STAR: Semantic-Traffic Alignment and Retrieval for Zero-Shot HTTPS Website Fingerprinting

Yifei Cheng[1,2], Yujia Zhu[1,2](✉), Baiyang Li[1,2], Xinhao Deng[3], Yitong Cai[1,2], Yaochen Ren[1,2], Qingyun Liu[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China
Email: {chengyifei, zhuyujia, libaiyang, caiyitong, renyaochen, liuqingyun}@iie.ac.cn, xinhaodeng.thu@gmail.com

***Abstract*—Modern HTTPS mechanisms such as Encrypted Client Hello (ECH) and encrypted DNS improve privacy but remain vulnerable to website fingerprinting (WF) attacks, where adversaries infer visited sites from encrypted traffic patterns. Existing WF methods rely on supervised learning with site-specific labeled traces, which limits scalability and fails to handle previously unseen websites. We address these limitations by reformulating WF as a zero-shot cross-modal retrieval problem and introducing STAR. STAR learns a joint embedding space for encrypted traffic traces and crawl-time logic profiles using a dual-encoder architecture. Trained on 150K automatically collected traffic–logic pairs with contrastive and consistency objectives and structure-aware augmentation, STAR retrieves the most semantically aligned profile for a trace without requiring target-side traffic during training. Experiments on 1,600 unseen websites show that STAR achieves 87.9% top-1 accuracy and 0.963 AUC in open-world detection, outperforming supervised and few-shot baselines. Adding an Adapter with only four labeled traces per site further boosts top-5 accuracy to 98.8%. Our analysis reveals intrinsic semantic–traffic alignment in modern web protocols, identifying semantic leakage as the dominant privacy risk in encrypted HTTPS traffic. We release STAR's datasets and code to support reproducibility and future research[1].**

***Index Terms*—Website fingerprinting, Zero-shot learning, Cross-modal retrieval**

I. Introduction

As modern HTTPS evolves, traditional protocol-visible identifiers such as Server Name Indication (SNI) and DNS queries are increasingly concealed by mechanisms like Encrypted Client Hello (ECH) [1] and encrypted DNS [2]. This shift limits the effectiveness of conventional web inference techniques that rely on such metadata [3]. However, even when both payloads and headers are fully encrypted, traffic traces still reveal structural patterns—such as packet sizes, timing, and burst behaviors—that reflect the underlying resource structure of websites [4]. Website fingerprinting (WF) approaches [5]–[10] exploit these residual features to infer the site being visited, without requiring access to any plaintext identifiers. In this context, WF has emerged as one of the few remaining passive techniques for web-level inference under full encryption.

Existing WF approaches, however, face fundamental limitations that hinder their scalability and practicality for real-world deployment. Specifically: (i) Traffic drift. Website content evolves dynamically over time [11], necessitating frequent recollection of labeled traffic data and retraining of models; (ii) Limited recognition capability. Current supervised learning–based approaches can only identify previously known websites, lacking the ability to generalize to newly emerging sites. These challenges significantly restrict the applicability of WF in operational settings.

To address these limitations, we introduce a novel approach that jointly exploits **traffic modality** features and **logical modality** features to enable scalable and generalizable WF against previously unseen websites. Logical modality features (e.g., URI lengths, response sizes, and protocol versions) can be automatically extracted through large-scale web crawling, capturing resource-level attributes that describe a website's semantic structure. By mapping both traffic modality features and logical modality features into a shared embedding space, we construct a large-scale website fingerprint database grounded in logical representations. Consequently, the task of identifying a website from unseen traffic can be reformulated as a cross-modal retrieval problem, wherein traffic modality features are matched to the most semantically relevant logical modality features stored in the fingerprint database.

We instantiate this formulation through **STAR** (Semantic–Traffic Alignment and Retrieval), a dual-encoder architecture that jointly embeds logic and traffic modalities into a unified latent space. STAR is trained on over 150K automatically collected logic–traffic pairs using a contrastive learning objective, with additional auxiliary losses to improve intra-class consistency and discriminability. To further enhance robustness against website evolution, we introduce a structure-aware data augmentation mechanism that perturbs both modalities in a semantically consistent manner. During inference, STAR retrieves the most semantically aligned logic profile for an encrypted traffic sample, using cosine similarity in the shared embedding space. This design enables zero-shot classification of encrypted traces with no prior access to traffic from target websites.

Beyond the system design, we also conduct a systematic

[1] https://github.com/2654400439/STAR-Website-Fingerprinting

investigation into **why semantic–traffic alignment is possible**. We identify three core alignment anchors—on the request side, response side, and transport protocol—each capturing a consistent mapping between traffic features and high-level website structures (§III-B). These anchors stem from the inherent design of modern web protocols (e.g., header compression, layered transport) and serve as empirical foundations for learning cross-modal associations, further supported by modality-level analyses of discriminability, stability, and cross-modal correlation (§V-C). Together, these findings not only validate the design rationale behind STAR, but also provide foundational evidence that cross-modal modeling is both feasible and effective for fingerprinting encrypted web traffic.

In summary, our contributions are as follows:

- We formalize **zero-shot website fingerprinting** under HTTPS as a cross-modal retrieval task, removing the need for per-site traffic collection and supporting generalization to unseen websites.
- We present **STAR**, the *first* dual-modality system that aligns crawl-time semantic logic with encrypted traffic traces through contrastive learning and structure-aware augmentation.
- We provide **empirical and statistical evidence** that semantic–traffic alignment is structurally grounded, revealing significant correlations across multiple alignment anchors.
- We perform **extensive closed- and open-world evaluations**, showing that zero-shot STAR achieves 87.9% accuracy over 1,600 unseen sites and a 0.963 AUC in an open-world test with millions of distractors, outperforming state-of-the-art supervised and few-shot baselines.
- We release the **STAR-200K dataset and source code** to facilitate future research on semantic inference under encrypted protocols [12].

These results highlight the feasibility of zero-shot traffic-based identification and demonstrate that **semantic leakage**, rather than header visibility, now constitutes the principal privacy risk in the encrypted web.

II. Background and Threat Model

A. Website Fingerprinting

Website fingerprinting (WF) infers a user's visited website by analyzing features of encrypted traffic—such as packet lengths, directions, and timing patterns. Introduced formally by Hintz [6], early WF methods used handcrafted features and classical classifiers [7], [13]. The rise of deep learning significantly boosted performance: models like Deep Fingerprinting (DF) [8] achieved high closed-world accuracy via CNNs, and later work explored GNNs [14], Transformers [15], and diffusion models [16].

Recent work has revisited WF under mainstream HTTPS, revealing that even minimal protocol interactions can leak identifying patterns. For example, Cebere et al. [17] analyzed leakage across TLS stages; Gao et al. [14] constructed resource graphs to model site structure; Cheng et al. [9] showed that HTTP version features form unique site-level sequences; Shen et al. [18] demonstrated the use of prior fingerprints to filter obfuscation flows. Other works revealed structural leakage in HTTP/3 and DoH traffic [19], [20].

To reduce the reliance on large labeled datasets, recent work has explored data-efficient strategies under *low-data regimes*. Few-shot approaches use contrastive learning to pretrain DF-style encoders [21], [22] or apply KNN over application-layer features [9]; generative methods augment training data with synthetic traces [23]. Some studies extend to zero-shot scenarios across network conditions (e.g., VPN changes) [15], [24], but still assume access to traffic from target websites.

