



USENIX

THE ADVANCED COMPUTING
SYSTEMS ASSOCIATION

MEPS: Privacy-preserving Edge-cloud Video Foundation Model Inference with Privacy Protectability

Siping Shi, *The University of Hong Kong*; Rui Lu, *The Hong Kong Polytechnic University*;
Dan Wang, *Hong Kong University of Science and Technology*;
Bihai Zhang, *The Hong Kong Polytechnic University*

<https://www.usenix.org/conference/usenixsecurity26/presentation/shi-siping>

**This paper is included in the Proceedings of the
35th USENIX Security Symposium.**

August 12–14, 2026 • Baltimore, MD, USA

ISBN 978-1-939133-58-8

Open access to the Proceedings of the
35th USENIX Security Symposium
is sponsored by



جامعة الملك عبد الله
للعلوم والتقنية
King Abdullah University of
Science and Technology

MEPS: Privacy-preserving Edge-cloud Video Foundation Model Inference with Privacy Protectability

Siping Shi^{*}, Rui Lu[†], Dan Wang[‡], Bihai Zhang[†]

^{*}The University of Hong Kong

[†]The Hong Kong Polytechnic University

[‡]Hong Kong University of Science and Technology

Abstract

Edge-cloud video understanding systems, which leverage video foundation models on the cloud for inference to serve downstream tasks, present significant privacy risks due to the exposure of video data containing sensitive information (e.g., human faces) from edge devices to the cloud. To mitigate privacy leakage, existing methods often involve perturbing sensitive information, which can unfortunately degrade the accuracy of video understanding tasks, especially when that sensitive information is relevant to the tasks. Alternatively, encryption-based private inference offers strong privacy guarantees without accuracy loss but imposes a substantial computational burden on resource-constrained edge devices. This paper addresses these privacy challenges by observing that while many video frames contain sensitive information, only a few contain sensitive information that is truly essential for the video understanding tasks. Thus, it is unnecessary and inefficient to apply universal, computation-intensive encryption for all video frames on the edge.

We propose MEPS, a novel privacy-preserving video understanding system designed to intelligently manage the trade-off between privacy, accuracy, and efficiency. MEPS selectively encrypts video frames only when the sensitive information they contain is crucial for the downstream tasks, thus ensuring robust privacy for essential data. For the majority of frames where sensitive information is non-essential to the task, MEPS applies lightweight perturbation schemes. This approach effectively protects privacy by obscuring sensitive details without negatively impacting task accuracy or overburdening edge devices. A core component of MEPS is its ability to discern whether sensitive information within a frame is essential for video understanding. To facilitate this, we propose privacy protectability, a novel metric rooted in information theory that evaluates the degree of overlap between sensitive information and the information required for video understanding tasks. We implement a prototype of MEPS and evaluate its performance with three public real-world video trace datasets. Theoretical analysis proves MEPS has rigorous privacy guarantee. Experimental results demonstrate

that MEPS can maintain the satisfied privacy protection and task accuracy, while significantly improve system efficiency, achieving a $48.8\times$ speedup.

1 Introduction

The increasing adoption of video foundation models (VFMs) stems from their demonstrated success in learning generic, robust representations from video data, applicable to diverse video understanding tasks. For example, driving assistance applications leverage VFMs to generate driving environmental descriptions [9] or perform Visual Question Answering (VQA) on elements like street signs [72]. Similarly, environmental intrusion detection systems can utilize VFMs to ascertain the causes of incidents (e.g., fire, animal or human intruders) by analyzing video content [32]. Given that VFMs are considerably larger and more computationally demanding than conventional video analytics models, their deployment frequently necessitates edge-cloud architectures. In these configurations, video data captured at the edge is transmitted to cloud infrastructure, where it is ingested by the VFM to produce the results for the designated video understanding tasks.

Privacy leakage is a recognized concern in edge-cloud video systems, where sensitive information, such as human faces, commonly present in video streams is susceptible to unauthorized access during cloud transmission and processing [26, 52, 70, 78]. Many privacy-preserving techniques address this by perturbing sensitive data [22, 35, 38, 65]; however, this can lead to significant degradation in the accuracy of video understanding tasks, especially if the sensitive information is pertinent to the video understanding objectives. Encryption-based schemes offer an alternative [37, 49, 69, 76], theoretically providing strong privacy protection while preserving task accuracy. Nevertheless, the high computational overhead of encryption imposes a substantial burden on resource-constrained edge devices. Consequently, to meet operational latency demands when applying encryption at the edge, measures such as reducing frame resolution can be

adopted, which can inadvertently diminish overall accuracy of video understanding tasks.

In this paper, we present quantitative findings from our measurements demonstrating a crucial characteristic of many video understanding applications: while sensitive information is prevalent in a high percentage of video frames (e.g., 70.9%), the subset of this information that is essential for achieving the video understanding objectives is considerably smaller (e.g., 7.3%). To illustrate, in a driving assistant application, details such as individual pedestrians or specific license plates, while sensitive, may not be requisite for tasks like environmental description or Visual Question Answering. Based on this observation, we argue that ubiquitous encryption is unnecessary for all frames in contemporary video systems. Instead, we propose a selective approach: encrypting only those frames where sensitive information is indispensable for the video understanding task. For the remaining frames, lightweight perturbation schemes [45] can be employed to obscure sensitive data (e.g., through noise addition) without negatively impacting task accuracy. It is important to recognize that perturbation scheme alone is insufficient to support the video understanding task when sensitive information is task-essential. For example, if a frame depicts a human signaling to cross the street, this sensitive information (the human and their gesture) is vital for accurate environmental summarization; perturbing it would substantially decrease task accuracy.

A key challenge to achieving our approach is effectively determining whether sensitive information within a video frame is essential for the given video understanding task. To address this, we introduce *privacy protectability*, a novel metric designed to characterize the degree to which a video frame can be privacy-protected without compromising the given video understanding task performance. Intrinsically, privacy protectability quantifies the “overlap” between sensitive information and the task-essential information that contributes to video understanding accuracy. Accordingly, computationally intensive encryption methods (e.g., homomorphic encryption [1]) are necessitated only when a frame exhibits low privacy protectability, indicating a significant overlap where sensitive information is essential.

We establish the theoretical foundation of privacy protectability with principles from information theory. Specifically, the accuracy of both video understanding tasks and the privacy inference attacks are determined by the information embedded in features of each frame [8] [5]. To operationalize this, we define a feature contribution score using Shapley Values. This score measures the marginal contribution of each feature to both video task accuracy (accuracy contribution) and potential privacy leakage (privacy contribution). A feature is considered “protectable” if the ratio of its privacy contribution to its accuracy contribution, representing the degree of undesirable overlap, falls below a predefined threshold. The overall privacy protectability of a video frame is then defined as the aggregated contributions of its protectable features.

We propose a Mixed Encryption and Perturbation System (MEPS) for privacy-preserving video understanding in edge-cloud environments. MEPS leverages privacy protectability metric to strategically combine the advantages of both encryption and perturbation schemes for optimal video understanding task performance, guided by two core design principles:

(1) **Efficient Privacy Protectability Estimation:** Exact calculation of the privacy protectability for each frame is computationally prohibitive, due to the exponential number of evaluations required to determine feature contribution scores (which involve assessing accuracy differences with and without each feature across all possible feature subsets). To overcome this, MEPS employs a privacy protectability estimation approach based on contrastive learning. This allows for the development of an accurate estimation model using a limited set of ground truth data.

(2) **Adaptive Runtime Privacy Protection:** To guarantee fast runtime execution, MEPS employs a hybrid strategy involving offline profiling and online adaptation. In the offline phase, various encryption and perturbation schemes are profiled to ascertain their accuracy levels under diverse configurations. At runtime, an optimization algorithm is deployed to dynamically adapt to the heterogeneous privacy protectability scores of incoming video frames, selecting optimal protection schemes and their configurations to minimize the latency associated with video understanding tasks while preserving privacy protection and task accuracy.

We developed a prototype of MEPS, packaged as a readily deployable Docker image suitable for various testbeds, including the Nvidia Jetson Nano. MEPS is configured via a simple file, allowing applications to adapt their settings prior to onboard deployment easily. We evaluate the performance of MEPS on an Nvidia Jetson Nano using three typical video understanding tasks: Traffic Report Captioning (TRC), Driving Assistance (DA), and Intrusion Detection (ID). Its performance was benchmarked against two encryption schemes (ConcreteML [75] and PECAM [69]), one perturbation scheme (DP-Forward [13]), and a plain (unprotected) baseline. MEPS consistently outperformed these baselines. For instance, in the DA task, while ConcreteML incurred a latency 34.67 times higher than the plain baseline due to its substantial encryption workload, MEPS achieved a latency only 2.26 times higher, which speeds up the system performance by 15.3 times. This significant improvement is attributable to the capability of MEPS to accurately quantify the privacy protection needs of each frame, and to selectively apply computationally intensive encryption only to frames that have a lower privacy protectability. Furthermore, MEPS demonstrated superior accuracy compared to perturbation-only approaches; specifically, the accuracy of DP-Forward was approximately 17.8% lower than MEPS. This discrepancy arises because DP-Forward indiscriminately injects noise, potentially corrupting sensitive information that is crucial for the video understanding task. The primary contributions of



Figure 1: Overview of edge-cloud VFM inference workflow.

this paper are summarized as follows:

- We observe that while most video frames in common video understanding tasks contain sensitive data, only a few contain sensitive information essential to the task, thereby exposing the inefficiency of current protection methods on resource-constrained edge devices and motivating this study (§2).
- We propose *privacy protectability*, a new metric to quantify the degree to which a frame can be protected for a given video understanding task. This metric uniquely delineates the appropriate conditions for applying encryption versus perturbation schemes (§3).
- We develop MEPS, a novel edge-cloud video understanding system that dynamically adapts its privacy protection strategy based on frame-level privacy protectability, selecting appropriate configurations to minimize video task latency (§4).
- We provide a theoretical privacy analysis and prove MEPS satisfy a rigorous privacy guarantee (§5). We also implement a prototype of MEPS and conduct comprehensive evaluations against established benchmarks, with a case study to validate its effectiveness (§6).

2 Background and Motivation

2.1 Background

Video foundation models. Video foundation models (VFMs) represent a class of large-scale models engineered to learn generic and robust representations from video data, applicable to diverse video understanding tasks such as video description and captioning, visual question answering (VQA), and video grounding and retrieval [41]. VFMs exhibit superior generalization capabilities in practical deployments compared to traditional deep video models, which are often constrained by task-specific designs and limited input-output spaces. Numerous VFMs have been proposed for various applications.

For instance, VisionGPT [29] targets open-world object detection and aims to provide safe visual navigation assistance for visually impaired individuals in complex environments. Lingo-1 [68], developed by Wayve, offers open-loop driving commentary, enhancing the interpretability of autonomous driving behavior. The process of *video foundation model inference* typically entails feeding the VFM with input videos and textual prompts (i.e., instructions) to generate the desired outputs for the specific video understanding task.

