

Appendix A Effect of Feature Selection

By removing each of the three self-supervised tasks (PFT, HPT, DMT) one by one, this experiment verifies the contribution of each task to the complete model and demonstrates the necessity of the feature selection.

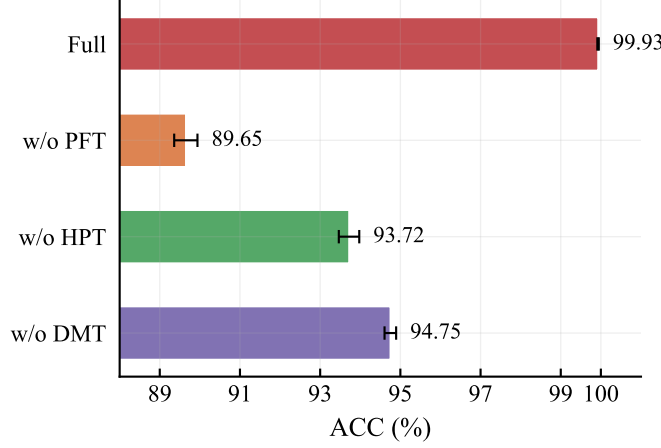


Fig. A1 Ablation experiment of removing self-supervised tasks.

As shown in Fig. A1, removing any self-supervised task leads to significant performance degradation, proving that all three tasks make irreplaceable contributions to the model. Among them, removing PFT causes the most severe *ACC* drop from 99.93% to 89.65%, a decrease of 10.28 percentage points. This is because PFT learns payload representations directly at the byte level and serves as the core task of the entire pre-training phase. Its absence deprives the model of basic ability to understand content features of encrypted traffic. Removing HPT (*ACC* drops to 93.72%, a decrease of 6.21 percentage points) also leads to notable performance decline, demonstrating that the homology prediction task helps the model capture semantic correlations of TCP/UDP header fields and inter-packet dependencies. Removing DMT (*ACC* drops to 94.75%, a decrease of 5.18 percentage points) indicates that the domain matching task is critical for aligning payload features with cross-domain traffic distributions. The three self-supervised tasks form a complementary representation learning system, and none is dispensable.

Appendix B Effect of Topology Modality Weights

Ablation experiment of topology modality weights systematically evaluates the impact of topological modality weight values ranging from 0.10 to 0.40 on detection accuracy to determine the optimal weight interval. Fig. B2 shows the experimental results.

When the topological modality weight is 0.10, the *ACC* is only 97.76%, indicating that an excessively low topological weight cannot fully leverage cross-cloud topological

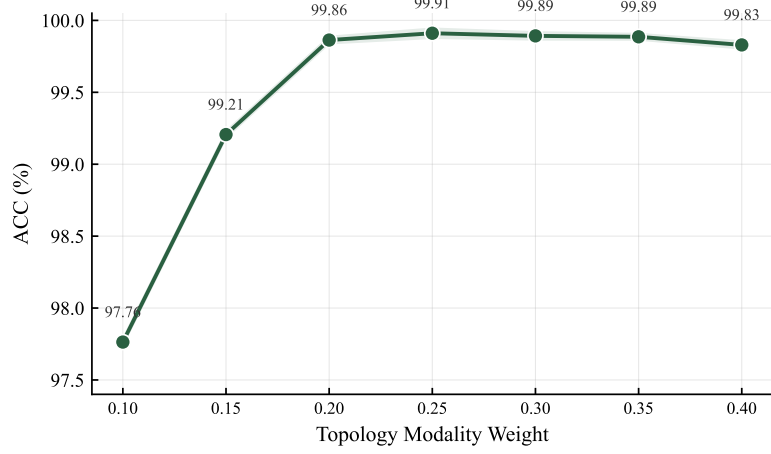


Fig. B2 Ablation experiment of topology modality weights.

information, making it difficult for the model to distinguish topological differences between malicious and normal traffic along service chain paths. As the weight increases from 0.10 to 0.25, the *ACC* rapidly rises to a peak of 99.91%, indicating that the fusion of topological information with payload and packet header features reaches a relatively optimal balance. When the weight further increases to 0.30–0.40, the *ACC* slightly decreases to 99.83%–99.89%. This is because an excessively high topological weight compresses the learning space of other modalities, limiting the model’s ability to comprehensively utilize multi-dimensional information. When the weight is below 0.15, the *ACC* drops sharply by more than 0.7 percentage points, demonstrating that topological information cannot be ignored. The model maintains near-optimal performance in the 0.20–0.30 interval, providing a clear theoretical basis for parameter selection in practical deployment.