This limitation motivates our proposed cross-modal approach, eliminating the necessity for target-site traffic collection entirely.

B. Threat Model and Problem Definition

We consider a standard passive adversary in the website fingerprinting (WF) setting [8], [9], [18]. The attacker resides on the network path between the user and the web server—such as an ISP or router—and is able to observe encrypted traffic but cannot modify, delay, inject, or decrypt any packets. The attacker's goal is to determine whether the user is visiting a monitored website and, if so, identify which one.

Unlike traditional WF approaches that formulate this task as a site-specific classification problem, we adopt a new threat model where the attacker performs **cross-modal retrieval** between encrypted traffic traces and semantic website representations. This allows the attacker to recognize previously unseen websites based on semantic-traffic alignment, without requiring labeled traffic samples for each monitored site.

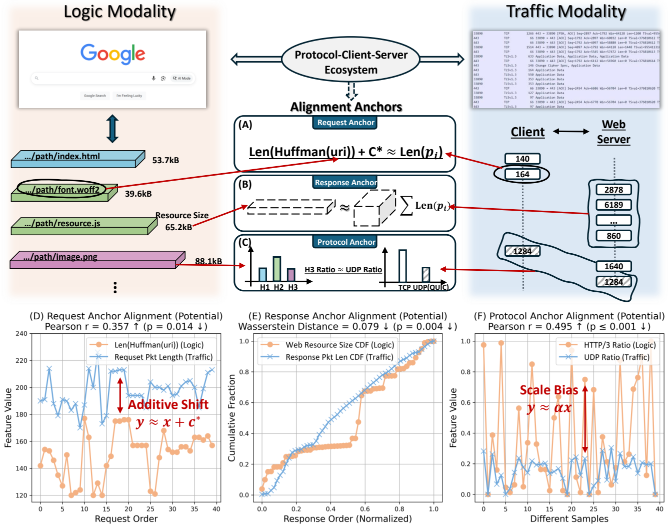
III. Core Observations and Cross-Modal Alignment Hypothesis

A. Cross-Modal Design Principles

Unlike typical cross-modal learning tasks [25] that align well-structured modalities such as natural language and images, our setting involves the alignment between encrypted traffic traces and abstracted semantic representations of websites. This **non-traditional modality pairing** introduces new challenges, particularly in the construction of effective input representations.

To guide the design of both traffic and logic modalities, we summarize three key principles that effective cross-modal representations should satisfy, inspired by prior work in fingerprinting and multi-modal retrieval:

- **P1: Discriminability (intra-modality)**
  Each modality should encode features that allow websites to be distinguished from one another, enabling the model to separate classes in both semantic and traffic spaces.
- **P2: Stability (intra-modality)**
  The modality representations should remain relatively consistent across repeated visits to the same site under similar conditions. High intra-class consistency is critical for generalization.
- **P3: Alignability (cross-modality)**



```
33890      TCP        1266 443 → 33890 [PSH, ACK] Seq=2897 Ack=1792 Win=64128 Len=1200 TSval=955
443        TCP          66 33890 → 443 [ACK] Seq=1792 Ack=2897 Win=60032 Len=0 TSval=376810612 T
443        TCP          66 33890 → 443 [ACK] Seq=1792 Ack=4097 Win=58880 Len=0 TSval=376810612 T
33890      TCP        1514 443 → 33890 [ACK] Seq=4097 Ack=1792 Win=64128 Len=1448 TSval=95541131
443        TCP          66 33890 → 443 [ACK] Seq=1792 Ack=5545 Win=57472 Len=0 TSval=376810613 T
33890      TLSv1.3     633 Application Data, Application Data, Application Data
443        TCP          66 33890 → 443 [ACK] Seq=1792 Ack=6112 Win=56960 Len=0 TSval=376810614 T
443        TLSv1.3     146 Change Cipher Spec, Application Data
443        TLSv1.3     164 Application Data
443        TLSv1.3     550 Application Data
33890      TLSv1.3     353 Application Data
33890      TLSv1.3     353 Application Data
443        TCP          66 33890 → 443 [ACK] Seq=2454 Ack=6686 Win=56704 Len=0 TSval=376810620 T
33890      TLSv1.3     127 Application Data
33890      TLSv1.3      97 Application Data
443        TCP          66 33890 → 443 [ACK] Seq=2454 Ack=6778 Win=56704 Len=0 TSval=376810620 T
443        TLSv1.3      97 Application Data
```

Fig. 1. Overview of cross-modality alignment anchors in our setting. (A–C) illustrate three hypothesized alignment anchors between website semantic logic (left) and encrypted traffic behavior (right): request-side, response-side, and transport protocol. (D–F) present empirical support for each anchor via Pearson correlation or Wasserstein distance on representative samples.

> There should exist identifiable structural relationships between the modalities, which we refer to as **alignment anchors**. These anchors serve as a learnable bridge that enables the model to connect encrypted traffic behavior with semantic site characteristics.

These principles form the foundation for our modality design choices and motivate the structural observations presented in the following sections.

B. Core Alignment Observations

We define a **cross-modal alignment anchor** as a feature or structure that exhibits semantic correspondence and measurable correlation between the logic and traffic modalities. In our setting, we identify three such alignment anchors—on the request side, response side, and transport protocol—each rooted in the design of modern web communication ecosystems (see Fig. 1 A–C).

1) Request Anchor: Our key observation is that the **length of HTTPS request packets** is linearly related to the **Huffman-encoded length of the resource URI**. This arises from protocol-level optimizations in HTTP/2 and HTTP/3, which employ static/dynamic header compression [26]. Most headers (e.g., User-Agent, Cookie) are replaced with compact indices, leaving the URI as the dominant uncompressed field, further compressed by a public Huffman table. Thus, request packet length can be approximated as:

$$Len(p_i) \approx \mathrm{Len}(\mathrm{Huffman}(uri_i)) + C \times H \tag{1}$$

where p_i is the packet length and H is the number of compressed headers. This alignment is visualized in Fig.1 A and supported in Fig.1 D.

TABLE I
EVALUATION OF ALIGNMENT ANCHORS ACROSS TOP-1000 WEBSITES.

Anchor	**Metric**	**Mean Value** ↑	**p-value** ↓	**Sig. (%)** ↑
Request	Pearson r	0.3114 ± 0.0730	0.0290	86
Response	1− Wasserstein	0.9109 ± 0.0278	0.0124	96
Protocol	Pearson r	0.5607 ± 0.1019	0.0017	100

2) Response Anchor: Response packets convey web content, and their cumulative size naturally reflects the sum of individual resource sizes [27]. We observe that:

$$\sum Len(p_i)^{(\mathrm{resp})} \approx \mathrm{Size}(resource_i) \tag{2}$$

enabling logic-to-traffic comparison of response behavior (Fig. 1B, E).

3) Protocol Anchor: HTTP/3 operates over QUIC/UDP, creating observable transport-layer patterns [9]. We compare the **UDP traffic ratio** with the **server-side HTTP/3 usage ratio**, forming a protocol anchor:

$$\mathrm{UDP\ Ratio} \approx \mathrm{HTTP/3\ Usage\ Ratio} \tag{3}$$

This structural similarity enables indirect inference of protocol usage (Fig. 1C, F).