Edge-cloud video understanding systems. Given that VFMs are substantially larger and more computationally demanding than traditional models, edge-cloud architectures have emerged as a common deployment strategy for video understanding systems (see Fig. 1). In this paradigm, a user at the edge initiates a video understanding task (e.g., captioning the driving environment). Video frames and textual prompts (e.g., “Please describe the driving environment”) captured by the edge device are transmitted to the cloud. There, VFMs, typically comprising vision and text encoders to extract multimodal features, process this input to generate the understanding result, which is then relayed back to the user. While this edge-cloud paradigm leverages powerful cloud resources to meet computational demands, it concurrently introduces significant privacy concerns due to the transmission of potentially sensitive video data and its subsequent processing in the cloud.

Privacy-preserving techniques for video. A primary challenge for edge-cloud video systems is safeguarding the privacy of video data generated at the edge. Consequently, research into privacy-preserving techniques for edge-cloud video analytics, which encompasses video understanding systems, has garnered considerable attention. Existing privacy-preserving approaches can be broadly categorized based on the underlying technology employed:

(1) **Encryption-based schemes:** Video frames are encrypted using cryptographic techniques such as homomorphic encryption (HE) [1, 49] or secure multiparty computation (SMC) [62, 73] prior to transmission. These schemes are well-suited for applications where sensitive information within each frame is indispensable for the video understanding task (e.g., face recognition). However, they are characterized by high computational complexity.

(2) **Perturbation-based schemes:** This approach involves perturbing sensitive information within video frames at the edge using techniques such as differential privacy [23, 65], object removal [35], or data transformation [38]. Perturbation-based schemes are suitable for scenarios where the sensitive information is non-essential to the video understanding task. Conceptually, this applies when sensitive details (e.g., pedestrians at the roadside) do not significantly “overlap” with the information critical for the task (e.g., identifying an upcoming crossroads). Though these schemes generally exhibit lower computational complexity, they introduce inevitable degraded task accuracy.

2.2 Threat model

We adopt a threat model commonly used in edge-cloud video analytics systems [21, 38, 69]. We consider an architecture in which an edge device captures video frames and applies privacy protection mechanisms before transmitting data to a cloud server for inference. The edge device is trusted by the user or system operator and executes the protection mechanisms correctly. The cloud server provides computational resources but is not trusted with sensitive visual content. Our threat model assumes a curious or semi-trusted cloud provider, including possible insider access. In this setting, MEPS enables users to use powerful reasoning capability of cloud-based VFMs while reducing privacy exposure without significantly sacrificing utility or latency.

Attacker’s goal. The attacker’s primary goal is to extract sensitive visual information from video data processed by the edge-cloud system. Such information includes personal attributes such as facial identity or characteristics, license plate numbers, and other visually identifiable cues that may reveal an individual’s identity, behavior, or location. Successful extraction may enable privacy-invasive activities, such as tracking individuals across locations, linking video observations with external data sources, or inferring private routines in surveillance or driving-assistance scenarios.

Attacker’s capabilities. We assume that the attacker has access to all data transmitted from the edge device to the cloud, either by acting as the cloud service provider or by passively observing network traffic, and it has full knowledge of the system design (Kerckhoffs’s principle). The attacker may leverage public datasets and state-of-the-art models (e.g., face recognition or object detection models) to perform attribute inference [43, 55], and is assumed to have sufficient computational resources to conduct such analysis. We assume that the attacker does *not* have access to the secret keys used for encryption on the edge device and cannot compromise with the edge device’s execution or modify the inference pipeline.

2.3 Motivation

In many video understanding applications, sensitive information can overlap with task-essential features. Although such overlap may occur infrequently, its impact can be significant. For example, in driving assistance systems, standard road signs typically contain no sensitive information. However, when critical cues originate from humans, such as a pedestrian’s hand gesture indicating an intention to cross the road, privacy-preserving perturbation can substantially degrade task accuracy. Similar challenges arise in applications like environmental event detection, where identifying the cause of damage (e.g., fire, animals, or human intruders) may require analyzing frames that contain both sensitive and task-essential information. As a result, conventional approaches often resort to encrypting all frames in which such overlap may occur.

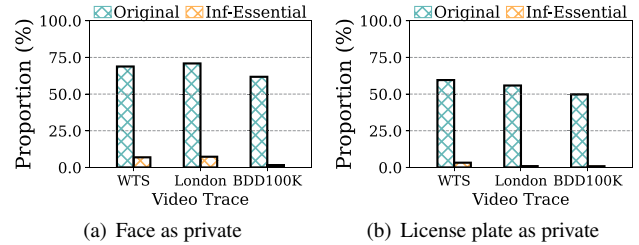


Figure 2: Proportion of sensitive frames under different attribute inference attacks.

Our research is motivated by the observation that this overlap between sensitive and task-essential information is typically occasional rather than constant. Consequently, encrypting all video frames in existing systems is often an overly conservative and inefficient design choice.

Measurement setup. We implement an edge-cloud video understanding system to conduct a measurement study on the driving commentary application. The edge device is Nvidia Jetson Nano [47], and the cloud server runs the pre-trained vision foundation model VisionGPT [67] on an Nvidia RTX 4090 GPU. We also collect three video traces that are captured by the cameras equipped on vehicles when driving in the streets of different cities (i.e., WTS [34], London [63], BDD100K [74]) as the input of the application. We consider human faces and license plate numbers as the sensitive information, and implement two attribute inference attacks to identify them in each video frame. One attack employs a SOTA model ArcFace [10] to identify faces, and the other attack adopts a pre-trained YOLOv8 model to recognize the license plate numbers [50].

Sensitive frames and Inf-Essential sensitive frames. We define a video frame as a sensitive frame if the attribute inference attack can extract sensitive information from it. To measure whether the sensitive information in a video frame overlaps with the driving commentary task-related information, we first perturb each video frame by removing the sensitive human faces or license plates and then perform the driving commentary task on the perturbed frames. We call a video frame an Inf-Essential sensitive frame if the task accuracy of the perturbed frame is decreased by more than 80% compared to the original one. The measurement results are shown in Fig. 2, and we have the key observation.

Key observation: Most video frames contain sensitive information while few frames contain sensitive information that is essential to the video understanding task.

Fig. 2(a) depicts that over half of frames contain sensitive human faces in most video traces, especially in the London trace, where more than 81.54% contain sensitive information. When considering Inf-essential sensitive frames for three video traces, the corresponding proportion of Inf-essential sen-

sitive frames is at most 3.25%, which is 56.29% less than the sensitive frames. Similar results can be found in Fig. 2(b) where license plates are considered as sensitive. The underlying reason is that the Inf-essential sensitive frames which require privacy protection with encryption-based schemes occasionally occur during the process of driving assistant.

The above observations show that encryption is not necessary applied on all video frames, and effectively identifying whether a frame is (non-)essential to the video understanding task can help to largely reduce the computation cost, as most frames in which the sensitive information is non-essential to the video understanding task can be protected with the lightweight perturbation-based scheme without sacrificing the task utility. To achieve this, we design a privacy protectability metric to effectively quantify the overlap degree between the sensitive information and the task-essential information for each video frame in the §3.

3 Privacy Protectability

We introduce a privacy protectability metric to quantify the degree of overlap between sensitive information and video understanding task-essential information within a given video frame, assessed through feature contribution scores. A higher privacy protectability value signifies a lower degree of overlap, indicating that the frame can be effectively protected using perturbation-based schemes with negligible impact on the utility of the video understanding task. Furthermore, we demonstrate how this privacy protectability metric can be utilized to judiciously determine the appropriate application of either perturbation-based or encryption-based privacy schemes.

3.1 Overview

We define a metric, privacy protectability, to measure to what degree a video frame can be protected for a given video understanding task. Intrinsically, a video frame can be represented by a set of features (i.e., a block of pixels can be seen as a feature). The information contained in a feature may not only contribute to predicting the video understanding task attribute but also contribute to predicting the sensitive attribute. For example, the information contained in the features related to a human is important to analyze both the video understanding task attribute of human behavior and the sensitive attributes of human identity. This indicates there exist features *overlap* between the sensitive information and the information contributing to the video understanding task. Intuitively, it is easier to protect the privacy of the frame if the contained features have fewer or no overlap between the sensitive information and the information contributing to the video understanding task, and we call these features *protectable features*. Therefore, we define privacy protectability as the proportion of accuracy contributions made by protectable features, as shown in Fig. 3. To achieve this, we first need to define the contribution

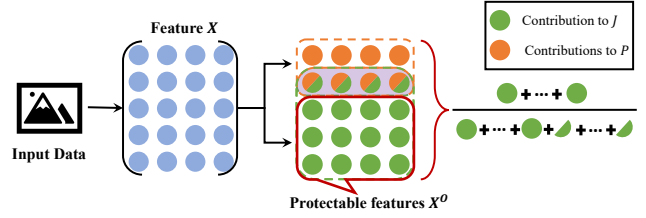


Figure 3: Privacy protectability modeling overview.

of each input feature for a given task with a deep neural network model. Then, we define the privacy protectability metric based on the proportion of accuracy contributions made by the protectable features.

Feature Contribution Score. For a given task model and a video frame consisting of a set of features, our purpose is to obtain the contributions of each feature made to the prediction of the task model. Intrinsically, features collaboratively make contributions to the task, and to obtain the contribution score of a feature, we need to measure the accuracy difference with and without this feature in all possible feature sets. Many feature contribution allocation methods are designed to achieve this in the model explanation domain. According to the model architecture that the methods can explain, existing works can be divided into two types: 1) model-specific methods, such as the coefficient magnitude [15] for linear models and Gini importance [39] for decision trees, 2) model-agnostic methods, such as local interpretable model-agnostic explanations (LIME) [79], permutation importance [4], and Shapley Values [24]. We adopt the popular model-agnostic method Shapley Value to measure the feature contribution as it provides a fair allocation of contribution among features and captures feature interactions. Specifically, the Shapley Values method assigns a value to each feature based on its marginal contribution when combined with other features. Suppose $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ to be a video frame consisting of n features, and $\mathbf{y}$ is the target attribute of the video frame $\mathbf{x}$. According to the Shapley Values, the score that the feature $\mathbf{x}_i$ contributes to predicting $\mathbf{y}$ can be represented as:

$$C_{\mathbf{x}_i}^{\mathbf{y}} = \sum_{\mathbf{x}_s \subseteq \mathbf{x} \setminus \{\mathbf{x}_i\}} w_s \cdot (v(\mathbf{x}_s \cup \mathbf{x}_i) - v(\mathbf{x}_s)), \quad (1)$$

where $w_s = \frac{|\mathbf{x}_s|!(|\mathbf{x}| - |\mathbf{x}_s| - 1)!}{|\mathbf{x}|!}$, and $v(\cdot)$ is the value function for a feature set.