Appendix C Effect of Weight Distribution

To intuitively demonstrate the adaptive superiority of the AWAf dynamic attention fusion mechanism, the attention weight distributions of the three modalities are visualized under five representative topological load conditions, as shown in Fig. C3. The five scenarios are defined according to the real-time transmission state vector $S=(s_1, s_2, s_3)$, where s_1 , s_2 , and s_3 denote the normalized node hop count, transmission delay, and packet loss rate, respectively. All values have been normalized for consistent comparison. For each scenario, 100 test samples are drawn, and the per-flow attention weights ($\omega_p, \omega_h, \omega_t$) output by the AWAf module are collected and averaged.

- **S1** (low-load short path): $s_1 \in [0.1, 0.3]$, $s_2 \in [0.05, 0.25]$, $s_3 \in [0.01, 0.05]$;
- **S2** (medium-load standard path): $s_1 \in [0.4, 0.8]$, $s_2 \in [0.30, 0.55]$, $s_3 \in [0.10, 0.25]$;
- **S3** (high-load long path): $s_1 \in [0.8, 1.0]$, $s_2 \in [0.60, 0.95]$, $s_3 \in [0.30, 0.70]$;
- **S4** (high packet-loss anomaly path): $s_1 \in [0.4, 0.8]$, $s_2 \in [0.30, 0.55]$, $s_3 \in [0.80, 0.95]$;
- **S5** (high-latency congestion path): $s_1 \in [0.4, 0.8]$, $s_2 \in [0.80, 0.95]$, $s_3 \in [0.30, 0.55]$.

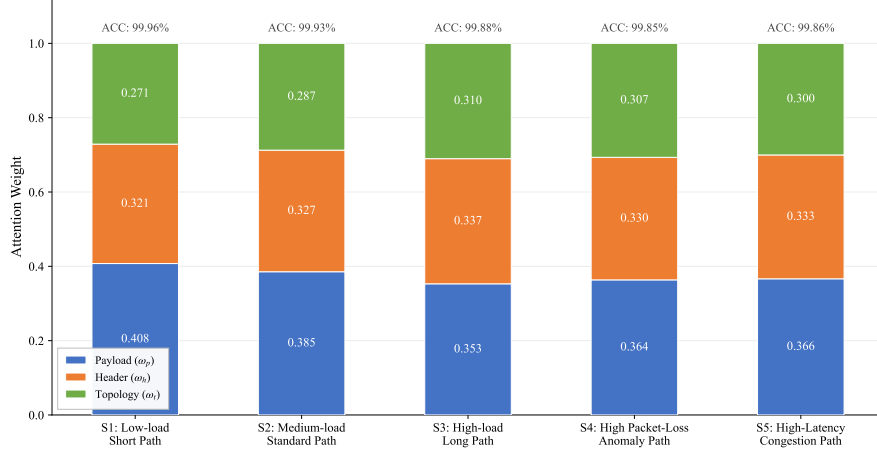


Fig. C3 Ablation experiment of weight distribution.

In the low-load short-path scenario (S1), the attention weight gap between Payload and Topology reaches its maximum ($\omega_p - \omega_t = 0.137$), reflecting that under favorable network conditions the content features alone are highly sufficient for accurate classification, and the supplementary contribution of topological features is relatively limited. As the transmission path becomes longer and more congested, this gap narrows substantially. From S1 to S3, the gap shrinks from 0.137 to 0.043, a reduction of 68.6%, accompanied by a concurrent increase in the Topology weight from 0.271 to 0.310 (a rise of 14.4%). Similarly, in the high packet-loss anomaly scenario (S4) and the high-latency congestion scenario (S5), the gap decreases to 0.057 and 0.066 (reductions of 58.4% and 51.8%), respectively. The AWAf mechanism captures this shift in feature reliability by dynamically increasing the Topology weight ω_t while maintaining the Payload weight ω_p as the leading contributor, thereby achieving a context-aware balance between content-based and topology-based discrimination.