To quantify these alignments, we perform statistical **hypothesis testing** on paired samples. For each anchor, we extract matched feature sequences from both modalities, apply normalization when necessary, and measure alignment using Pearson correlation or Wasserstein distance. Significance is evaluated via **permutation testing**. Aggregated results in Table I confirm statistically significant alignment across all three anchors, motivating the modality representations used in our framework (§IV).

IV. METHODOLOGY

This section presents the STAR framework for semantic–traffic alignment in website fingerprinting. As shown in Fig. 2, STAR maps paired inputs from two heterogeneous modalities—website logic and encrypted traffic—into a shared embedding space for unified retrieval and classification.

We first define modality-specific input representations (§IV-B), then design a dual-encoder architecture to embed them (§IV-C). The encoders are jointly trained with contrastive and auxiliary losses to promote semantic alignment (§IV-D). To enhance generalization, we introduce structure-aware data augmentation (§IV-E). The learned encoders support flexible downstream usage, including zero-shot retrieval and few-shot adaptation (see Fig. 3).

A. Framework Overview

Our STAR framework is built on a cross-modal dual-encoder design, as illustrated in Fig. 2 and Fig. 3. It takes as input two modalities for each website access: (i) a *logic modality*, which encodes semantic web resource structures—such as resource uri lengths, sizes, and protocol behaviors—and (ii) a *traffic modality*, which captures encrypted packet-level features during access. These paired inputs are processed

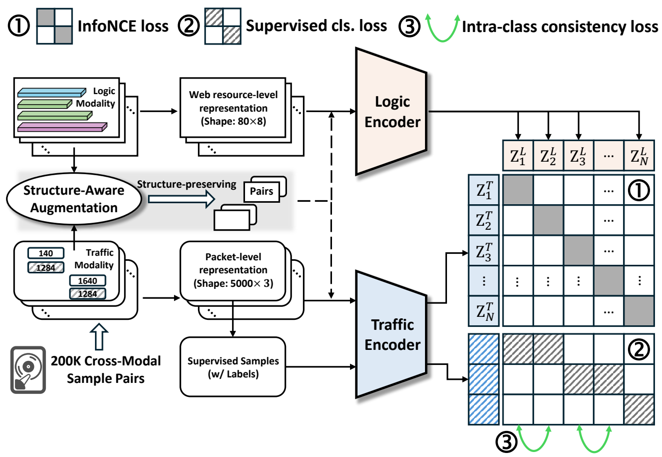

① InfoNCE loss ② Supervised cls. loss ③ Intra-class consistency loss
Logic Modality
Web resource-level representation (Shape: 80×8)
Logic Encoder
Structure-Aware Augmentation
Structure-preserving
Pairs
Traffic Modality
140
1284
1640
1284
Packet-level representation (Shape: 5000× 3)
Traffic Encoder
200K Cross-Modal Sample Pairs
Supervised Samples (w/ Labels)
Z_1^L Z_2^L Z_3^L ... Z_N^L
Z_1^T Z_2^T Z_3^T ⋮ Z_N^T
①
②
③

Fig. 2. **Training-stage framework of STAR.** Structure-aware logic–traffic sample pairs and labeled traffic samples are passed through the Logic Encoder and Traffic Encoder, whose weights are jointly learned so that (1) paired embeddings align via contrastive loss, (2) traffic embeddings support supervised classification, and (3) same-class traffic embeddings remain consistent.

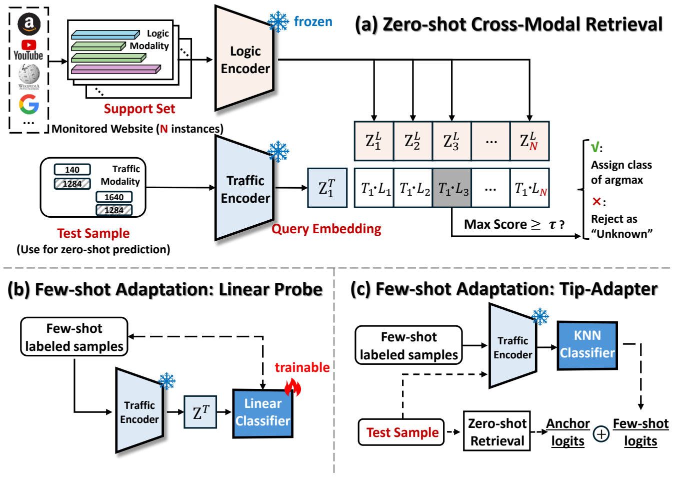


Fig. 3. **Inference-stage framework of STAR.** (a) **Zero-Shot Retrieval**: encode a test trace and match it against gallery logic embeddings; assign the top class if the similarity exceeds a threshold, otherwise reject as "Unknown." (b) **Few-Shot Linear Probe**: train a linear classifier on few-shot traffic embeddings with the encoder frozen. (c) **Few-Shot Tip-Adapter**: fuse anchor-based logits from logic retrieval with k-NN logits from a few-shot traffic memory for final prediction.

by separate encoders and projected into a shared embedding space. To enforce semantic alignment, we apply an InfoNCE-based contrastive loss, while auxiliary classification and consistency losses promote inter-class separability and intra-class coherence.

During training (Fig. 2), STAR is optimized on large-scale cross-modal sample pairs collected via automated crawling and traffic capture, further expanded through structure-aware data augmentation. The learning objectives jointly update both encoders to align paired embeddings, classify traffic samples, and cluster instances from the same class.

At inference time (Fig. 3), the trained encoders support multiple downstream scenarios. For *zero-shot classification*, a test trace is encoded by the traffic encoder and compared—via cosine similarity in the shared embedding space—against a gallery of logic-side prototypes pre-computed from crawl-time profiles. The top-matched class is returned if the similarity exceeds a decision threshold; otherwise, the input is rejected as unmonitored. For *few-shot adaptation*, STAR integrates with plug-and-play strategies such as linear probing [25] or Tip-Adapter-style fusion [28], both operating over frozen encoders [2]. This retrieval-based formulation enables scalable, flexible deployment in open-world scenarios without requiring retraining or per-target traffic collection.

B. Modality Representation Construction

To enable reliable cross-modal alignment, we design compact yet expressive representations for encrypted traffic and site logic. Each is structured as a fixed-length sequence with features selected to reflect the core alignment anchors identified in §III-B, while maintaining generalizability and model efficiency.

1) Traffic Modality: We represent encrypted traffic traces as a sequence of packet-level features, defining a feature matrix $T \in \mathbb{R}^{5000\times 3}$, where each row $f_i^{(T)} = [\mathrm{dir}(p_i),\ v_i,\ s_i]$ corresponds to a packet P_i:

- $\mathrm{dir}(p_i)$ is the *directional packet length*, where client-to-server packets are positive and server-to-client packets are negative. This single value reflects both request and response behaviors, preserving alignment signals from both ends.
- $v_i \in \{1, 2, 3\}$ is the *inferred HTTP version*. Inspired by [9], we heuristically assign this per-packet label based on transport-layer characteristics: UDP packets are labeled HTTP/3; TCP packets are marked HTTP/2 if two consecutive packets begin with TLS content-type `0x17`[3], otherwise HTTP/1.1.
- $s_i \in \mathbb{Z}^+$ is the *flow index*, indicating the bidirectional connection to which the packet belongs, enabling coarse-grained structural grouping within traces.