3.2 Definition of Privacy Protectability

Privacy Protectability. For a given video frame, we attempt to measure to what degree it can be protected for a given video understanding task. Since a video frame is composed of features, the protectable degree of a frame can be represented by the accuracy contributions made by the protectable features. Intrinsically, each feature contained in the video frame can

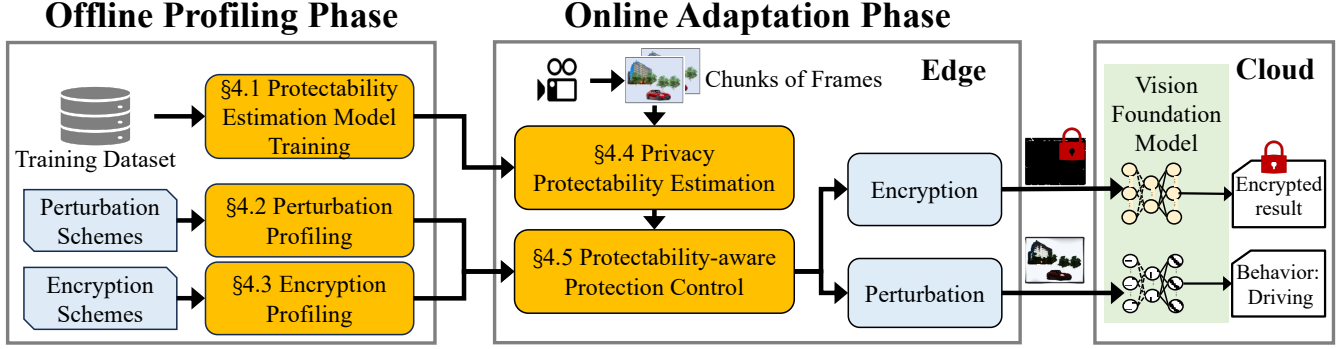


Figure 4: Overview of the MEPS system.

make either accuracy contributions or privacy contributions, or both. A feature can be easily protected when it makes no or few privacy contributions. While it is much more difficult to be protected if the feature makes both many accuracy contributions and privacy contributions. To capture this characteristic, we can measure the protectable degree of a feature with the *privacy-accuracy contribution ratio*. Formally, let C_i^J and C_i^P be the contribution score of feature $\mathbf{x}_i$ made to the video understanding task J and privacy attack P , respectively, then the privacy-accuracy contribution ratio of $\mathbf{x}_i$ is $\mathcal{R}_i = C_i^P / C_i^J$. In the extreme case where feature $\mathbf{x}_i$ contributes to the accuracy of privacy attack and does not contribute to video understanding task accuracy, we have $\mathcal{R}_i = \infty$, and in the other extreme case, $\mathcal{R}_i = 0$. A feature is protectable if $\mathcal{R}_i$ is no more than a threshold δ . Intuitively, the more accuracy contributions made by the protectable features, the higher the privacy protectability of the frame. Thus, we can define the privacy protectability of a frame as the proportion of accuracy contributions made by the protectable features as compared to the whole feature set in the following.

Definition 1. (*Privacy protectability*) Let $\mathbf{x}$ be the video frame with n features, and C_i^P and C_i^J be the contribution score of feature $\mathbf{x}_i$ to video understanding task J and privacy attack P . The privacy protectability of $\mathbf{x}$ denoted by $Pr(\mathbf{x}, J, P)$, is defined as:

$$Pr(\mathbf{x}, J, P) = \frac{\sum_{\mathbf{x}_i \in \mathbf{x}^O} C_i^J}{\sum_{\mathbf{x}_j \in \mathbf{x}} C_j^J}, \quad (2)$$

where $\mathbf{x}^O = \{\mathbf{x}_i | \frac{C_i^P}{C_i^J} \leq \delta, \mathbf{x}_i \in \mathbf{x}\}$.

Interpretation of privacy protectability. The privacy protectability metric is built on Shapley-based feature contribution scores. For each feature, it uses a privacy-side score to measure its contribution to the attacker and a task-side score to measure its contribution to task utility. Their ratio therefore quantifies the feature’s privacy sensitivity relative to task utility. A high ratio means the feature is less suitable for the lightweight perturbation mechanism and may require stronger protection, such as encryption. In this sense,

privacy protectability characterizes the suitability of a frame for lightweight perturbation protection within MEPS.

3.3 Application of Privacy Protectability

Privacy protectability is a new metric of a frame $\mathbf{x}$. We now apply this metric to quantify whether a perturbation scheme can or cannot be applied for this frame.

Assume a video understanding task J has an accuracy requirement Δ and a privacy leakage requirement σ for a certain privacy attack P . Assume that there is a perturbation scheme $\mathcal{P}$ can achieve a certain video understanding task accuracy level A while satisfying the given privacy leakage level ϵ , we can illustrate this as $A = \mathcal{F}_{\mathcal{P}}(\epsilon)$. Note that $\mathcal{F}_{\mathcal{P}}$ is determined by the internal perturbation design of $\mathcal{P}$ and we do not make assumptions on any specific $\mathcal{F}_{\mathcal{P}}$.

Theorem 1. If $Pr(\mathbf{x}, J, P) < \mathcal{F}_{\mathcal{P}}(\sigma)$, $\mathcal{P}$ cannot achieve the accuracy requirement Δ and the privacy leakage requirement σ simultaneously.

Remarks: We omit the proof of this theorem which can be derived from the definitions with simple twists. Its key implication is that for a perturbation scheme, there are frames it cannot protect while satisfying both accuracy and privacy requirements. Our privacy protectability metric quantifies this limitation which is independent of resources. While encryption schemes can achieve both accuracy and privacy, they may be constrained by resource limitations, such as needing to reduce frame resolution, which can lower video understanding accuracy. However, the key difference is that certain accuracy levels can be achieved with sufficient resources (encryption schemes), whereas they are unattainable with perturbation schemes due to the "overlap" of sensitive information.

4 MEPS Design

We design MEPS, an edge–cloud privacy-preserving video understanding system that protects sensitive video content

captured at the edge while minimizing task latency under computing and communication constraints. MEPS leverages the privacy protectability metric to adaptively apply encryption-based or perturbation-based protection on a per-frame basis.

Fig. 4 depicts an overview of MEPS which consists of an **offline profiling phase** and an **online adaptation phase**. In the offline phase, MEPS constructs three profiles: (1) a **protectability estimation model training module** (§4.1) takes the sampled video data as input to obtain a model to predict frame-level privacy protectability; (2) a **perturbation profiling module** (§4.2) that maps task accuracy to different perturbation configurations (e.g., resolution and privacy levels); and (3) a **encryption profiling module** (§4.3) that maps task accuracy to encryption configurations (e.g., resolution). The offline module can be computed in the cloud periodically to adapt to the video content shifts that occur during online deployment.

In the online adaptation phase, incoming video frame chunks are first fed into the **privacy protectability estimation module** (§4.4), to obtain frame-level protectability scores, and then the **protectability-aware control module** (§4.5) selects the appropriate protection scheme and configuration for each frame based on these scores, the offline profiles, privacy requirements, resource constraints, and latency targets, using the protection scheme scheduling and configuration (PSSC) algorithm. Finally, the video frames are subsequently protected via either encryption or perturbation and transmitted to the cloud for inference with a VFM. Particularly, the cloud maintains two specialized VFM backends: a standard inference pipeline for perturbed frames and a private inference pipeline for encrypted frames using the corresponding HE libraries. All the inference results are returned to the edge.

4.1 Privacy Protectability Estimation Model Training with Contrastive Learning

We adopt contrastive learning to train the privacy protectability estimation model. We first introduce the model structure design and then demonstrate the model training approach.

Privacy protectability estimator structure. The privacy protectability is computed based on the contribution score of each feature in the frame. When an input frame composed of a set of features is fed into the estimator model, the output of the model should also be a set of contribution scores for the corresponding feature set. Therefore, the designed privacy protectability model should make the feature-to-feature prediction available. Many feature-to-feature model structures are proposed in the computer vision domain for image segmentation, such as GAN [3], U-Net [54], etc. We adopt U-Net as it is much more efficient compared to other methods, which is suitable for the resource-constrained edge device. Specifically, the privacy protectability estimation model is composed of two parts: an encoder for extracting the latent representation of the input frame and a head to predict the contribution

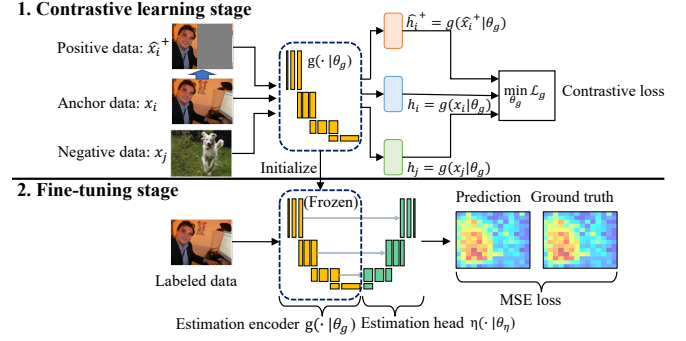


Figure 5: Privacy protectability estimation model training.

of each feature in the frame.

Privacy protectability estimator training. The key challenge of training the privacy protectability estimator is to obtain the ground truth of the feature contribution score, as it is computationally intensive to compute the accuracy difference with and without each feature. We adopt a contrastive learning approach to overcome this as it enables us to train a model with a few ground truths. Contrastive learning aims to learn representations by contrasting similar and dissimilar pairs of samples in an unsupervised manner [6] and has been adopted in many areas, such as Natural Language Processing (NLP) [18] and recommendation systems [80]. Fig. 5 shows the protectability estimation model training process which includes two stages: 1) In **contrastive learning stage**, we aim to learn an estimation encoder to extract the latent representation of the input frame. Specifically, we construct a positive pair and several negative pairs for each anchor data x_i in the training data set. A positive pair includes the anchor data x_i and the synthetic data $\hat{x}_i^+$ which has the most similar feature contributions as x_i . To obtain $\hat{x}_i^+$, we need to generate a synthetic positive sample set $\mathcal{X}_i^+$ via m times of independent perturbations on the data instance x_i , and select the synthetic sample which has the minimum prediction difference with x_i as in [7]. A negative pair includes the anchor data x_i and randomly selected data x_j where the feature contributions mostly differ from x_i . With these positive and negative pairs, we can train the estimation encoder $g(\cdot|\theta_g)$ by minimizing the contrastive loss $\mathcal{L}_g$ as follows:

$$\mathcal{L}_g = -\log \frac{\exp(g(x_i|\theta_g) \cdot g(\hat{x}_i^+|\theta_g)/\tau)}{\sum_{j=1}^N \exp(g(x_i|\theta_g) \cdot g(x_j|\theta_g)/\tau)}, \quad (3)$$

where τ is a temperature hyper-parameter and N denotes the batch size for training; 2) In **fine-tuning stage**, we add an estimation head $\eta(\cdot|\theta_\eta)$ after the trained estimation encoder $g(\cdot|\theta_g)$ to predict privacy protectability. Specifically, we freeze $g(\cdot|\theta_g)$ and tune the head $\eta(\cdot|\theta_\eta)$ with a small amount of ground truth by minimizing the MSE loss between the prediction and the ground truth until $\eta(\cdot|\theta_\eta)$ is converged.

4.2 Perturbation Profiling

A perturbation scheme attempts to add appropriate noise to the input video frame to achieve a privacy protection level and maximize the video understanding task accuracy¹. Since the relationship between the task accuracy and a perturbation scheme cannot be explicitly expressed and may vary under different input video frame configurations, we need to establish a mapping profile between input configurations and the task accuracy for the given perturbation scheme. With this profile, we can quickly find the maximum task accuracy a perturbation scheme can achieve for a given video frame with the corresponding configurations in runtime.

