available benchmarks specifically designed for short-form video membership inference. As future work, we plan to construct a dedicated short-form video benchmark and conduct systematic evaluations to further validate the effectiveness and generalizability of VideoMIA in this emerging domain.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 62425203, 62502044; Beijing Natural Science Foundation under Grant number L253005; CCF-SANGFOR Research Fund under Grant number 20240202; Research Initiation Project for Introduced Talents of BUPT under Grant number 2025KYQD11; and the Beijing Municipal Science & Technology Commission, the Administrative Commission of Zhongguancun Science Park under Grant number Z251100003625014.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *Proceedings of European Conference on Computer Vision*, pages 370–387. Springer.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? In *Proceedings of the First Conference on Language Modeling*.

André V Duarte, Xuandong Zhao, Arlindo L Oliveira, and Lei Li. 2025. Dis-co: Discovering copyrighted content in vlms training data. In *Proceedings of International Conference on Machine Learning*, pages 14807–14832.

Maria Dolores Esteban and Domingo Morales. 1995. A summary on entropy statistics. *Kybernetika*, 31(4):337–346.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in llms: Tracing data contamination in large language models. In *Proceedings of the 12th International Conference on Learning Representations*.

David J Hand and Robert J Till. 2001. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45(2):171–186.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2022. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37.

Yuke Hu, Zheng Li, Zhihao Liu, Yang Zhang, Zhan Qin, Kui Ren, and Chun Chen. 2025. Membership inference attacks against vision-language models. In *Proceedings of the 34th USENIX Security Symposium (USENIX Security 25)*, pages 1589–1608.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Bargav Jayaraman, Chuan Guo, and Kamalika Chaudhuri. 2024. Déjà vu memorization in vision–language models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 50722–50749.

Narine Kokhlikyan, Bargav Jayaraman, Florian Bordes, Chuan Guo, and Kamalika Chaudhuri. 2024. Measuring deja vu memorization efficiently. In *Proceedings of Advances in Neural Information Processing Systems*, pages 30015–30040.

Qi Li, Cheng-Long Wang, Yinzhi Cao, and Di Wang. 2024a. Data lineage inference: Uncovering privacy vulnerabilities of dataset pruning. *arXiv preprint arXiv:2411.15796*.

Qi Li, Runpeng Yu, and Xinchao Wang. 2025. Vid-sme: Membership inference attacks against large video understanding models. In *Proceedings of Annual Conference on Neural Information Processing Systems*.

Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. 2024b. Membership inference attacks against large vision-language models. In *Proceedings of Advances*

in Neural Information Processing Systems, pages 98645–98674.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of Advances in Neural Information Processing Systems*, pages 34892–34916.

Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. 2018. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 12585–12602.

Yonatan Oren, Nicole Meister, Niladri S Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. Proving test set contamination in black-box language models. In *Proceedings of the 12th International Conference on Learning Representations*.

Francesco Pinto, Nathalie Rauschmayr, Florian Tramèr, Philip Torr, and Federico Tombari. 2024. Extracting training data from document-based vqa models. In *Proceedings of International Conference on Machine Learning*, pages 40813–40826.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *Proceedings of the 12th International Conference on Learning Representations*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE.

Liwei Song and Prateek Mittal. 2021. Systematic evaluation of privacy risks of machine learning models. In *Proceedings of the 30th USENIX Security Symposium (USENIX Security 21)*, pages 2615–2632.

Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Membership inference attacks against adversarially robust deep learning models. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Cheng-Long Wang, Qi Li, Zihang Xiang, Yinzhi Cao, and Di Wang. 2025. Towards lifecycle unlearning commitment management: Measuring sample-level unlearning completeness. In *Proceedings of 34th USENIX Security Symposium (USENIX Security 25)*, pages 6481–6500.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, and 1 others. 2024b. Cogvlm: Visual expert for pretrained language models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 121475–121499.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proceedings of the 31st IEEE Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Chunjiang Ge, Yan Yang, Yuhan Dai, Yongdong Luo, Tong Xu, Caifeng Shan, and Enhong Chen. 2026. Sparrow: Data-efficient video-llm with text-to-image augmentation. *IEEE Transactions on Multimedia*.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. 2024. Low-cost high-power membership inference attacks. In *Proceedings of the 41st International Conference on Machine Learning*, pages 58244–58282.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025. Llava-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*.