2) Logic Modality: The logic modality encodes a website's resource-level structure as a semantic matrix $L \in \mathbb{R}^{80\times 8}$, where each row $f_j^{(L)}$ corresponds to a web resource as observed during page load. These resource vectors capture the website's high-level semantics and are extracted from browser developer logs [30] via automated scripts.

We group the eight features into three semantic categories:

- **Identifier length indicators**: Huffman-encoded and raw URI lengths provide a compact representation of the resource path size and align with the request-side packet lengths in the traffic modality.

[2] Detailed implementation-level descriptions of inference procedures are provided in an online technical appendix: https://github.com/2654400439/STAR-Website-Fingerprinting/blob/main/docs/TechnicalAppendix/STAR_Technical_Appendix.pdf.

[3] 0x17 is the TLS `content_type` value indicating *application data* [29]

- **Content indicators**: Response size and header length describe the volume of returned data per resource, enabling alignment with response-side traffic features.
- **Protocol-level context**: HTTP version, alternative service flag (for HTTP/3 support), MIME type category, and server IP index encode protocol semantics and content-type variability across resources, supporting protocol-aware mapping and resource grouping.

3) Normalization & Encoding: All inputs are formatted as fixed-length matrices: traffic traces are truncated or zero-padded to 5000 packets, and logic traces to 80 resources. Continuous features (e.g., lengths and sizes) are log-scaled, categorical fields (e.g., MIME type, stream index) are embedded via learnable vectors, and Boolean values are represented as binary integers.

These design choices ensure that the learned representations remain compact, semantically meaningful, and structurally aligned across modalities.

C. Dual-Encoder Architecture

To bridge the modality gap between encrypted traffic traces and website logic structures, we adopt a dual-encoder architecture to project each modality into a shared embedding space. This architecture is inspired by the CLIP [25] paradigm, where modality-specific encoders are used to preserve intra-modality semantics while enabling cross-modal alignment via **contrastive training**.

a) Traffic Encoder: For the traffic modality, we build upon the DFNet [8] backbone, a deep convolutional network widely adopted in website fingerprinting literature due to its strong discriminative capacity under encrypted traffic. Given a packet-level input matrix $\mathbf{T} \in \mathbb{R}^{5000\times 3}$, we replace the original 1D convolutional layers with three-channel convolutions to accommodate the 3-dimensional packet features. We remove the classification head of DFNet and preserve the penultimate hidden representation as the traffic embedding. A subsequent projection head f_T maps the encoder output to a normalized embedding:

$$\mathbf{z}_i^T = \frac{f_T\left(\text{DFEnc}(\mathbf{T}_i)\right)}{\left|f_T\left(\text{DFEnc}(\mathbf{T}_i)\right)\right|_2} \tag{4}$$

b) Logic Encoder: For the logic modality, we employ a Transformer encoder [31] to effectively process structured sequences of web resources. Given a resource-level input matrix $\mathbf{L} \in \mathbb{R}^{80\times 8}$, the encoder utilizes multi-head self-attention to capture feature-wise and resource-wise dependencies, allowing the model to learn hierarchical importance among resources. The output representations are aggregated via masked average pooling, followed by a projection head f_L that yields the normalized logic embedding:

$$\mathbf{z}_i^L = \frac{f_L\left(\text{TransEnc}(\mathbf{L}_i)\right)}{\left|f_L\left(\text{TransEnc}(\mathbf{L}_i)\right)\right|_2} \tag{5}$$

Each embedding $\mathbf{z}_i^L$ or $\mathbf{z}_i^T$ resides in a shared latent space $\mathbb{R}^d$ (we set $d = 256$), which serves as the basis for cross-modal contrastive alignment.

Algorithm 1 Structure-Aware Cross-Modal Augmentation

```
Require: Logic modality $\mathbf{R} = \{r_1, r_2, \ldots, r_n\}$ with IP tags;
    Traffic modality $\mathbf{P} = \{p_1, p_2, \ldots, p_m\}$ with IP tags
Ensure: Augmented pair $(\hat{\mathbf{R}}, \hat{\mathbf{P}})$
Group $\mathbf{R}$ by server IP: $\mathcal{S} \leftarrow \{s_1, \ldots, s_k\}$ with resource groups $\mathcal{G}(s_i)$
Compute IP selection scores: $\omega(s_i) \leftarrow 1 - \frac{|\mathcal{G}(s_i)|}{|\mathbf{R}|}$
Sample deletion threshold $T \sim \mathcal{N}(\mu = 0.3, \sigma = 0.1) \cdot |\mathbf{R}|$
Initialize $\mathcal{S}_{\text{del}} \leftarrow \emptyset$, $\hat{\mathbf{R}} \leftarrow \mathbf{R}$, $\hat{\mathbf{P}} \leftarrow \mathbf{P}$
while total deleted resources $< T$ do
    Sample IP $s \sim \omega(s)$ with weighted probability
    if $s \notin \mathcal{S}_{\text{del}}$ then
        Remove $\mathcal{G}(s)$ from $\hat{\mathbf{R}}$; remove packets with IP $s$ from $\hat{\mathbf{P}}$
        $\mathcal{S}_{\text{del}} \leftarrow \mathcal{S}_{\text{del}} \cup \{s\}$
return $(\hat{\mathbf{R}}, \hat{\mathbf{P}})$
```

D. Cross-Modal Contrastive Training

To align the logic and traffic modalities, we adopt a multi-objective training strategy centered on InfoNCE loss [32] and supplemented by auxiliary supervision. The goal is to ensure that matched logic-traffic pairs are closer in the embedding space than mismatched pairs.

1) InfoNCE Loss for Cross-Modal Alignment: We leverage the standard contrastive loss over a batch of N paired samples. For each traffic embedding $\mathbf{z}_i^T$ and its corresponding logic embedding $\mathbf{z}_i^L$, the InfoNCE objective encourages the inner product $\langle \mathbf{z}i^T, \mathbf{z}i^L \rangle$ to be higher than that of any non-matching pair. The loss is given by:

$$\mathcal{L}_{\text{InfoNCE}} = -\sum_{i=1}^{N} \log \frac{\exp\left(\langle \mathbf{z}_i^T, \mathbf{z}i^L \rangle / \tau\right)}{\sum_{j=1}^{N} \exp\left(\langle \mathbf{z}_i^T, \mathbf{z}_j^L \rangle / \tau\right)} \tag{6}$$

where τ is a temperature hyperparameter. Unlike conventional supervised contrastive learning, all negatives in the denominator are guaranteed to be true negatives due to the use of large-scale unlabeled pairs across diverse websites (each pair from a distinct site, see §V-A), preventing semantic ambiguity.

2) Supervised Contrastive Loss for Discrimination: To further enhance class-discriminative capacity in the traffic modality, we incorporate a supervised contrastive loss using labeled fingerprinting datasets. Following *SupCon* [33], the loss encourages embeddings from the same class to be closer, while keeping different classes apart:

$$\mathcal{L}_{\text{SupCon}} = -\sum_{i\in\mathcal{I}} \frac{1}{|\mathcal{P}(i)|} \sum_{p\in\mathcal{P}(i)} \log \frac{\exp\left(\langle \mathbf{z}_i^T, \mathbf{z}_p^T \rangle / \tau\right)}{\sum_{a\in\mathcal{A}(i)} \exp\left(\langle \mathbf{z}_i^T, \mathbf{z}_a^T \rangle / \tau\right)} \tag{7}$$

where $\mathcal{P}(i)$ denotes the set of positives (same class), and $\mathcal{A}(i)$ the set of all anchors except i.