Our profiling mainly focuses on configurations of the perturbation scheme $\mathcal{P}$, which may affect the task accuracy A when achieving a privacy protection level ϵ . We do not pay attention to the specific noise injection designs of the perturbation scheme. For frames with different privacy protectability Pr , the proportion of the protectable features may vary, thus leading to different task accuracy with the same perturbation scheme. In addition, for a given video frame, different values of resolution R greatly impact the amount of injected noise, which also leads to a different task accuracy. For example, a smaller amount of noise can be injected to achieve the same privacy leakage level for a frame with a resolution of 540p as compared to a greater amount of noise to be injected into a frame with a resolution of 1080p. In summary, our profiling problem is to construct the mapping $\mathcal{P} \rightarrow (\epsilon, Pr, R, A)$, or in other words, $A = \mathcal{M}(\mathcal{P}, \epsilon, Pr, R)$.

We adopt a data-driven approach to construct the perturbation profiling map. Specifically, video frames captured at the edge are processed using the perturbation scheme under different resolutions and privacy budgets, producing perturbed frames that are then evaluated by the VFM to measure task accuracy. Each configuration tuple $\langle Pr, R, \epsilon \rangle$, together with its averaged accuracy A , forms an entry in the profiling map. While finer-grained configurations improve profiling accuracy, we balance profiling cost-effectiveness by controlling the granularity of resolution and privacy levels based on available storage and search resources.

4.3 Encryption Profiling

In an encryption scheme, the input data is first encrypted with a private key to protect the sensitive information, and then the encrypted data is fed to the VFM to output the inference result. When the computation resources are sufficient, the encryption scheme can preserve both task accuracy and video privacy. Otherwise, the resolution of the input frame is downgraded to satisfy the latency requirement, resulting in decreased video understanding task accuracy. The relationship between the configuration (i.e., resolution) of the encryption scheme and

¹Implicitly, it is understood that the accuracy of a perturbation scheme is measured on a set of frames and we do not explicitly present this.

the task accuracy can not be obtained analytically, and we need to profile it for a given encryption scheme.

To generate the configuration-accuracy mapping profile of an encryption scheme, we adopt a similar approach as in §4.2. The collected frames are first encrypted with different resolutions and then fed into the VFM to output the video understanding result for measuring the task accuracy. Finally, each configuration with the corresponding average accuracy is recorded in the profile.

4.4 Privacy Protectability Estimation

We calculate the privacy protectability of each frame in the frame chunks with the pre-trained privacy protectability estimation model. Specifically, each frame is fed into the privacy protectability estimation model and performs inference to output the estimated privacy protectability score. The computation overhead introduced by the protectability estimation model inference is negligible. The latency of running the estimation model with U-Net is only about 40ms on the Jetson Nano device, while the computation cost of encryption and neural network-based perturbation method is far more than the protectability estimation model inference. We can optimize the privacy protectability estimation model to adapt to the dynamic video content shifts that occurred during deployment through an online learning approach. Specifically, we can re-train and update the deployed model with the online sampled video frames when the video content shifts occur. Online learning is not the research focus of our work and has been extensively explored in literature [33] [46] [31], and we can implement the online privacy protectability model update by leveraging existing methods.

4.5 Protectability-aware Protection Control

4.5.1 Problem Formulation

For an edge-cloud video understanding system, a camera is equipped on an edge device which has constrained computation and communication resources. A chunk of m video frames $\mathcal{V}$ is captured by the camera and fed into the video understanding system to obtain the result. Each frame has a minimum accuracy requirement α . To protect the privacy of the video frames, a perturbation scheme $\mathcal{P}$ and an encryption scheme $\mathcal{E}$ are installed in the video understanding system. For frames with different protectability Pr , different configurations $\mathbf{G}$ (i.e., resolution) of the protection scheme can lead to a different video understanding task accuracy. Therefore, we have **The Protectability-aware Video Understanding (PVU) Problem**. Given video frames $\mathcal{V}$, the video understanding task J , the privacy protectability scores $\mathbf{Pr}$ of $\mathcal{V}$, the computation and communication resources R^C, R^B of the edge device, the profiles $\mathcal{M}^{\mathcal{P}}, \mathcal{M}^{\mathcal{E}}$ of the scheme $\mathcal{P}$ and $\mathcal{E}$, respectively, the privacy protection level L^{max} , and the minimum accuracy requirement for each frame α , we need to determine

Algorithm 1: Protection Scheme Scheduling and Configuration (PSSC)

Input: $\mathcal{V}, \mathbf{Pr}, R^C, R^B, \mathcal{M}^P, \mathcal{M}^E, L^{max}, \alpha$.

Output: Protection scheme configurations $\mathbf{G}$

```

1 Sort  $\mathcal{V}$  according to decreasing protectability score in  $\mathbf{Pr}$ ;
2 Construct configuration set  $\mathcal{G}_i$  for frame  $V_i$  with  $Pr_i, L^{max}$ , and  $\alpha$ ;
3 Sort  $\mathcal{G}_i$  according to increasing computation cost for all  $V_i \in \mathcal{V}$ ;
4 for  $V_i \in \mathcal{V}$  do
5   for  $j = 1, \dots, |\mathcal{G}_i|$  do
6      $\tilde{T}_{i,j} = T_{i,j} - T_{i,j-1}, \tilde{r}_{i,j}^C = r_{i,j}^C - r_{i,j-1}^C$ ;
7    $G_i = g_1, \bar{R}^C = \bar{R}^C - r_{i,1}^C$ ;
8 Sort  $\mathcal{G}_i$  according to decreasing improvement  $\tilde{T}_{i,j}/\tilde{r}_{i,j}^C$ ;
9  $N = \max_{i=1, \dots, m} |\mathcal{G}_i|$ ;
10 for  $V_i \in \mathcal{V}$  do
11   for  $k = 2, \dots, N$  do
12     if  $r_{i,k}^C \leq \bar{R}^C$  then
13        $G_i = g_k, \bar{R}^C = \bar{R}^C - r_{i,k}^C$ ;
14 Output  $\mathbf{G} = \{G_i | i = 1, \dots, m\}$ ;
```

the configurations $\mathbf{G}$ of the protection scheme, to minimize the video understanding task latency $\sum_{i=1}^m T_i$ on $\mathcal{V}$.

4.5.2 The Protection Scheme Scheduling and Configuration (PSSC) Algorithm

We design a protection scheme scheduling and configuration algorithm to solve the PVU problem. The key idea is to select the configuration of the protection scheme with the minimum latency of unit resource consumption for each frame while satisfying the accuracy requirement. Details of our designs are demonstrated in Algorithm 1. Specifically, the PSSC algorithm first constructs the available configuration candidates set for each frame, in which each satisfies the privacy protection level and minimum frame accuracy requirement (Line 2). Then, we sort the configuration candidates set of each frame according to the latency of unit resource consumption with the decreasing order (Line 4-8). Finally, we iteratively assign the configuration that has the minimum latency to each frame until all the resource is consumed to satisfy the latency requirement (Line 10-13). Our PSSC algorithm has a sublinear search complexity as the configuration profile is prepared offline, which will not incur a burden in runtime execution.

5 Privacy Analysis

MEPS provides a hybrid, per-frame privacy guarantee that adapts to the content of each frame. The formal guarantee depends on the protection mechanism applied. For frames protected via encryption, the guarantee is *semantic security* derived from the standard cryptographic assumptions of the underlying HE scheme, which provides the strongest level of data confidentiality. For frames protected via perturbation scheme, particularly the Differentially Private Pixelization

(DP-Pix) [17] algorithm in our implementation, the privacy guarantee can be formally proved to satisfy ϵ -DP.

Firstly, we need to define neighboring images as following.

Definition 2. (*m-neighbouring images*) Two images I_1 and I_2 are neighboring images if they have the same dimension and they differ by at most m pixels.

Then, we can calculate the global sensitivity of the DP-Pix algorithm. In this algorithm, an input image first performs pixelization and applies Laplace perturbation to the pixelized image. Specifically, an $M \times N$ image I is partitioned into $K = \lceil \frac{M}{b} \rceil \lceil \frac{N}{b} \rceil$ grid cells, where each grid cell contains $b \times b$ pixels. Let c_k denote the k -th grid cell. The pixelization of an image I can be denoted as a vector of length K , i.e.,

$$P_b(I) = \left\{ \frac{1}{b^2} \sum_{(x,y) \in c_1} I(x,y), \dots, \frac{1}{b^2} \sum_{(x,y) \in c_K} I(x,y) \right\}. \quad (4)$$

The global sensitivity of P_b is thus $\Delta P_b = \max_{I_1, I_2} |P_b(I_1) - P_b(I_2)| = \frac{255m}{b^2}$, as the difference between any two pixels is at most 255, and up to m pixels can differ between any neighboring images I_1 and I_2 . Let $\tilde{N} = \{\tilde{N}_1, \tilde{N}_2, \dots, \tilde{N}_K\}$ and each $\tilde{N}_k (k \in \{1, \dots, K\})$ is randomly drawn from a Laplace distribution with mean of 0 and a scale of $\frac{255m}{b^2\epsilon}$. The differentially private pixelization algorithm $\tilde{P}_b$ can be represented as $\tilde{P}_b(I) = P_b(I) + \tilde{N}, \forall I$, which achieves ϵ -differential privacy according to the Laplace mechanism definition [16].

Finally, since each frame is perturbed independently with a privacy budget ϵ , we can have the following privacy guarantee of MEPS based on *parallel composition* [42].

Theorem 2. MEPS satisfies ϵ -differential privacy.

Proof. Assume MEPS perturbs C frames by individually applying the differentially private pixelization algorithm $\tilde{P}_b$ with privacy budget ϵ . According to Theorem 1 in [17], each perturbed frame satisfies ϵ -differential privacy. Since frames are perturbed independently, according to the parallel composition, MEPS satisfy ϵ -differential privacy. $\square$

Side-channel Attack in MEPS. The selective-encryption policy of MEPS may leak the temporal privacy of events in captured video, which is a known side-channel attack. This attack can be mitigated by established techniques, such as injecting randomness [57] into the selection policy (e.g., encrypting a small, random fraction of high protectability frames). While our primary focus is on attribute inference attacks, we can implement MEPS to defend against the side-channel attack by adopting existing methods.

Privacy analysis scope and feature interactions. MEPS performs privacy analysis at the *frame level* and targets scenarios with *strong frame-level heterogeneity*, such as driving assistance, where privacy relevance varies significantly across

frames. Within each frame, MEPS does not assume strict independence among privacy-relevant features, since the privacy protectability metric is based on the *Shapley-value framework*, which captures feature interactions through *marginal contributions* over feature subsets. At the *sequence level*, however, MEPS does not explicitly model cross-frame correlations, such as repeated appearances of the same sensitive attribute across correlated frames. We acknowledge that such temporal correlations may matter in other scenarios, and leave extending MEPS from frame-level adaptive protection to sequence-level privacy control as future work.

Discussion on homogeneous perturbation. MEPS currently applies differential privacy at the frame level using a uniform perturbation scheme across the whole frame. We adopt this design because it aligns our focus on frame-level adaptive protection under strong frame-level heterogeneity, and it makes privacy accounting more tractable. The homogeneous perturbation also has a limitation: when sensitive content is localized, such as a face or license plate, uniform perturbation may add noise to non-sensitive regions and unnecessarily reduce task utility. A more fine-grained design, such as region-aware or object-aware perturbation, could better improve the privacy-utility tradeoff, but would require reliable region localization and more complex end-to-end privacy accounting. We leave this extension to future work.