Appendix D Effect of Task Loss Weights

As described in Section A, the contribution ranking of $PFT > HPT > DMT$ further validates the rationality of assigning the highest loss weight to PFT. As shown in Fig. D4, detection accuracy varies noticeably with different loss weight configuration schemes. Ranked in descending order of ACC, the 0.4/0.3/0.3 scheme attains peak performance at 99.93%, succeeded by the 0.5/0.25/0.25 scheme (99.91%), the uniformly weighted 0.33/0.33/0.33 scheme (99.90%), the 0.3/0.4/0.3 scheme (99.88%), and the 0.3/0.3/0.4 scheme at 99.83%. As the most fundamental self-supervised task, PFT directly learns byte-level representations of encrypted payloads. A higher weight for PFT yields better model performance, which aligns with the essential requirement of malicious traffic detection: payload byte features carry critical discriminative information such as encryption protocol handshakes, TLS extensions, and certificate fields. The equal-weight configuration (0.33/0.33/0.33) is close to optimal but does not reach the peak, proving the necessity of a differentiated weighting strategy.

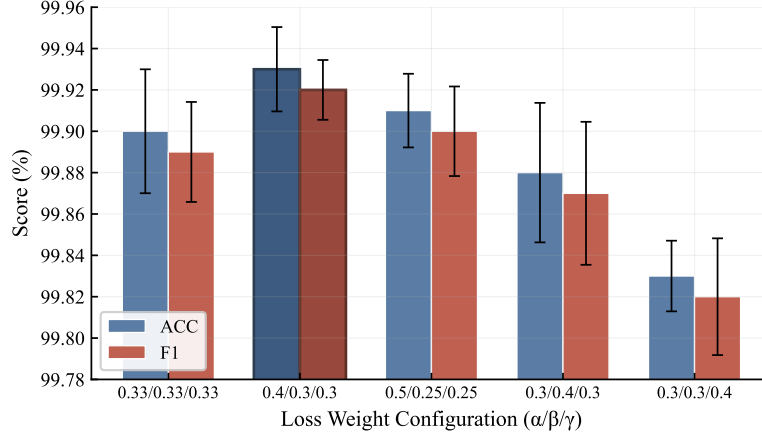


Fig. D4 Ablation experiment of self-supervised task loss weight configuration.

The optimal weights are $\alpha=0.4$, $\beta=0.3$, and $\gamma=0.3$. The PFT task contributes the most to the model. This configuration balances in-depth mining of payload byte features with auxiliary representation learning from homology prediction and domain matching.

Appendix E Effect of Packet Truncation Parameters

Ablation experiment of the number of packets per flow evaluates the impact of sampling 5 to 50 packets per flow on detection accuracy and time to determine the optimal sampling strategy. As shown in Fig. E5, the number of sampled packets per flow directly affects the amount of information available to the model and detection latency. When only 5 packets are sampled, the *ACC* is 80.63%. This is because a small number of packets cannot cover the complete encrypted session handshake process, and the model may miss critical protocol negotiation features. When the number increases to 20 packets, the *ACC* rises to 99.92%, as the model can capture the complete TLS handshake sequence and traffic patterns in the initial data transmission phase. Further increasing to 30–50 packets results in almost no change in *ACC* (99.93%–99.94%), but detection time increases linearly from 2.28 ms to 4.82 ms, indicating high information redundancy in the additional packets.

The optimal sampling number is 20 packets per flow, under which the model obtains sufficient discriminative information. Beyond 20 packets, the marginal contribution of additional information approaches zero, while detection latency increases significantly. This conclusion provides quantitative support for flow sampling strategies in real-time traffic detection systems. The payload truncation length determines the byte range of each packet involved in feature extraction, directly affecting the input dimension and information volume of the model. With the number of sampled packets per flow fixed at 20, the following experiment investigates the effects of different payload truncation lengths on detection performance.