3) Consistency Loss for Stability: Given the inherent instability of encrypted traffic, even within the same class, we introduce an intra-class consistency loss to promote local smoothness among traffic embeddings. Specifically, we minimize pairwise distance among all traffic embeddings with the same class label:

$$\mathcal{L}_{\text{Consistency}} = \sum_{(i,j)\in\mathcal{C}} \left|\mathbf{z}_i^T - \mathbf{z}_j^T\right|_2^2 \tag{8}$$

where $\mathcal{C}$ is the set of intra-class traffic pairs.

4) Final Objective.: The full training objective combines all three components with weighting coefficients:

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \lambda \text{sup} \mathcal{L}_{\text{SupCon}} + \lambda \text{cons} \mathcal{L}_{\text{Consistency}} \tag{9}$$

This hybrid objective enables us to exploit both large-scale weakly-aligned web pairs and reliable supervised samples to improve alignment quality and generalization.

E. Structure-Aware Cross-Modal Augmentation

To enhance generalization under site evolution, we introduce a structure-aware augmentation method that perturbs both modalities in a consistent manner. This approach generates realistic logic–traffic sub-pairs while preserving the structural alignment necessary for contrastive training. Unlike traditional view-level augmentations or modality-specific transformations [34], our method exploits a shared structural anchor—**server IP addresses**—that appears in both modalities and governs subsets of web resources and traffic packets.

The augmentation operates by selectively dropping all resources in the logic modality that are associated with a sampled set of server IPs. The corresponding traffic packets linked to the same IPs are then removed from the traffic modality, producing a semantically valid and internally consistent subpair. To avoid excessive content removal, IPs are sampled with inverse probability proportional to their resource count, and deletions continue until a stochastic threshold is met. This threshold is drawn from a Gaussian prior to introduce controlled variation. The full procedure is outlined in Algorithm 1.

The resulting augmented pairs preserve partial yet coherent cross-modal alignment and are seamlessly integrated into the training set as additional samples for contrastive learning.

V. Evaluation

In this section, we conduct comprehensive experiments to evaluate the effectiveness of **STAR** across both closed-world and open-world settings. We first describe the datasets used in our experiments (§V-A) and introduce competitive baselines (§V-B). We then perform a series of experiments, including modality design analysis (§V-C), classification under closed-world (§V-D) and open-world (§V-E) settings, as well as in-depth ablation and interpretability analysis (§V-F).

A. Datasets

We utilize two types of datasets in our experiments: (1) a large-scale cross-modal dataset constructed by ourselves, and (2) an existing labeled fingerprinting dataset used for evaluation and auxiliary supervision.

(1) Cross-Modal Dataset (STAR-200K). We collect a large-scale dataset of website-level cross-modal samples based on the top 200,000 sites from the Tranco list[4] [35]. Data collection is performed on ten geographically distributed AWS EC2 instances across North America, Europe, and Asia. For each site, we use Selenium-controlled `Chrome browsers` (selected for its market share >60% [36]) to access the homepage, extracting browser logs for logical modality representation. Concurrently, raw traffic is captured using `tcpdump` to build the encrypted traffic modality. Each site is accessed once; after filtering for failures (e.g., connection issues, CAPTCHAs), we obtain over 170K valid sample pairs. We refer to this dataset as **STAR-200K**. We use 150K pairs for training STAR, and reserve 20K disjoint pairs for open-world evaluation.

TABLE II
Zero-shot classification results and modality properties with different traffic representations.

Modality Representation		Task Accuracy		Modality Properties		
Logic	**Traffic**	**Top-1 (%)**	**Top-5 (%)**	**AMI (P1)**	**FDR (P2)**	**dCor (P3)**
Ours	CUMUL [13]	36.69	62.48	0.3539	0.3611	0.3811
	Trace [8]	52.44	80.19	0.2389	0.2445	**0.6312**
	H123 [9]	50.50	76.62	**0.6748**	**1.909**	0.4466
	TAM [37]	12.09	42.03	0.4121	1.175	0.2791
	WTCM [38]	18.76	50.18	0.4893	1.2088	0.2329
	Ours	**87.87**	**96.94**	0.6228	1.5744	0.5906

(2) Labeled Fingerprinting Dataset (H&W-1600). We use the public dataset from [9], which provides 40 traffic samples for each of 2,240 HTTPS websites across three groups: `popular`, `random`, and `censorship`. We select the `popular` subset (1,600 websites) for closed-world evaluation. The remaining samples are used as labeled data for supervised training modules (§IV-D). To prevent data leakage, we ensure all evaluation websites are disjoint from the STAR-200K pretraining and labeled training sets.

B. Baselines

To demonstrate the effectiveness of **STAR**, the first *zero-shot* website fingerprinting method without access to target traffic, we compare against representative state-of-the-art baselines from three categories.

Standard WF methods include CUMUL [13], which uses cumulative packet lengths with an SVM classifier; DF+ [8], a CNN-based model extended to directional packet lengths for HTTPS settings; RF [37], which utilizes fixed-time aggregation matrices for deep classification; and CountMamba [38], which models coarse-grained count matrices using a state space model for robust, early-stage classification.

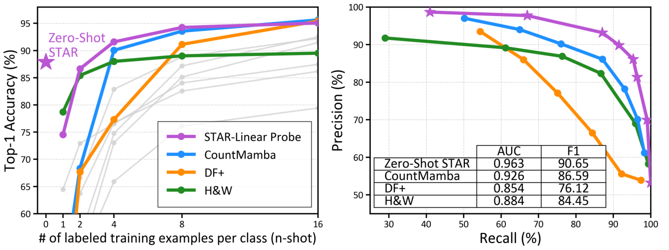
Few-shot methods include TF [21] and NetCLR [22], both of which pretrain DF-based encoders using contrastive learning (NetCLR adds self-supervised tasks), and H&W [9], which matches application-layer features via KNN.

Fine-grained methods include FineWP [27], using statistical features with random forests, and Oscar [39], which applies multi-label metric learning for precise web page classification.

[4] Available at https://tranco-list.eu/list/5XYPN.