6 Implementation and Evaluation

6.1 Prototype Implementation

We implement an MEPS prototype based on Microsoft Rocket [30], a configurable platform for neural network inference on videos. The prototype is implemented in Python and packaged as a Docker image for easy deployment. To use MEPS, applications only need to specify basic configurations, such as the camera USB port or video streaming URL, the application model paths, cloud server IP address, etc., before launching the container.

We deploy an MEPS prototype on an Nvidia Jetson Nano [47], a widely used edge AI platform equipped with a 128-core Maxwell GPU, a quad-core ARM Cortex-A57 CPU, and 4 GB memory running Ubuntu. MEPS is installed on the Jetson Nano using Docker. For cloud-side inference, we use Microsoft Azure infrastructure with dual Nvidia A100 GPUs to meet the computational demands of VFMs. The edge device connects to the cloud via a 1 Gbps LAN cable through a TP-Link router, running OpenWrt, which uses scripted bandwidth control to emulate different network conditions.

MEPS includes a modular library supporting both perturbation-based and encryption-based protection schemes. For perturbation schemes, the profiling parameters include: (1) the privacy leakage level $\epsilon \in [0, 1]$; (2) the frame resolution $R \in 120p, 280p, 540p, 720p$; and (3) the privacy protectability score Pr , discretized into intervals $0-0.03, \dots, 0.96-1.0$.

In our prototype, we adopt DP-Pix [17], which applies Laplace noise to sensitive regions based on the estimated protectability, implemented using the PyTorch distribution library [51].

For encryption schemes, the profiling parameter is the frame resolution $R \in 120p, 280p, 540p, 720p$. We implement encryption using ConcreteML [75], a homomorphic encryption framework developed by ZAMA that automatically converts machine learning models into homomorphic equivalents for private inference.

Deployment platform and limitation. MEPS is designed for lightweight edge deployment and has been implemented on Jetson Nano, a representative resource-constrained edge device. Its practical hardware requirement is application- and configuration-dependent, such as frame resolution and the selected protection scheme. MEPS can adapt to different resource budgets through offline profiling and online adaptation, but deployment on more constrained commodity devices may incur higher latency, which can reduce utility. We have evaluated the scalability of MEPS under different dynamic computation and communication resources in § 6.2.3.

6.2 Evaluation

In this section, we evaluate MEPS performance, aiming to answer the following questions:

Q1: Overall performance. How does MEPS balance task accuracy, privacy leakage, and latency across applications and different sensitive attributes? (§ 6.2.2)

Q2: Scalability. How does our designed system perform under different dynamic computation and communication resources? (§ 6.2.3)

Q3: Ablation study. How do the privacy protection level, protectable-feature threshold, and protectability estimator design affect the overall system performance? (§ 6.2.4)

Q4: End-to-end operation performance. How does our designed system conduct adaptation for real-world applications in runtime? (§ 6.2.5)

6.2.1 Methodology

Applications and datasets. We use three representative video understanding applications to evaluate MEPS and the baselines: 1) *Traffic Report Captioning (TRC)*, detects and comprehensive traffic situations on roads from the videos captured by static traffic cameras, also the driver views in the *Woven Traffic Safety (WTS) Dataset* [34] using TrafficVLM [12] to emphasize detailed behaviors of both vehicles and pedestrians, including accidents. 2) *Driving Assistance (DA)* estimates the driving surroundings and gives driving advice by VisionGPT [67] from videos captured by advanced driver-assistance cameras in *London* [63], and 3) *Intrusion Detection (ID)*, utilizes Video-llama [77] for identifying unauthorized access events from parking lot CCTV surveillance footage within the *Gebhardt* dataset [19].

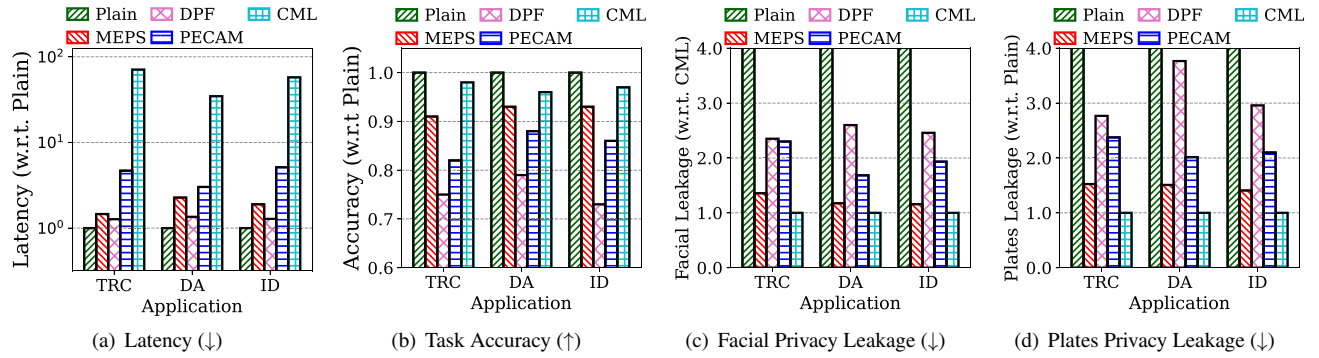


Figure 6: Overall performance under different applications.

Network Trace. We select three public network trace datasets, Belgium 4G/LTE dataset [64], Raca dataset [53] and Ookla Internet Speed Dataset [27] for 4G/LTE, 5G/NR, and WiFi-2.4G, respectively. We also set up a wired network with 200 Mbps to investigate the stable network scenarios.

Privacy Protectability Estimation Model Training. We implement the privacy protectability estimation model with the U-Net model structure, which contains an estimation encoder with four blocks and an estimation decoder with three blocks. We train a privacy protectability estimation model for each application on the corresponding dataset. In each dataset, the proportion of data for training and validation is 80% and 20% for testing, respectively, and only 5% of data have labels. We first pre-train the estimation model with the contrastive loss on the entire dataset. Then, we fine-tune the estimation decoder on the labeled data.

Baselines. We compare MEPS with *open-source* and SOTA video privacy protection approaches.

- **Plain:** It is a no-protection setup that directly transmits video to the cloud to serve as the lower bound of privacy protection and the upper bound of inference accuracy. The Plain baseline is a reference to demonstrate how defense methods affect efficiency and task utility.
- **ConcreteML (CML)** [75]: The edge device encrypts intermediate computational results via HE and transmits them to the cloud. The cloud performs the attention mechanism on these encrypted data and sends the encrypted outputs back. Because CML operates exclusively on integer data, the VFM requires quantization to an integer format before inference can proceed. This methodology allows for secure inference, ensuring the cloud learns nothing about the sensitive input data.
- **DP-Forward (DPF)** [14]: It incorporates additive Laplace noise distributions into embeddings of LLM to reduce the sensitive information to protect privacy while maintaining the accuracy of the inference task.
- **PECAM** [69]: It leverages a Generative Adversarial Network to perform a securely reversible transformation on

the data, which enables the concealing of the sensitive information in the data to achieve privacy protection.

Evaluation Metrics. We evaluate the performance of our proposed system from both accuracy and privacy protection aspects with the following two metrics:

- **Task Accuracy:** For the application TRC, we measure its accuracy with the BLEU [48], which is the word-level similarity. For the applications DA and ID, the F1-score is employed as the task accuracy function since these applications care about both recall and precision.
- **Privacy Leakages:** We quantify privacy leakage as the success rate of a attribute-specific detector or attribute classifier on protected frames, where lower success rates indicate stronger privacy protection. Specifically, we use ArcFace [10] for *Face*, a YOLOv8-based license-plate detector [50] for *License Plate*, FairFace [28] for *Gender*, apparent *Race/Ethnicity*, and apparent *Age Group*, and YOLOv8 vehicle detection [25] with dominant-color classification for *Vehicle Color*.

To better demonstrate the difference in system performance, we plot the figures below *with regard to* (w.r.t.) the Plain.

6.2.2 Overall Performance

Performance of Latency. Figure 6(a) compares the latency of MEPS with other privacy-preserving approaches across three video understanding tasks (TRC, DA, ID), normalized to the Plain (no protection) baseline. MEPS incurs only modest latency overhead (e.g., $1.45\times$ in TRC), whereas full encryption schemes such as CML impose substantial overhead (e.g., $34.67\times$ in DA and up to $70.79\times$ in TRC), and PECAM also incurs notable latency (e.g., $3.54\times$ in DA). Overall, MEPS achieves up to $48.82\times$ lower latency than CML and $1.57\times$ lower latency than PECAM. This improvement stems from MEPS’s adaptive design, which selectively applies expensive encryption only to frames where sensitive information is task-essential, avoiding the high cost of universal encryption. Although the perturbation-only scheme DPF achieves lower latency, MEPS provides a more balanced trade-off between

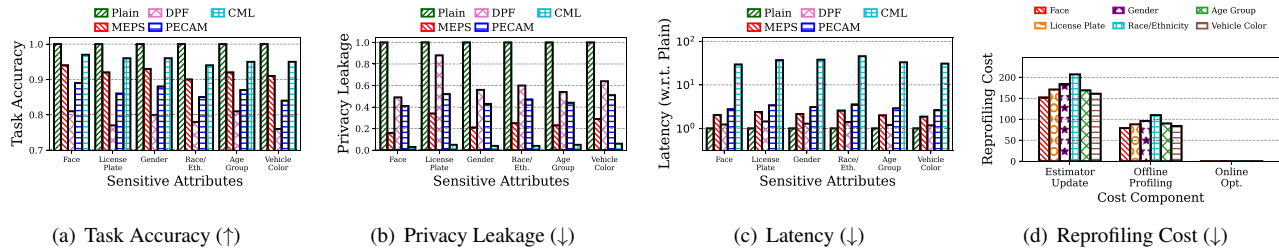


Figure 7: Generalization performance on different sensitive attributes in DA.

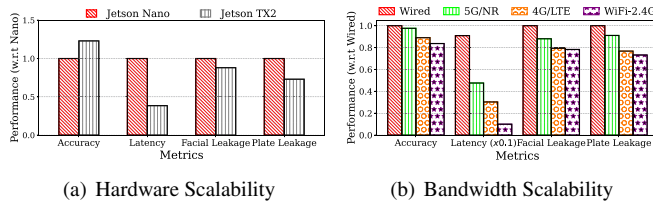


Figure 8: Performance under different resources in DA.

latency, privacy protection, and task accuracy, making it better suited for resource-constrained edge deployments.

Performance of Task Accuracy. Figure 6(b) shows the task accuracy of different methods across the three applications. MEPS consistently achieves accuracy comparable to the Plain baseline and the encryption-heavy CML, with scores above 0.9 (e.g., 0.91 in TRC and 0.93 in both DA and ID). This demonstrates that MEPS preserves task-essential features by selectively encrypting sensitive frames and perturbing non-essential ones. In contrast, the perturbation-only DPF approach suffers notable accuracy degradation across all tasks (e.g., 0.75 in TRC and 0.73 in ID). MEPS also outperforms PECAM in accuracy, and while CML attains slightly higher raw accuracy, MEPS delivers comparable accuracy with substantially lower latency, yielding a more favorable overall trade-off for practical edge–cloud deployments.