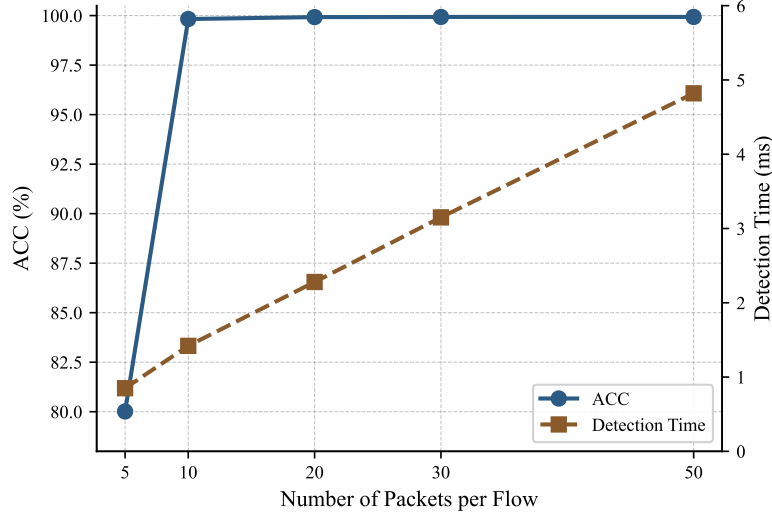


Fig. E5 Ablation experiment of the number of packets per flow.

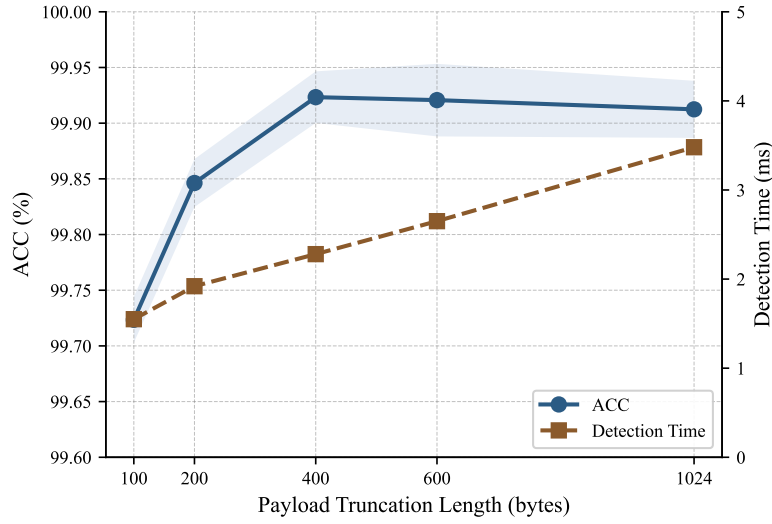


Fig. E6 Ablation experiment of payload truncation length.

As shown in Fig. E6, a 100-byte truncation only achieves an *ACC* of 99.72%, as this length only covers the IP header and part of the TLS Client Hello, and cannot capture messages rich in fingerprint information such as Server Hello and Certificate. A 200-byte truncation improves *ACC* to 99.85%, and a 400-byte truncation reaches the peak of 99.92%. This is because key discriminative features of encrypted traffic—TLS extension lists, SNI fields, cipher suite negotiation, certificate length, etc.—are mainly concentrated in the first 400 bytes of the payload. The *ACC* no longer improves

and even slightly decreases (99.91%) with 600-byte and 1024-byte truncation, indicating that deeper payload content brings no additional contribution to classification but introduces processing noise instead. The first 400 bytes exactly cover complete handshake-phase information.

Appendix F Tuning of Model Hyperparameters

The number of GAT attention heads has a decisive influence on the quality of topological modality feature extraction.

