

TGM-Zero: Text-Guided Zero-Day Detection and Fine-Grained Classification of DoS/DDoS Attacks

Tran L. T. Le¹, David Yau¹, Qun Song²
¹ Singapore University of Technology and Design, Singapore
² City University of Hong Kong

ABSTRACT & MOTIVATION

Current IDS systems face a closed-set limitation, where they fail to identify attack variants not present in training data. TGM-Zero bridges this gap by aligning network telemetry with textual semantics from LLMs, allowing for zero-shot reasoning over "unknown" threats.

THE FRAMEWORK GOAL

Enable robust zero-shot identification and rapid few-shot adaptation for evolving IoT-based volumetric flood attacks.

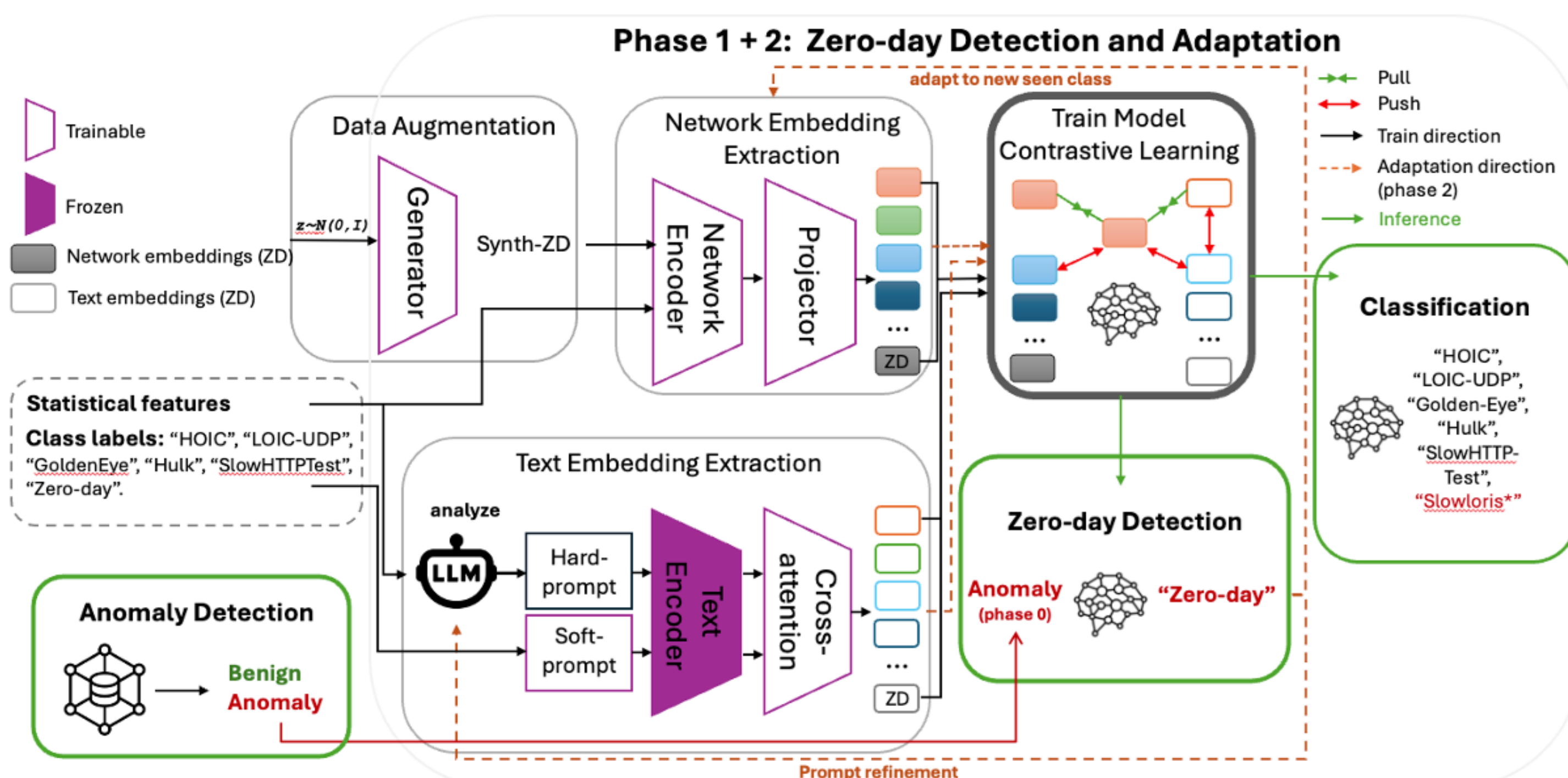


Fig. 1 : Framework Architecture. Diagram showing the flow from raw traffic to LLM-guided text embeddings and final classification.