Performance of Privacy Protection. Figures 6(c) and 6(d) report privacy leakage for facial information and license plates, respectively. MEPS consistently achieves lower leakage than the perturbation-based DPF and the encryption-based PECAM across all three applications (TRC, DA, ID). For example, in the DA task, MEPS reduces facial leakage by 81.9% and plate leakage by 66.7% relative to the Plain baseline, compared to reductions of 53.8% and 7.5% for DPF, and 57.5% and 50.1% for PECAM. While CML provides the strongest privacy protection and serves as a benchmark, MEPS achieves privacy leakage substantially closer to CML than to DPF or PECAM. Importantly, MEPS delivers this strong privacy protection with significantly lower latency and comparable task accuracy, demonstrating that its adaptive strategy effectively

balances privacy, accuracy, and efficiency without the high overhead of universal encryption.

Performance of Generalization to Different Sensitive Attributes. We evaluate how MEPS behaves when the protected sensitive attributes changes on the DA task. In addition to the default *Face* and *License Plate* attributes used in the main evaluation, we consider four additional sensitive attributes: *Gender*, apparent *Race/Ethnicity*, apparent *Age Group*, and *Vehicle Color*. Particularly, *Race/Ethnicity* and *Age Group* are used only as visual leakage targets inferred from people in the frame rather than verified demographic labels. The results are shown in Fig. 7(a)–(c). Across all six attributes, MEPS maintains high task accuracy of 0.90–0.94, close to Plain and CML, while substantially reducing privacy leakage to 0.16–0.34. In comparison, DPF leaves a leakage rate of 0.49–0.88, and PECAM leaves 0.41–0.52. Meanwhile, MEPS incurs only $1.86\text{--}2.57\times$ latency relative to Plain, which is much lower than the $29.51\text{--}45.71\times$ latency of CML. These results demonstrate that MEPS generalizes across different sensitive attributes while preserving the trade-off among task accuracy, privacy protection, and system latency. Fig. 7(d) further reports the reprofitting cost for adapting MEPS to different sensitive attributes. Updating the privacy protectability estimator and offline protection profiles incurs only one-time costs, while the online optimization cost remains below 0.2 seconds per update. This indicates that MEPS can support different sensitive attributes with modest reprofitting overhead.

6.2.3 Scalability

We evaluate the scalability of MEPS with different computation and communication resources by deploying MEPS on another typical edge device, the Nvidia Jetson TX2, which has about 2.2 times the computational capabilities of the Jetson Nano. We choose DA application for this evaluation. Fig. 8(a) shows the performance of MEPS on Jetson TX2 with regard to Jetson Nano. MEPS demonstrates superior performance in both latency reduction and privacy protection. Specifically, deploying MEPS on the more powerful Jetson TX2 compared to the Jetson Nano results in a significant latency reduction (to approximately 0.38 times of the Nano latency) and improved privacy protection (lower facial and plate

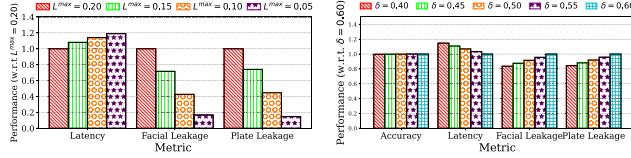


Figure 9: Impact of privacy protection level L^{max} .

Figure 10: Impact of threshold δ in DA.

leakage), while consistently maintaining high task accuracy; this demonstrates the ability of MEPS to effectively leverage enhanced computational resources for better efficiency and privacy. Similar trends are observed when communication resources are reduced, as shown in Fig. 8(b). MEPS exhibits remarkable resilience by sustaining high task accuracy and low privacy leakage across a spectrum of network conditions, from high-bandwidth Wired connections to lower-bandwidth WiFi-2.4G. While latency understandably degrades with decreasing bandwidth (with WiFi showing the highest latency values, approximately 0.91 times of Wired for the scaled metric), the consistent preservation of crucial metrics like accuracy and privacy underscores the adaptability of MEPS and its capacity to deliver reliable core performance even when network resources are limited, affirming its practical utility for real-world edge-cloud deployments.

6.2.4 Ablation Study

Impact of Privacy Protection Level L^{max} . We investigated the influence of the privacy protection level parameter, L^{max} , on system performance by evaluating latency and privacy leakage across its different settings. Fig. 9 illustrates the performance of MEPS on the three applications under varying L^{max} values. The results indicate that task accuracy exhibits a slight decrease as L^{max} is reduced. Notably, when the privacy leakage level (presumably a target or outcome influenced by L^{max}) decreases from 0.2 to 0.05, the system latency remains nearly constant. Furthermore, Fig. 9 clearly demonstrates a strong correlation between privacy leakage and the configured L^{max} , with smaller L^{max} values resulting in enhanced privacy protection (i.e., lower leakage). These findings suggest that selecting a smaller L^{max} in practical deployments is a viable strategy to achieve robust privacy protection without incurring significant latency.

Impact of Protectable-feature Threshold δ . We evaluate how the protectable-feature threshold δ affects MEPS on the DA task. In our defined privacy protectability metric, the threshold δ determines whether a feature is regarded as protectable. A smaller δ applies a stricter criterion, causing fewer privacy-overlapping features to be treated as protectable and leading MEPS to select stronger protection more frequently. We vary δ from 0.40 to 0.60 with a step size of 0.05 and report the normalized accuracy, latency, and privacy leakage in Fig. 10. When δ decreases from 0.60 to 0.40, i.e., a 33.3%

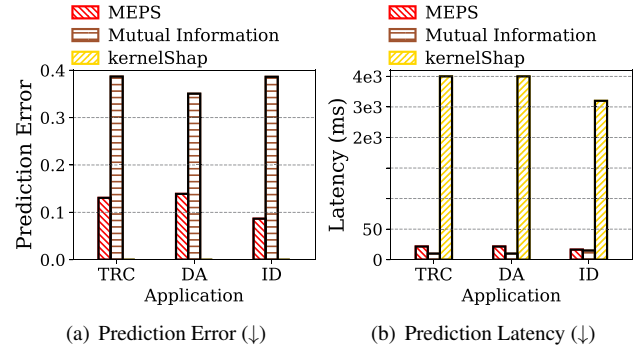


Figure 11: Impact of Protectability estimator.

decrease, latency increases by 14.7% because more frames are assigned to stronger protection configurations. However, the task accuracy remains stable, decreasing only slightly from 0.931 to 0.928. Meanwhile, facial leakage and plate leakage are reduced to 83.6% and 84.4% of their values under $\delta = 0.60$, respectively. These results show that MEPS is robust to the choice of δ within the tested range. The threshold mainly provides a controllable privacy-latency trade-off, where smaller values strengthen privacy protection with modest latency overhead.

Impact of Protectability Estimator Selection. Our investigation focused on the impact of different designs of the Protectability Estimator by comparatively analyzing prediction accuracy and latency. We considered three methodologies: (i) *KernelSHAP* [56], which facilitates exact computation of privacy protectability; (ii) *Mutual Information*, a computationally simpler approximation technique; and (iii) the proposed contrastive learning-based estimator for MEPS. The empirical results are depicted in Fig. 11. We can observe that KernelSHAP achieves the highest estimation accuracy but incurs prohibitively high latency for practical video understanding workloads. In contrast, the Mutual Information approach has minimal latency but suffers large prediction errors, exceeding KernelSHAP by 65.4%–71.1%. The MEPS estimator achieves a favorable balance, incurring only a 6.8% average increase in error relative to KernelSHAP while reducing latency by more than 138 $\times$. Furthermore, when compared to Mutual Information, it reduces prediction error by 32.6%–48.8% while operating at a comparable latency. These findings collectively underscore the superior cost-effectiveness of the contrastive learning model for performing privacy protectability estimation within the MEPS architecture.

6.2.5 End-to-end Operation Performance

In this section, we evaluate MEPS in an end-to-end operation on the Driving Assistance task, utilizing VisionGPT [67]. In this task, video frames are processed to provide driving advice, including environmental descriptions, identification

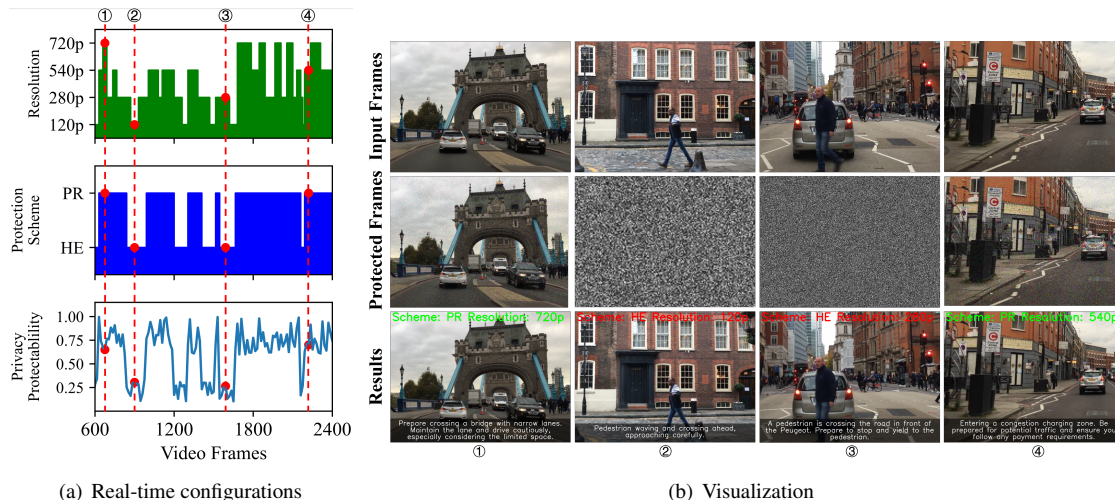


Figure 12: The end-to-end operations of MEPS.

of potential traffic risks, and suggestions for driver behavior. Concurrently, human features, such as facial attributes, are treated as private data, making them vulnerable to attribute inference attacks. We simulate this application with one-hour video traces *London* captured by the cameras on drivers' view using 4G network trace [27]. Subsequently, we present a segment encompassing four minutes, equivalent to 2400 frames, demonstrating the internal variables and configurations inherent to the MEPS. This exhibition is visualized through Fig. 12. In Fig. 12(a), we can see the three subfigures represent the Protectability Scores, the Protection Scheme, and the Resolution of the profile MEPS chooses, respectively. Notably, our observations confirm the adeptness of MEPS in accurately selecting the appropriate protection scheme and resolution configuration for the given video frames. For example, in video frame ①, the driver is approaching a bridge. The privacy-sensitive information in this frame includes the license plates of the cars ahead, resulting in high privacy protectability. This frame can be protected using the perturbation scheme. In contrast, in video frames ② and ③, pedestrians are waving and attempting to cross the road. This scenario creates a significant overlap between privacy attributes and inference features, leading to noticeably lower protectability. As a result, MEPS applies an encryption scheme to protect these frames. Finally, in video frame ④, with almost non-privacy features, MEPS uses the perturbation scheme with 540p resolution, effectively demonstrating its ability to adapt protection schemes based on the protectability of frames.