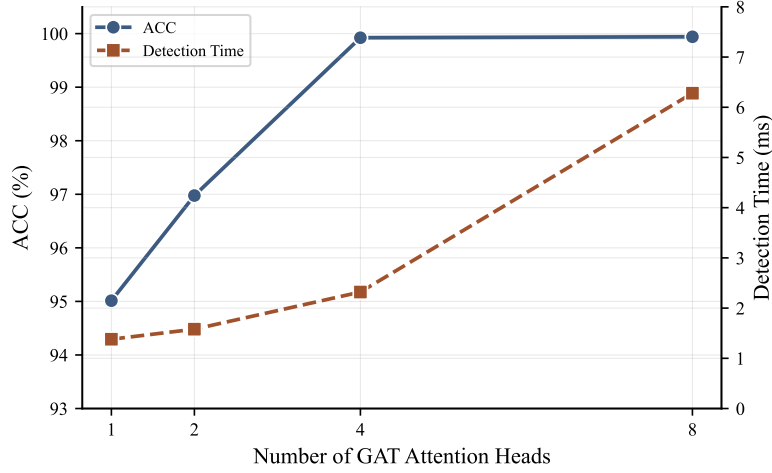


Fig. F7 Ablation experiment of GAT attention heads.

As shown in Fig. F7, single-head attention only achieves an *ACC* of 95.01%, indicating that single-channel attention cannot simultaneously focus on multi-dimensional features of service chain topology. The 2-head configuration improves *ACC* to 96.98%, but is still insufficient to capture all structural information of the topology graph. The 4-head configuration reaches 99.92%, as each attention channel can focus on different types of topological attributes—for example, one channel focuses on latency-sensitive paths, while another pays attention to links with abnormal packet loss rates. The 8-head configuration slightly improves accuracy to 99.94% (only +0.02%), but detection time soars from 2.32 ms to 6.28 ms (a 2.7-fold increase), indicating that the complexity of service chain topology does not require excessive parallel attention channels. The 4-head configuration achieves the best balance between accuracy and efficiency.

The dimension of AWAf has an extremely significant impact on model accuracy. As shown in Fig. F8, the optimal dimension is 128, under which the model achieves the highest accuracy with a detection time of 2.32 ms. When the dimension is 16, the *ACC* drops to 85.01%. This is because the low-dimensional space severely limits the representation capacity of adaptive attention weight fusion, preventing the model

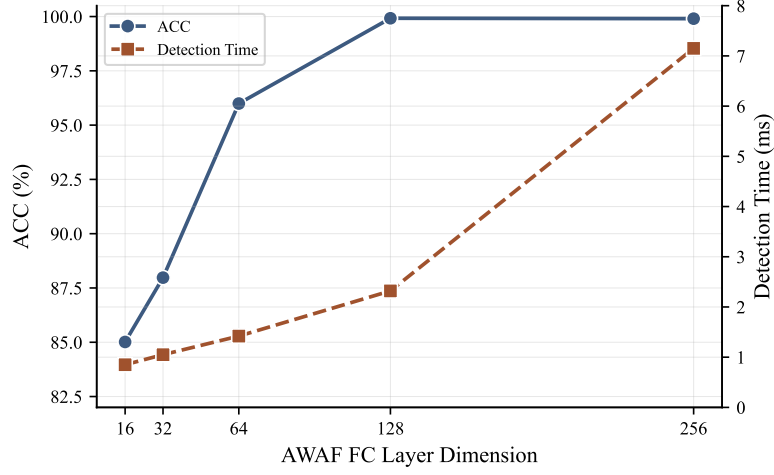


Fig. F8 Ablation experiment of AWAF Fully-Connected layer dimension.

from effectively learning complex interactive relationships among the three modalities. As the dimension increases to 128, the *ACC* sharply rises to 99.92%, indicating that a 128-dimensional representation space fully satisfies cross-modal fusion requirements. However, when the dimension further increases from 128 to 256, no accuracy improvement is observed (99.90%), while detection time surges from 2.32 ms to 7.15 ms (a 3.1-fold increase). The *dim*=256 configuration is a typical over-parameterized case. The additional 128 dimensions bring no accuracy gain but introduce significant computational overhead, which is unfavorable for real-time deployment scenarios.