TABLE III
CLOSED-WORLD WEBSITE-FINGERPRINTING ACCURACY (%)

Method	0-shot		1-shot		2-shot		4-shot		8-shot		16-shot	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
CUMUL [13]	/	/	64.46	73.61	72.93	79.80	76.52	85.13	84.04	90.77	87.45	92.18
DF [8]	/	/	17.48	37.44	51.76	76.71	73.04	90.72	85.13	96.39	91.46	98.50
DF+	/	/	33.10	57.18	67.67	85.92	77.31	91.75	91.13	98.36	95.41	99.35
RF [37]	/	/	26.63	47.84	51.29	75.49	65.91	86.99	76.10	92.66	79.43	94.82
CountMamba [38]	/	/	47.16	66.68	68.34	85.09	90.04	97.73	93.56	98.94	**95.62**	**99.62**
H&W [9]	/	/	78.70	89.08	85.42	91.96	88.01	93.70	89.02	93.98	89.51	94.12
NetCLR [22]	/	/	36.67	58.74	55.60	77.35	74.77	91.02	87.17	96.62	92.51	98.22
TF [21]	/	/	59.21	59.33	69.87	78.42	76.69	89.29	80.24	92.58	82.03	93.27
FineWP [27]	/	/	35.00	53.72	66.63	82.83	82.88	93.66	88.84	96.33	92.21	97.74
Oscar [39]	/	/	46.94	64.34	63.71	79.08	76.70	88.32	82.55	91.98	86.16	93.95
Clustering + Hungarian	30.04	/	/	/	/	/	/	/	/	/	/	/
STAR-Linear Probe	**87.87**	**96.94**	74.53	92.21	86.63	97.12	91.59	98.78	**94.24**	**99.38**	95.06	99.39
STAR-Tip Adapter			**88.26**	**97.20**	**90.93**	**98.52**	**91.92**	**98.84**	93.42	98.95	94.11	99.09



	AUC	F1
Zero-Shot STAR	0.963	90.65
CountMamba	0.926	86.59
DF+	0.854	76.12
H&W	0.884	84.45

(a) Few-shot Performance in Closed-world Setting (b) Precision-Recall Curves in Open-world Setting

Fig. 4. **Closed-world and open-world performance comparison. (a)** Top-1 accuracy under different n-shot settings in the closed-world scenario. Zero-shot STAR is marked with a purple star, and the top-3 few-shot baselines are color-highlighted; others are shown in grey. **(b)** Precision-recall curves in the open-world 4-shot setting, comparing Zero-shot STAR with the top-3 baselines. AUC and best F1 scores are shown in the table.

All baselines are implemented using official or WFlib [40] code with default settings.

C. Evaluation of Modality Representations

To validate the effectiveness of our cross-modal formulation, we begin with a systematic evaluation of different modality representation choices.

• **Experimental Setup.** We adopt the proposed STAR training paradigm and explore its behavior under various traffic modality representations. The logic modality is fixed to our proposed 8-dimensional web resource-level representation, while the traffic modality varies across prior designs in the website fingerprinting literature. For example, we include Trace sequences [8], flow-level statistical summaries (H123) [9], the Traffic Aggregation Matrix (TAM) [37], and the Windowed Traffic Counting Matrix (WTCM) [38]. For each modality combination, we perform full model training with our multi-loss objective and structure-aware augmentation, then evaluate zero-shot classification performance on the H&W-1600 dataset.

To complement accuracy metrics, we additionally assess three modality design criteria introduced in §III-A:

- **Inter-class discriminability (P1)** is quantified via **Adjusted Mutual Information (AMI)**, which measures how well the traffic embeddings can be clustered into groups that match the true class labels.
- **Intra-class stability (P2)** is estimated by the **Fisher Discriminant Ratio (FDR)**, comparing between-class and within-class variances of traffic embeddings.
- **Cross-modal alignability (P3)** is captured using **Distance Correlation (dCor)** [41] between normalized embeddings $\mathbf{z}_i^T$ and $\mathbf{z}_i^L$. dCor equals zero if and only if two random variables are statistically independent.

All statistics are computed on traffic embeddings normalized to zero mean and unit variance. AMI and dCor are scale-invariant, while FDR is already a scale-free ratio.

• **Results.** As shown in Table II, Trace-based representations achieve the highest dCor, benefiting from explicit request-response anchoring that naturally aligns with logic semantics. H123 obtains the best AMI and FDR scores, thanks to its protocol-aware descriptors and flow-level aggregation that enhance class discriminability and intra-class consistency.

However, these statistical strengths do not directly translate into strong task performance—neither H123 nor Trace reaches competitive zero-shot accuracy. TAM and WTCM, despite their popularity in prior single-modal tasks, show poor alignment and low performance in our cross-modal setup. This is likely due to their use of fixed-size sliding windows, which obscure packet-level semantic anchors and disrupt alignment with logic-side representations.

In contrast, our proposed traffic encoding preserves both protocol semantics and alignment structures, achieving balanced modality properties and significantly superior classification accuracy. These results validate our cross-modal formulation and modality design as crucial to enabling effective zero-shot retrieval.

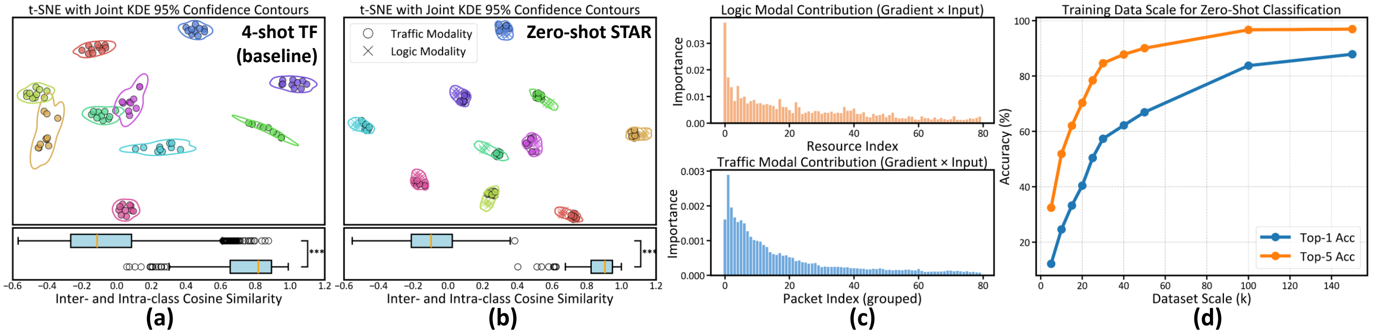


Fig. 5. **Analysis of the STAR model.** (a, b) t-SNE visualization and cosine similarity statistics of modality representations learned by TF (baseline) and STAR, respectively. (c) Gradient-based importance scores over input positions reveal that both modalities exhibit localized discriminative patterns. (d) Impact of training data scale on zero-shot classification accuracy, showing rapid performance saturation after 100k samples.

D. *Closed-World Website Fingerprinting*

To evaluate STAR under a standard closed-world setting, we conduct experiments where the client accesses a fixed set of monitored websites. Results are summarized in Table III.

- **Experimental Setup.** We train STAR on a mixture of three datasets: the STAR-200K cross-modal dataset, structure-aware augmented pairs and the labeled training portion of H&W, combined at a 10:3:3 ratio. To assess zero-shot performance—i.e., recognizing websites unseen during training in the traffic modality—we construct disjoint training and evaluation website sets. The model is optimized with the objective in Eq. 9 for 200 epochs on 5 NVIDIA A100 GPUs, requiring about 4 hours. During inference, we follow a CLIP-style retrieval procedure: each traffic sample is embedded using the traffic encoder and projection head, and cosine similarity is computed against 1,600 logic-side anchors, each representing an embedding of a logic modality sample from the test set. Classification is determined by nearest-neighbor retrieval, and we report both top-1 and top-5 accuracy.

Since no existing website fingerprinting approach supports zero-shot classification, we build a baseline using k-means clustering ($K = 1,600$) with optimal label assignment obtained via the Hungarian algorithm. For few-shot evaluation, we follow the standard n-shot setting on H&W-1600, using n labeled samples per class. Competing methods (e.g., TF, NetCLR, H&W) are trained on the n-shot subset, while STAR uses lightweight adaptation via a linear probe and Tip-Adapter.