7 Related Work

Privacy-preserving video understanding. Video understanding applications typically rely on the generalization and reasoning capability of video foundation models, which are

trained with large-scale datasets and models with billion-scale parameters, and are computationally intensive when performing inference on resource-constrained edge devices. When leveraging cloud computing resources, the privacy leakage problem occurs when transmitting the videos to the cloud for inference. Existing works attempt to solve this problem mainly from two aspects: 1) edge-side computing with video foundation models compressed with techniques such as quantization [11, 20, 36, 58], pruning [2, 40, 44, 66, 71], etc., which can avoid video transmission to protect privacy at the cost of video task accuracy, and cannot be adopted to apply on our targeted video understanding applications like driving assistant with high accuracy requirement; 2) edge-cloud privacy-preserving computing with various privacy-preserving inference schemes such as Homomorphic Encryption is adopted in ConcreteML [75] to perform foundation model inference with encrypted input data, and a generative adversarial network (cycleGAN) is designed in PECAM [69] to encrypt the sensitive information in video frames before transmission, and well-designed DP noises are applied in DP-Forward [13] to perturb the sensitive information contained in the inference data. Our work falls into the edge-cloud privacy-preserving computing category. Compared to existing edge-cloud privacy-preserving methods, our design can achieve a higher video understanding task accuracy (compared to perturbation-based method DP-forward) and a significantly lower application latency (compared to encryption-based methods ConcreteML and PECAM at the same time when deploying on the resource-constrained edge devices).

Information theory for performance quantification in neural network models. In the research area of explainable AI, studies try to explain the reasoning and decision-making processes of neural network models. To generate the explanations of the model performance, feature importance is used as a

metric to differentiate the impact of each feature on the accuracy of the model prediction [5] [8]. The more information contained in the feature useful to the model prediction, the higher the feature importance scores.

In the research area of transfer learning, studies try to use the knowledge learned from a source task to improve the performance of a related but different target task. To capture the relationship between the source task and the target task, a transferability metric [59] [60] [61] is designed to measure to what extent the knowledge or learned features from the source task can be effectively applied to a different but related target task. A high transferability indicates that the knowledge gained from the source task can be successfully transferred and utilized in the target task.

Our proposed privacy protectability metric is also information-based, but it uniquely addresses the domain of privacy in video understanding. It establishes a theoretical foundation for quantitatively distinguishing between encryption and perturbation schemes based on their efficacy in protecting a given video frame. This metric enables the development of our MEPS system, which is specifically designed for scenarios where, although most video frames contain sensitive information, only a small fraction of this sensitive data is essential for the video understanding task.

8 Conclusion

This paper introduced MEPS, a novel privacy-preserving video understanding system for edge-cloud architectures, specifically designed to manage the trade-off between privacy, accuracy, and efficiency on resource-constrained edge device. Based on our observation that only a small portion of sensitive information in video frames is typically essential for the video understanding tasks, we propose MEPS to adaptively apply encryption on selected frames and leverages lightweight perturbation for most frames. We developed a new privacy protectability metric to guide this differentiated protection. For efficient operation, MEPS utilizes an offline profiling and online adaptation strategy, allowing it to adjust to varying frame characteristics and minimize task latency. A theoretical privacy analysis proves MEPS satisfies ϵ -DP in our implementation. And comprehensive evaluations of our MEPS prototype demonstrates its superior performance as compared to existing systems.

Acknowledgments

We thank the USENIX Security reviewers for their detailed feedback, with special thanks to our anonymous shepherd for thoughtful guidance that substantially improved the final manuscript. The work of Dan Wang is supported in part by RGC GRF 15201322, 15230624, 15239925, CRF C5020-25GF, ITC ITS/052/23MX, and a HKUST start up grant.

Ethical Considerations

Stakeholder Analysis and Process Impact. This paper involves four primary stakeholder groups: 1) *Data subjects*, which are individuals whose visual information (e.g., faces, bodies, license plates) may appear in video data processed by edge-cloud video understanding systems. During the research process, these individuals rely on system designers and researchers to ensure that their sensitive attributes are not exposed; 2) *Data Owners and System Operators*, who deploy cameras and edge-cloud video systems (e.g., transportation operators, infrastructure managers). They face competing requirements for privacy protection, system efficiency, and task accuracy, as well as legal and regulatory constraints related to data protection; 3) *Researchers and Practitioners*, who design, evaluate, and deploy video foundation models. They depend on rigorous, reproducible methods and clear threat models to assess privacy risks and defenses; 4) *Society at Large*, whose privacy concerns are affected by the widespread deployment of video understanding systems, particularly in public or semi-public spaces. During the research process, we exclusively use **publicly available datasets** that have been widely studied in prior work. We do not collect new video data, do not interact with human subjects, and do not deploy MEPS in real-world operational settings. As such, no individuals were directly exposed to additional privacy risks as a result of our experiments. We note that the public documentation of the datasets does not clearly specify whether all visible subjects were aware of, or consented to, the original recording and subsequent research use of the data. These datasets may contain sensitive visual content, including faces, license plates, and other identifiable cues. While newly collecting similar data would raise significant ethical concerns, our use introduces limited additional risk because the data sources are already publicly available and have been used for research or public viewing. We do not collect new videos, identify individuals, link subjects across videos, or infer personal information beyond aggregate privacy-leakage measurements. We also do not redistribute raw videos in our artifact; users must obtain them from the original sources and comply with the corresponding terms of use. Given these safeguards, and considering that MEPS is designed to reduce privacy exposure in edge-cloud video inference, we believe that the research benefits outweigh the limited additional risks in this controlled evaluation setting.

Impact of the Research. This research has both positive and negative impacts on stakeholders.

Positive impacts: 1) MEPS applies either encryption or perturbation based privacy protection mechanism on the content of each video frame. This improves privacy protection for *data subjects* compared to systems that transmit raw video or rely solely on perturbation; 2) MEPS lowers the barrier for deploying privacy-preserving video systems on resource-constrained edge devices by balancing privacy, accuracy, and

efficiency, which enables *system operators* to adopt stronger privacy protections without prohibitive computational cost; 3) Formalization of privacy protectability and threat model provide a principled framework for analyzing privacy–utility trade-offs in video systems, supporting reproducible investigations for *researchers*.

Negative impacts: 1) Potential misuse of the proposed selectively protecting technique can have impacts on *data subjects and society*. This risk is inherent to many privacy-preserving system designs; 2) Selective protection policies may introduce side-channel leakage (e.g., inferring events from when encryption is triggered), which could reveal sensitive temporal information if left unmitigated.

Mitigations. To address these risks, we implement and document the following mitigation strategies:

Methodological Mitigations (Implemented). 1) Formal Privacy Guarantees: Frames protected via encryption inherit cryptographic semantic security, while perturbed frames satisfy differential privacy guarantees. This limits both direct leakage and inference-based attacks within our threat model; 2) Restricted Threat Model and Scope: We explicitly limit our study to honest-but-curious cloud servers and attribute inference attacks, avoiding invasive adversarial settings that could increase real-world risk; 3) Use of public datasets only: we avoid introducing new privacy risks or exposing previously private data by only use publicly available datasets.

Recommended Deployment Measures (Not Implemented). 1) Side-Channel Mitigation: We recommend incorporating randomized encryption policies or padding strategies to reduce temporal side-channel leakage in real-world deployments; 2) Governance and Auditing: We encourage system operators to pair MEPS with organizational safeguards, such as auditing and access controls, to prevent misuse.

Justification for Research and Publication. This work addresses a critical challenge in modern edge–cloud video systems: achieving meaningful privacy protection without sacrificing accuracy or feasibility on constrained devices. From a beneficence perspective, MEPS demonstrably reduces privacy leakage while improving system efficiency. From a respect-for-persons perspective, it minimizes unnecessary exposure of sensitive visual attributes. Although the work has dual-use potential, we believe that transparent publication, along with explicit discussion of limitations and mitigations, better serves the public interest than withholding these findings. We therefore conclude that conducting and publishing this research is ethically justified.

Open Science

In compliance with the Open Science Policy, we will releasing the artifacts necessary to reproduce the main results of this paper. Specifically, the video datasets and the video foundation models (i.e., TrafficVLM, VisionGPT, and Video-llama) can be obtained from their official repositories. We will further

release the implementation of MEPS, including the privacy protectability estimation model, protection control logic, and evaluation scripts. The released artifacts is available through Zenodo² to support future research.

References

- [1] Abbas Acar, Hidayet Aksu, A. Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys*, 51(4):1–35, July 2018.
- [2] Atakan Aral, Melike Erol-Kantarci, and Ivona Brandić. Staleness control for edge data analytics. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(2), June 2020.
- [3] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proc. of the IEEE/CVF ICCV’19*, Seoul, Korea, Oct. 2019.
- [4] Strobl Carolin, Boulesteix Anne-Laure, Kneib Thomas, Augustin Thomas, and Zeileis Achim. Conditional variable importance for random forests. *ACM Computing Surveys*, 9(1):307, July 2008.
- [5] Amnon Catav, Boyang Fu, Yazeed Zoabi, Ahuva Libi Weiss Meilik, Noam Shomron, Jason Ernst, Sri-ram Sankararaman, and Ran Gilad-Bachrach. Marginal contribution feature importance - an axiomatic approach for explaining data. In *Proceedings of the 38th International Conference on Machine Learning*, Virtual Event, July 2021.
- [6] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debaised contrastive learning. In *Proc. of NeurIPS’20*, Virtual Event, 33 2020.
- [7] Yu-Neng Chuang, Guanchu Wang, Fan Yang, Quan Zhou, Pushkar Tripathi, Xuanning Cai, and Xia Hu. CoRTX: Contrastive framework for real-time explanation. In *Proc. of ICLR’23*, Kigali Rwanda, May 2023.
- [8] Ian C. Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *Proc. of NeurIPS’20*, Vancouver, Canada, Dec. 2020.
- [9] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng.