A Instruction Prompts

We present the instruction prompt X_{ins} used in FrameIdentify Task and TempRank Task.

Prompt for FrameIdentify:
Select the candidate frame that appears earliest in the video after the context frames. Output only the index (e.g. 0, 1, 2, or 3) after '### Answer:'.
Prompt for TempRank:
Sort the candidate frames in chronological order based on their occurrence in the video after the given contexts. Output only the sorted indices of the candidates as a list (e.g., [X X X X]) after '### Answer:'.

B Details for Benchmarks

VideoInstruct/Video-MME and LLaVA-178K/Video-MME. For VideoInstruct/Video-MME and LLaVA-178K/Video-MME, members are sampled from the open training datasets VideoInstruct-100K or LLaVA-Video-178K, while non-members are drawn from the open-source evaluation dataset Video-MME. The details of these benchmarks are summarized in Table 4 and Table 5. For videos in these two benchmarks, the default sampling rate is 10 frames per minute, and the resolution is 1280×720.

YouTubeTection. In YouTubeTection, the default sampling rate is 10 frames per minute, and the default resolution is 1280×720. Each video is categorized by duration into three types: long (> 20 minutes), medium ($5-20$ minutes), and short (< 5 minutes). The selection of videos included in the benchmark is guided by their view counts, based on the assumption that highly popular videos, due to their widespread availability, are more likely to appear in training datasets.

MovieTection. In MovieTection, the default sampling rate for movie videos is one frame per minute, and the default resolution is 1126×512.

C Model Knowledge Cutoff

Notably, different models have distinct knowledge cut-off dates. These cut-off dates are used to separate clean and suspected samples in the benchmarks, particularly in the previously proposed MovieTection, where suspected samples are filtered according to each model's specific knowledge cut-off date (shown in Tabel 6).

Benchmark	Member	Non-Member	Scale
VideoInstruct /Video-MME	VideoInstruct -100K	Video-MME	520/520
LLaVA-178K /Video-MME	LLaVA-Video -178K	Video-MME	520/520

Table 4: Details of VideoInstruct/Video-MME and LLaVA-178K/Video-MME.

Benchmark	Target Model
VideoInstruct /Video-MME	LLaVA-Next-Video-7B LLaVA-Next-Video-34B-hf MiniCPM-V-200K-Video-Finetune (9B)
LLaVA-178K /Video-MME	LLaVA-Video-7B-Qwen2

Table 5: Benchmarks and corresponding target models.

Models	Knowledage Cut-off Date
GPT-4o	2024.06
GPT-4o-Mini	2024.06
Claude-3.7	2024.11
Gemini-2.0-flash	2024.06
Qwen2-VL-7B	2023.06
Qwen2.5-VL-32B	2024.10

Table 6: The knowledge cut-off date for each model. For each model, we use its knowledge cutoff date as a boundary for clean and suspected samples in benchmarks accordingly.

D Details for Experiments

For closed-source models, inference is conducted through API calls, with the generation temperature fixed at 0 to ensure deterministic outputs. For the main results, we use equal weights $(0.5/0.5)$ to combine the detection scores from FrameIdentify and TempRank.

E Unbiased Results

E.1 Benchmark Analysis

To directly test whether member and non-member videos are distinguishable purely due to dataset-level distribution differences, we conduct an experiment using only video content embeddings. Specifically, we extract frame-level embeddings for each

Benchmark	Accuracy
VideoInstruct / Video-MME	0.5290
LLaVA-178K / Video-MME	0.5568

Table 7: Content-based classification accuracy for member and non-member.

Models	VideoMIA	VideoMIA (Unbias)
LLaVA-7B	0.5772	0.5672
LLaVA-34B	0.5878	0.5790
MiniCPM-Finetune	0.5952	0.5902
LLaVA-Qwen2	0.6077	0.6004

Table 8: Performance of unbiased VideoMIA.

video and compute the mean embedding to obtain a single video-level representation. We then train a simple classifier to distinguish members from non-members based solely on these averaged embeddings. If the classifier were able to separate the two groups, this would suggest that dataset-level distribution differences could account for the membership signal. Conversely, failure to separate them would indicate that such distributional bias is not the primary factor. Based on the results reported in Table 7, we observe that the resulting AUC is close to chance level (approximately 0.5x), indicating that the two groups are not easily separable based on raw visual distribution alone. This suggests that large-scale dataset-level bias is unlikely to be the primary driver of the observed MIA performance.