PHASE 0: ANOMALY DETECTION

A binary filter distinguishes Benign from Anomaly behavior, serving as the first gate in the pipeline.

	Predicted: Attack	Predicted: Benign
Actual: Attack	3631	2
Actual: Benign	47	472

GOAL

Maximize recall to ensure all potential malicious flows are captured while reducing the processing burden on classification stages.

PHASE 1: ZERO-DAY DETECTION

1. Multimodal Alignment

TRAFFIC ENCODER
EfficientNet-B2

TEXT ENCODER
CyBERT/ TinyLLaMA

Network flows are mapped into a latent space shared with textual attack descriptions. Similarity scores are used to classify attacks.

GOAL

Map heterogeneous network telemetry to human-readable domain knowledge for semantic-based classification.

2. Hybrid Prompting

Hard Prompts: Traffic-statistics templates.

1) **Negative Prompts:** Attributes set to "unknown" for out-of-class anchoring.

2) **Descriptive Prompts:** LLM-generated summaries conditioned on real traffic statistics.

Soft Prompts: Learnable task-specific embeddings.

GOAL

Inject expert domain knowledge into the model while allowing it to optimize alignment for specific network contexts.

Prompt Pipeline	Negative (ours)	Descriptive	
GPT-4o → CyBERT	91.70	71.02	Phase 1: Negative prompts improve zero-day accuracy (up to +20%).
GPT-4o → TinyLLaMA	94.52	93.64	
Gemini 2.5 Pro → CyBERT	73.32	44.17	
Gemini 2.5 Pro → TinyLLaMA	95.10	94.88	

3. Synthetic Zero-day Augmentation

Using a Conditional Wasserstein GAN (CLSWGAN), we generate synthetic embeddings for unseen classes using "Negative Prompts" to define what an attack is *not*.

GOAL

Mitigate bias toward seen classes by populating the representation space for zero-day threats before real samples are available.

Ablation Study: Impact of Combining Soft Prompts (SP) and Synth-ZD on Classification Accuracy