- **Results.** STAR delivers strong *zero-shot* performance, achieving **87.87% top-1** and **96.94% top-5 accuracy** over 1,600 website classes, despite not seeing any traffic samples from the evaluation set. This confirms the effectiveness and generalization ability of the learned cross-modal alignment. With few-shot adaptation, STAR's performance improves further, reaching 95.06% top-1 accuracy with a linear probe at 16-shot and up to 99.09% with Tip-Adapter. Compared with existing few-shot methods, STAR provides both higher upper-bound accuracy and better zero-shot generalization. For example, H&W and TF remain competitive under few-shot settings but still plateau below STAR's adapted results. Notably, As shown in Fig. 4(a), STAR's zero-shot accuracy already matches the average 8-shot performance of other methods, which typically require over 100 hours of traffic collection on a single machine [21], highlighting its advantage in low-data and real-time deployment scenarios.

E. *Open-World Website Fingerprinting*

Website fingerprinting in the open-world setting offers a more realistic evaluation paradigm, as it assumes the client may access websites outside the attacker's monitored set. In this scenario, effective attack methods must possess the capability to reject inputs from unknown or unmonitored websites—an open-set recognition challenge.

- **Experimental Setup.** To evaluate performance under open-world conditions, we select the top-performing baselines from the closed-world experiments: CountMamba, DF+, and H&W, representing state-of-the-art methods across different paradigms. Each is trained under a 4-shot setting (i.e., 4 labeled examples per monitored class), and compared against our *zero-shot* **STAR** model.

These four methods span two representative open-world handling strategies: **(i)** Threshold-based similarity rejection: STAR and H&W directly compute similarity scores between test samples and support samples, rejecting a test input if its top similarity falls below a predefined threshold. **(ii)** Explicit background class training: CountMamba and DF+ introduce a "non-monitored" class during training by sampling from a large unmonitored website set, enabling the model to learn a decision boundary between monitored and unmonitored traffic.

Following established practice [9], we adopt a binary classification evaluation strategy between monitored and unmonitored samples. At test time, we construct a balanced evaluation set, ensuring a 1:1 ratio of monitored and unmonitored samples. We report precision and recall at varying decision thresholds, along with the overall AUC and the best F1 score for each method.

- **Results.** Experimental results are summarized in Fig. 4(b). *zero-shot* STAR achieves the best open-world detection performance among all evaluated methods, attaining an **AUC of 0.963**, significantly outperforming CountMamba (0.926), DF+ (0.854), and H&W (0.884). The corresponding best F1 score

TABLE IV
ABLATION STUDY RESULTS.

Method	Closed-World		Open-World	
	Top-1	**Top-5**	**AUC**	**F1**
Base[1]	69.56	91.06	0.850	82.63
Base + CMA[2]	80.31	92.91	0.897	85.91
Base + CMA + OT_{Cls}[3]	82.19	94.75	0.916	87.03
Base + CMA + OT_{Cons}	84.06	95.87	0.929	87.12
Base + CMA + $\text{OT}_{\text{Hybrid}}$	**87.87**	**96.94**	**0.963**	**90.65**

[1] Basic cross-modal alignment trained with large-scale sample pairs using the InfoNCE loss.
[2] **CMA**: Cross-Modal Augmentation.
[3] **OT**: Optimization Targets (see §IV-D).

of STAR is 90.65, indicating both high precision and strong recall.

We attribute this performance gain to the **cross-modal alignment learning paradigm** employed by STAR. Unlike traditional classification-based approaches—which focus on optimizing decision boundaries over a fixed set of monitored classes—our model is trained on large-scale cross-modal sample pairs, allowing it to learn a more **generalizable alignment space** between website-level semantic features and encrypted traffic patterns. This alignment is not bound to specific class labels, but rather captures discriminative structure across the broader web domain. As a result, even in open-world settings where unseen websites appear, STAR can reliably identify whether a test sample aligns well with any monitored site—without requiring explicit negative class supervision during training. These findings highlight the unique advantage of retrieval-based, modality-aligned approaches in realistic, open-set fingerprinting scenarios.

F. Ablation and Interpretability Analysis

To better understand the design, performance gain, and behavior of STAR, we conduct an in-depth analysis covering its key components, learned representations, and the effect of training scale.

- **Ablation Study.** Table IV reports ablation results under closed- and open-world settings. Starting from the base cross-modal alignment model trained with InfoNCE loss, incorporating **Cross-Modal Augmentation (CMA)** consistently improves performance. Additional gains are achieved by introducing **optimization targets** for classification (OT_{Cls}), consistency (OT_{Cons}), and their hybrid form, with $\text{OT}_{\text{Hybrid}}$ yielding the best overall results.
- **Representation Analysis.** We visualize learned embeddings using t-SNE in **Fig. 5(a-b)**. Compared with the TF baseline, STAR produces tighter intra-class clusters and stronger alignment between traffic and logic embeddings. Cosine similarity distributions further confirm higher intra-class similarity and improved separability.
- **Attribution and Scale Analysis.** *Gradient*×*Input* attribution [42] (**Fig. 5(c)**) reveals that STAR exploits localized discriminative cues in both modalities. In the logic modality, influential features are concentrated in early resource slots, often corresponding to primary page elements, while in the traffic modality, early packet groups contribute disproportionately to alignment. **Fig. 5(d)** further shows that zero-shot accuracy improves rapidly with training scale and saturates beyond approximately 100k samples, suggesting diminishing returns once sufficient cross-modal diversity is learned.

Overall, this analysis demonstrates that STAR's superior performance arises not from any single component, but from the joint effect of robust cross-modal pretraining, effective alignment optimization, and scale-driven generalization.

VI. DISCUSSION

A. Scope and Limitations

This work redefines website fingerprinting as a cross-modal retrieval problem and presents STAR as a first realization of this paradigm. Our evaluation is scoped to standard HTTPS browsing sessions and validates the approach under typical conditions. More complex settings—such as multi-tab access, cross-network variability, and alternative encryption tunnels like VPN or Tor—remain beyond the scope of this initial study. We also focus on Chrome-based traffic traces, given its prevalence in practice; generalization to other browsers like Firefox, Safari remains to be assessed. These scenarios introduce additional factors that may affect alignment robustness and are left for future exploration.

B. Implications and Future Directions

STAR demonstrates that semantic–traffic alignment enables scalable, zero-shot fingerprinting without target-side traffic, revealing structural leakage as a persistent privacy risk even under full encryption. Beyond facilitating low-overhead deployment, our formulation offers a lens to analyze semantic leakage and guide defense design. Potential countermeasures include perturbing resource structures or obfuscating alignment anchors via traffic shaping, though their effectiveness and associated bandwidth overhead remain open challenges. Future work may extend STAR to multi-page tracking, dynamic contexts, or hybrid inference tasks involving both structure and behavior.

VII. CONCLUSION

We reframed website fingerprinting under HTTPS as a zero-shot cross-modal retrieval problem and introduced STAR, a dual-encoder system that aligns semantic resource logic with encrypted traffic. Trained on large-scale logic–traffic pairs with structure-aware augmentation, STAR achieves strong zero-shot classification without target-side traffic collection. It surpasses state-of-the-art baselines across closed- and open-world settings, highlighting semantic–traffic alignment as a new axis of vulnerability. We release our dataset and implementation to support future research in both attack and defense directions.