E.2 Unbiased Performance

To further reduce any residual distribution mismatch, we construct a more strictly controlled benchmark by applying distribution matching and filtering procedures to the positive and negative samples. Concretely, we align the member and non-member sets based on embedding similarity, and filter samples to obtain a unbias subset. After this filtering process, the embedding-based classifier yields an AUC extremely close to 0.5000, confirming that the resulting dataset is effectively distribution-neutral. Importantly, when we re-evaluate VideoMIA on this fully distribution-matched subset (in Table 8), the membership inference performance remains largely unchanged. This demonstrates that the temporal reasoning signal exploited by our method is not an artifact of cross-dataset distribution differences, but instead reflects genuine memorization-related effects.

Models	Member	Non-member
LLaVA-7B	0.3261	0.2694
LLaVA-34B	0.3247	0.2608
MiniCPM-Finetune	0.3428	0.2408
LLaVA-Qwen2	0.3564	0.2782

Table 9: Accuracy of four models on member and non-member samples in Frame Identification Task.

Models	VideoMIA	VideoMIA (Filter)
LLaVA-7B	0.5772	0.5791
LLaVA-34B	0.5878	0.5890
MiniCPM-Finetune	0.5952	0.5976
LLaVA-Qwen2	0.6077	0.6080

Table 10: VideoMIA performance after applying the general temporal reasoning filter.

F General-knowledge Filtering

F.1 General Reasoning Performance

We report the results of comparing members and non-members on temporal reasoning tasks, and the results demonstrate that success in these tasks necessarily implies memorization. For example, in our Nearest-Frame Identification Task (n=4), non-member samples achieve an average accuracy of 0.25, which corresponds to the chance level for this 4-option selection task, whereas member samples consistently score higher. Table 9 shows that member and non-member samples are clearly separable. Non-members exhibit random sequence-recovery performance, while members demonstrate consistent, beyond-random behavior. This analysis demonstrates that frame queries can not be solved from only general knowledge.

F.2 Performance with Filtering

We conduct experiments to exclude the potential effects of general temporal reasoning. Specifically, we extract the frames and convert both the context and candidate frames into textual descriptions. We then prompt a strong LLM (GPT-4o) to solve our two temporal tasks purely based on these textual descriptions. Importantly, this reasoning process is performed entirely in the text modality and is independent of any video data. Therefore,

API Model	Image Price ($/image)	Input Text Price ($/1M tokens)	Output Text Price ($/1M tokens)	Number of API Calls	Estimated Cost ($/video)
GPT-4o	0.000638	2.5	10	80	0.0632
Gemini	0.001315	2	12	80	0.1112
Claude	0.00042	3	15	80	0.0480

Table 11: Estimated API costs for membership inference.

it relies solely on the LLM's general reasoning and world knowledge, rather than memorization of video content. To quantify how well a task can be inferred using general knowledge alone, we repeat the prompting process multiple times for each task and compute the average accuracy score. A lower score indicates that the task is difficult to justify based solely on general knowledge. Based on these scores, we retain only the top-N tasks with low general-knowledge recoverability for subsequent MIA evaluation. The resulting MIA performance with filtering, reported in Table 10, shows little difference from our original results. This demonstrates that there was minimal bias in the original experiments, and the membership signal primarily stems from video-specific memorization rather than general temporal reasoning ability.

G Cost Estimation for Attack

Our method involves multiple samplings and repeated queries. However, the computational overhead is controllable. The additional cost arises only at the inference stage and does not require any extra training or model fine-tuning. In practice, the number of sampled permutations per video is limited and fixed, making the overall query budget predictable. Compared to white-box approaches that require access to model gradients, intermediate representations, or retraining shadow models, our method operates purely in a black-box query setting. As a result, it avoids the substantial computational and implementation overhead associated with white-box attacks. Based on current pricing for representative commercial LLM APIs, we summarize the estimated query cost in Table 11.