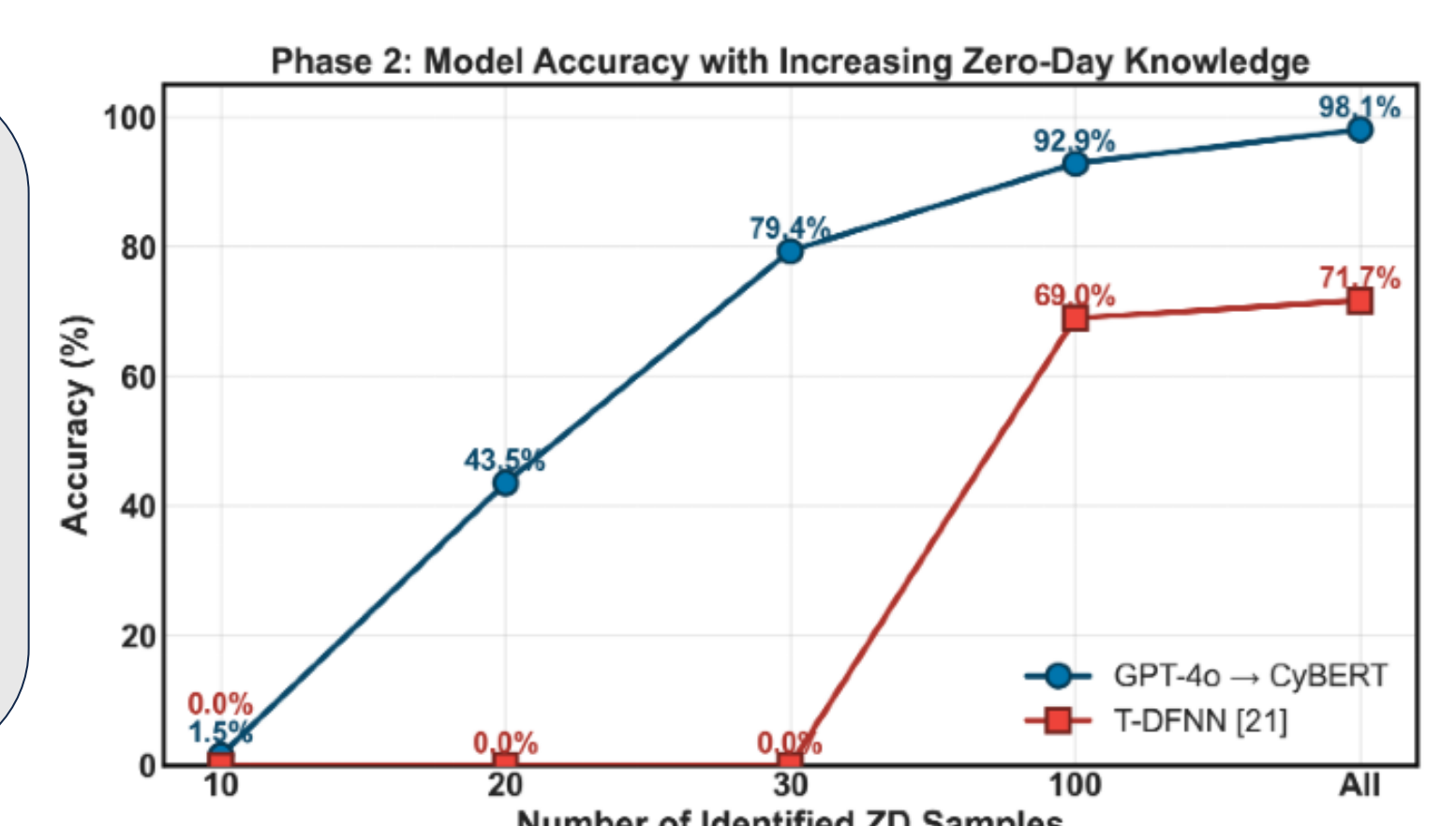
Prompt Pipeline	S.P	Synth ZD	Slowloris*	Overall ACC	HOIC	LOIC UDP	LOIC HTTP	GoldenEye	Hulk	Slow HTTPTest
GPT-4o→CyBERT	✓	✓	98.07	97.66	100.0	99.81	99.23	97.30	96.34	100.0
	✗	✓	69.17	94.07	100.0	100.0	99.61	98.46	99.81	100.0
GPT-4o→TinyLLaMA	✓	✗	†	84.42	100.0	100.0	99.42	99.23	100.0	100.0
	✓	✓	97.69	98.21	100.0	100.0	99.61	99.23	99.81	100.0
Gemini-2.5 Pro→CyBERT	✗	✓	32.95	47.28	100.0	97.49	95.76	6.36	0.00	2.89
	✓	✗	†	84.45	100.0	100.0	99.61	99.42	99.81	100.0
Gemini-2.5 Pro→TinyLLaMA	✓	✓	97.30	98.02	100.0	99.42	99.23	99.04	99.23	100.0
	✗	✓	71.87	93.96	100.0	97.87	99.61	96.34	99.81	100.0
Gemini-2.5 Pro→CyBERT	✓	✗	4.82	85.05	100.0	100.0	99.42	99.42	98.84	100.0
	✓	✓	97.50	98.04	100.0	99.81	98.84	99.23	99.81	100.0
Gemini-2.5 Pro→TinyLLaMA	✗	✓	0.39	33.2	36.03	100.0	0.0	0.0	99.61	0.0
	✓	✗	†	84.48	100.0	100.0	99.61	99.61	99.81	100.0

PHASE 2: FEW-SHOT ADAPTATION

Once a few of real samples are identified, the model undergoes few-shot fine-tuning.

GOAL

Seamlessly transition from detecting "unknowns" to accurately classifying them as new classes with minimal data overhead.



EXPERIMENTAL SETUP AND DATASET

Benchmark: CIC-IDS-2018 (~12,000 samples) with benign traffic and eight DoS/DDoS classes spanning volumetric and application-layer attacks.

Protocol: Leave-One-Class-Out—one attack (Slowloris*) is fully excluded during training and treated as an unseen zero-day.

Evaluation: Train on benign + remaining classes; test zero-shot identification of the held-out attack.

CONCLUSION

Phase 0: Anomaly Detection. Successfully identifies almost all malicious flows, including unknown variants

Phase 1: Zero-Shot Identification. Achieves over 97% detection accuracy for zero-day threats by reasoning through semantic similarity.

Phase 2: Rapid Adaptation. Reaches up to 98% accuracy post-adaptation while maintaining performance on previously known attack classes.