ACKNOWLEDGMENT

We sincerely thank the anonymous reviewers for their valuable comments and suggestions, which helped improve the quality of this paper.

References

[1] E. Rescorla, K. Oku, N. Sullivan, and C. A. Wood, "TLS Encrypted Client Hello," Internet Engineering Task Force, Internet-Draft draft-ietf-tls-esni-25, Jun. 2025, work in Progress. [Online]. Available: https://datatracker.ietf.org/doc/draft-ietf-tls-esni/25/

[2] P. E. Hoffman and P. McManus, "DNS Queries over HTTPS (DoH)," RFC 8484, Oct. 2018. [Online]. Available: https://www.rfc-editor.org/info/rfc8484

[3] N. Niere, F. Lange, N. Heitmann, and J. Somorovsky, "Encrypted client hello (ech) in censorship circumvention," *Free and Open Communications on the Internet*, 2025.

[4] M. Shen, K. Ye, X. Liu, L. Zhu, J. Kang, S. Yu, Q. Li, and K. Xu, "Machine learning-powered encrypted network traffic analysis: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 791–824, 2022.

[5] D. Wagner, B. Schneier *et al.*, "Analysis of the ssl 3.0 protocol," in *The second USENIX workshop on electronic commerce proceedings*, vol. 1, no. 1, 1996, pp. 29–40.

[6] A. Hintz, "Fingerprinting websites using traffic analysis," in *International workshop on privacy enhancing technologies*. Springer, 2002, pp. 171–178.

[7] J. Hayes and G. Danezis, "k-fingerprinting: A robust scalable website fingerprinting technique," in *25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 1187–1203.

[8] P. Sirinam, M. Imani, M. Juarez, and M. Wright, "Deep fingerprinting: Undermining website fingerprinting defenses with deep learning," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 1928–1943.

[9] Y. Cheng, Y. Zhu, B. Li, P. Sun, Y. Ding, X. Deng, and Q. Liu, "Holmes & watson: A robust and lightweight https website fingerprinting through http version parallelism," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 1078–1092.

[10] X. Deng, Q. Yin, Z. Liu, X. Zhao, Q. Li, M. Xu, K. Xu, and J. Wu, "Robust multi-tab website fingerprinting attacks in the wild," in *2023 IEEE symposium on security and privacy (SP)*. IEEE, 2023, pp. 1005–1022.

[11] M. Juarez, S. Afroz, G. Acar, C. Diaz, and R. Greenstadt, "A critical evaluation of website fingerprinting attacks," in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, 2014, pp. 263–274.

[12] Y. Cheng, "STAR website fingerprinting code repository," https://github.com/2654400439/STAR-Website-Fingerprinting, 2025, accessed: 2025-12-17.

[13] A. Panchenko, F. Lanze, J. Pennekamp, T. Engel, A. Zinnen, M. Henze, and K. Wehrle, "Website fingerprinting at internet scale." in *NDSS*, vol. 1, 2016, p. 23477.

[14] B. Gao, W. Liu, G. Liu, F. Nie, and J. Huang, "Multi-level resource-coherented graph learning for website fingerprinting attacks," *IEEE Transactions on Information Forensics and Security*, 2024.

[15] J. Li, D. Wang, Y. Liu, Y. Gao, X. Zhang, Z. Lin, X. Ma, X. Luo, and X. Guan, "Cross-environmental website fingerprinting," in *IEEE INFOCOM 2025-IEEE Conference on Computer Communications*. IEEE, 2025, pp. 1–10.

[16] C. Yang, X. Xiao, G. Hu, Z. Ling, H. Li, and B. Zhang, "Dtpn: A diffusion-based traffic purification network for tor website fingerprinting," in *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, 2025, pp. 642–650.

[17] B. Cebere and C. Rossow, "Understanding web fingerprinting with a protocol-centric approach," in *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*, 2024, pp. 17–34.

[18] G. Huang, C. Ma, M. Ding, Y. Qian, C. Ge, L. Fang, and Z. Liu, "Efficient and low overhead website fingerprinting attacks and defenses based on tcp/ip traffic," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 1991–1999.

[19] S. Siby, L. Barman, C. Wood, M. Fayed, N. Sullivan, and C. Troncoso, "Evaluating practical quic website fingerprinting defenses for the masses," *Proceedings on Privacy Enhancing Technologies*, 2023.

[20] S. Siby, M. Juarez, C. Diaz, N. Vallina-Rodriguez, and C. Troncoso, "Encrypted dns → privacy? a traffic analysis perspective," *arXiv preprint arXiv:1906.09682*, 2019.

[21] P. Sirinam, N. Mathews, M. S. Rahman, and M. Wright, "Triplet fingerprinting: More practical and portable website fingerprinting with n-shot learning," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 1131–1148.

[22] A. Bahramali, A. Bozorgi, and A. Houmansadr, "Realistic website fingerprinting by augmenting network traces," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 1035–1049.

[23] S. E. Oh, N. Mathews, M. S. Rahman, M. Wright, and N. Hopper, "Gandalf: Gan for data-limited fingerprinting," *Proceedings on privacy enhancing technologies*, vol. 2021, no. 2, 2021.

[24] D. Li, C. Gu, and Y. Zhu, "Gene fingerprinting: Cracking encrypted tunnel with zero-shot learning," *IEICE TRANSACTIONS on Information and Systems*, vol. 105, no. 6, pp. 1172–1184, 2022.

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

[26] R. Peon and H. Ruellan, "HPACK: Header Compression for HTTP/2," RFC 7541, May 2015. [Online]. Available: https://www.rfc-editor.org/info/rfc7541

[27] M. Shen, Y. Liu, L. Zhu, X. Du, and J. Hu, "Fine-grained webpage fingerprinting using only packet length information of encrypted traffic," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2046–2059, 2020.

[28] R. Zhang, R. Fang, W. Zhang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," *arXiv preprint arXiv:2111.03930*, 2021.

[29] E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.3," RFC 8446, Aug. 2018. [Online]. Available: https://www.rfc-editor.org/info/rfc8446

[30] Google Chrome Developers, "Performance log - chromedriver," https://developer.chrome.com/docs/chromedriver/logging/performance-log?hl=en, 2024, accessed: 2025-07-29.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[33] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[34] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," *arXiv preprint arXiv:2110.05208*, 2021.

[35] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation," in *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, ser. NDSS 2019, Feb. 2019.

[36] Cloudflare, Inc., "Browser usage statistics – cloudflare radar," https://radar.cloudflare.com/adoption-and-usage, 2025, accessed: 2025-07-19.

[37] M. Shen, K. Ji, Z. Gao, Q. Li, L. Zhu, and K. Xu, "Subverting website fingerprinting defenses with robust traffic representation," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 607–624.

[38] X. Deng, R. Zhao, Y. Wang, M. Zhan, Z. Xue, and Y. Wang, "Countmamba: A generalized website fingerprinting attack via coarse-grained representation and fine-grained prediction," in *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2025, pp. 1419–1437.

[39] X. Zhao, X. Deng, Q. Li, Y. Liu, Z. Liu, K. Sun, and K. Xu, "Towards fine-grained webpage fingerprinting at scale," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 423–436.

[40] X. Deng, "Website-fingerprinting-library," https://github.com/Xinhao-Deng/Website-Fingerprinting-Library, 2023, accessed: 2025-07-19.

[41] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, Dec. 2007.

[42] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.