²<https://zenodo.org/records/20342549>

- A survey on multimodal large language models for autonomous driving. In *Proceedings of IEEE/CVF WACV'24 Workshops*, pages 958–979, January 2024.
- [10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. of the IEEE/CVF CVPR'19*, Long Beach, USA, June 2019.
 - [11] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems (NIPS'24)*, 36, 2024.
 - [12] Quang Minh Dinh, Minh Khoi Ho, Anh Quan Dang, and Hung Phong Tran. Trafficvlm: A controllable visual language model for traffic video captioning. In *Proc. of the IEEE/CVF CVPR'24*, pages 7134–7143, 2024.
 - [13] Minxin Du, Xiang Yue, Sherman S. M. Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *Proceedings of ACM SIGSAC CCS'23*, page 2665–2679, New York, NY, USA, 2023.
 - [14] Minxin Du, Xiang Yue, Sherman SM Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *Proc. of ACM SIGSAC'23*, 2023.
 - [15] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, Jan. 2023.
 - [16] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, page 265–284, 2006.
 - [17] Liyue Fan. Image pixelization with differential privacy. In *Data and Applications Security and Privacy XXXII*, page 148–162, July 16–18 2018.
 - [18] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv:2006.03659*, 2020.
 - [19] Gebhardt Insurance Group. Gebhardt Insurance Traffic Cam Round Trip Bike Shop.
 - [20] Cong Guo, Jiaming Tang, Weiming Hu, Jingwen Leng, Chen Zhang, Fan Yang, Yunxin Liu, Minyi Guo, and Yuhao Zhu. Olive: Accelerating large language models via hardware-friendly outlier-victim pair quantization. In *Proceedings of ISCA '23*, 2023.
 - [21] Chien Hao-Jen, Khalili Hossein, Hass Amin, and Sehatbakhsh Nader. Enc2: Privacy-preserving inference for tiny iots via encoding and encryption. In *Proc. of ACM MobiCom'23*, Madrid, Spain, Oct. 2023.
 - [22] Long Huang and Chen Wang. Sniffing location privacy of video conference users using free audio channels. In *Proc. of IEEE S&P'25*, Los Alamitos, CA, May 2025.
 - [23] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1895–1912, Santa Clara, CA, Aug. 2019.
 - [24] Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-time shapley value estimation. In *Proc. of ICLR'22*, Virtual Event, April 2022.
 - [25] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.
 - [26] Ehsanul Kabir, Lucas Craig, and Shagufta Mehnaz. Disparate privacy vulnerability: targeted attribute inference attacks and defenses. In *Proceedings of the 34th USENIX Conference on Security Symposium*, 2025.
 - [27] Kaggle. Ookla internet speed dataset, 2022.
 - [28] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1548–1558, January 2021.
 - [29] Chris Kelly, Luhui Hu, Bang Yang, Yu Tian, Deshun Yang, Cindy Yang, Zaoshan Huang, Zihao Li, Jiayin Hu, and Yuexian Zou. Visiongpt: Vision-language understanding agent using generalized multimodal framework. *arXiv*, 2024.
 - [30] Avi Kewalramani. Live Video Analytics with Microsoft Rocket for reducing edge compute costs, 2020.
 - [31] Mehrdad Khani, Ganesh Ananthanarayanan, Kevin Hsieh, Junchen Jiang, Ravi Netravali, Yuanchao Shu, Mohammad Alizadeh, and Victor Bahl. RECL: Responsive Resource-Efficient continuous learning for video analytics. In *Proc. of USENIX NSDI'23*, Boston, MA, Apr. 2023.
 - [32] Hamza Kheddar. Transformers and large language models for efficient intrusion detection systems: A comprehensive survey. *arXiv preprint arXiv:2408.07583*, 2024.
 - [33] Jaehong Kim, Youngmok Jung, Hyunho Yeo, Juncheol Ye, and Dongsu Han. Neural-enhanced live streaming:

- Improving live video ingest via online learning. In *Proc. of ACM SIGCOMM'20*, New York, NY, Aug. 2020.
- [34] Quan Kong, Yuki Kawana, Rajat Saini, Ashutosh Kumar, Jingjing Pan, Ta Gu, Yohei Ozao, Balazs Opra, David C. Anastasiu, Yoichi Sato, and Norimasa Kobori. Wts: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding, 2024.
- [35] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *Proc. of IEEE/CVF CVPR'19 workshops*, Long Beach, CA, USA, June 2019.
- [36] Yanjing Li, Sheng Xu, Baochang Zhang, Xianbin Cao, Peng Gao, and Guodong Guo. Q-vit: Accurate and fully quantized low-bit vision transformer. *Advances in neural information processing systems (NIPS'24)*, 35:34451–34463, 2022.
- [37] Yuelin Liu, Huangxun Chen, and Zhice Yang. Enforcing end-to-end security for remote conference applications. In *Proc. of IEEE S&P'24*, Los Alamitos, CA, USA, May 2024.
- [38] Rui Lu, Siping Shi, Dan Wang, Chuang Hu, and Bihai Zhang. Preva: Protecting inference privacy through policy-based video-frame transformation. In *Proc. of IEEE/ACM 7th SEC'22*, Seattle, USA, Dec. 2022.
- [39] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, Jan. 2020.
- [40] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems (NIPS'23)*, 36:21702–21720, 2023.
- [41] Neelu Madan, Andreas Møgelmoose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*, 2024.
- [42] Frank D. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, page 19–30, 2009.
- [43] Shaguftha Mehnaz, Sayanton V. Dibbo, Ehsanul Kabir, Ninghui Li, and Elisa Bertino. Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In *USENIX Security'22*, Boston, USA, Aug. 2022.
- [44] Xupeng Miao, Gabriele Oliaro, Xinhao Cheng, Mengdi Wu, Colin Unger, and Zhihao Jia. Flexllm: A system for co-serving large language model inference and parameter-efficient finetuning. *arXiv preprint arXiv:2402.18789*, 2024.
- [45] Fatemehsadat Mireshghallah, Mohammadkazem Taram, Prakash Ramrakhiani, Ali Jalali, Dean Tullsen, and Hadi Esmaeilzadeh. Shredder: Learning noise distributions to protect inference privacy. In *Proc. of ACM ASPLOS'20*, Lausanne, Switzerland, Mar. 2020.
- [46] Ravi Teja Mullapudi, Steven Chen, Keyi Zhang, Deva Ramanan, and Kayvon Fatahalian. Online model distillation for efficient video inference. In *Proc. of the IEEE/CVF ICCV'19*, Los Alamitos, USA, Nov. 2019.
- [47] NVIDIA. Nvidia jetson nano developer kit, 2023.
- [48] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [49] Rishabh Poddar, Ganesh Ananthanarayanan, Srinath Setty, Stavros Volos, and Raluca Ada Popa. Visor: Privacy-Preserving video analytics as a cloud service. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1039–1056, Aug. 2020.
- [50] Mike Polinowski. YOLOV8 License Plate Detection | Mike Polinowski, 9 2023.
- [51] PyTorch. PyTorch documentation — PyTorch 2.0 documentation, 2023.
- [52] Wenjie Qu, Yuguang Zhou, Yongji Wu, Tingsong Xiao, Binhang Yuan, Yiming Li, and Jiaheng Zhang. Prompt inversion attack against collaborative inference of large language models. In *Proc. of IEEE S&P'25*, Los Alamitos, CA, May 2025.
- [53] Darijo Raca, Dylan Leahy, Cormac J. Sreenan, and Jason J. Quinlan. Beyond throughput, the next generation: A 5g dataset with channel and context metrics. In *Proc. of MMSys'20*, 2020.
- [54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. of MICCAI'15*, Munich, Germany, Oct. 2015.
- [55] Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *Proc. of IEEE/CVF CVPR'19*, Long Beach, USA, June 2019.

- [56] SHAP. kernalshap, 2023.
- [57] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying location privacy. In *2011 IEEE Symposium on Security and Privacy*, 2011.
- [58] Yu Sun, Gaojian Xiong, Jianhua Liu, Zheng Liu, and Jian Cui. Tsqp: Safeguarding real-time inference for quantization neural networks on edge devices. In *Proc. of IEEE S&P'25*, Los Alamitos, CA, USA, May 2025.
- [59] Yang Tan, Yang Li, and Shao-Lun Huang. Otce: A transferability metric for cross-domain cross-task representations. In *Proc. of the IEEE/CVF CVPR*, Virtual Event, June 2021.
- [60] Xinyi Tong, Xiangxiang Xu, Shao-Lun Huang, and Lizhong Zheng. A mathematical framework for quantifying transferability in multi-source transfer learning. In *Proc. of NeurIPS'21*, Virtual Event, Dec. 2021.
- [61] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proc. of IEEE/CVF ICCV'19*, Los Alamitos, USA, Nov. 2019.
- [62] Anh-Tu Tran, The-Dung Luong, Jessada Karnjana, and Van-Nam Huynh. An efficient approach for privacy preserving decentralized deep learning models based on secure multi-party computation. *Neurocomputing*, 422:245–262, January 2021.
- [63] J Utah. London 4k - skyscraper district drive - city of london, 2019.
- [64] J. vanderHooft. 4g/lte bandwidth logs, November 2011.
- [65] Han Wang, Shangyu Xie, and Yuan Hong. Videodp: A flexible platform for video analytics with differential privacy. *Proceedings on Privacy Enhancing Technologies*, 2020(4):277–296, 2020.
- [66] Hanrui Wang, Zhekai Zhang, and Song Han. Spatten: Efficient sparse attention architecture with cascade token and head pruning. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 97–110. IEEE, 2021.
- [67] Hao Wang, Jiayou Qin, Ashish Bastola, Xiwen Chen, John Suchanek, Zihao Gong, and Abolfazl Razi. Visiongpt: Llm-assisted real-time anomaly detection for safe visual navigation. *arXiv preprint arXiv:2403.12415*, 2024.
- [68] WAYVE. LINGO-1: Exploring Natural Language for Autonomous Driving, 2023.
- [69] Hao Wu, Xuejin Tian, Minghao Li, Yunxin Liu, Ganesh Ananthanarayanan, Fengyuan Xu, and Sheng Zhong. Pecam: privacy-enhanced video streaming and analytics via securely-reversible transformation. In *Proc. of ACM MobiCom'21*, New Orleans, USA, Mar. 2021.
- [70] Xiaoyu Xia, Ziqi Wang, Ruoxi Sun, Bowen Liu, Ibrahim Khalil, and Minhui Xue. Edge unlearning is not "on edge"! an adaptive exact unlearning system on resource-constrained devices. In *Proc. of IEEE S&P'25*, Los Alamitos, CA, May 2025.
- [71] Mengwei Xu, Tiantu Xu, Yunxin Liu, and Felix Xiaozhu Lin. Video analytics with zero-streaming cameras. In *Proc. of USENIX ATC'21*, Virtual Event, July 2021.
- [72] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024.
- [73] Guang Yan, Yuhui Zhang, Zimu Guo, Lutan Zhao, Xiaojun Chen, Chen Wang, Wenhao Wang, Dan Meng, and Rui Hou. Comet: Accelerating private inference for large language model by predicting activation sparsity. In *Proc. of IEEE S&P'25*, Los Alamitos, CA, USA, May 2025.
- [74] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [75] Zama. Concrete ML: a privacy-preserving machine learning library using fully homomorphic encryption for data scientists, 2022. <https://github.com/zama-ai/concrete-ml>.
- [76] Chengru Zhang, Xiao Yang, David Oswald, Mark Ryan, and Philipp Jovanovic. Eva: Efficient privacy-preserving proof of authenticity for lossily encoded videos. In *Proc. of IEEE S&P'25*, Los Alamitos, CA, USA, May 2025.
- [77] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.
- [78] Xiaohan Zhang, Haoqi Ye, Ziqi Huang, Xiao Ye, Yinzhi Cao, Yuan Zhang, and Min Yang. Understanding the (in)security of cross-side face verification systems in mobile apps: A system perspective. In *Proc. of IEEE S&P'23*, 2023.

- [79] Xingyu Zhao, Wei Huang, Xiaowei Huang, Valentin Robu, and David Flynn. Baylime: Bayesian local interpretable model-agnostic explanations. In *Proc. of UAI'21*, Virtual Event, July 2021.
- [80] Chang Zhou, Jianxin Ma, Jianwei Zhang, Jingren Zhou, and Hongxia Yang. Contrastive learning for debiased candidate generation in large-scale recommender systems. In *Proc. of ACM SIGKDD'21*, Madrid, Spain, Jul. 2022.