ViScamDial: Psychology-Driven Adversarial Synthetic Data for Vietnamese Vishing Detection

Anonymous ACL submission

Abstract

We introduce **ViScamDial**, a large-scale Vietnamese vishing corpus designed to address data scarcity in low-resource fraud detection. Unlike synthetic datasets based mainly on templates or direct prompting, ViScamDial is generated through a theory-grounded adversarial multi-agent framework that integrates nine behavioral psychology principles to model attacker–victim interactions. The resulting corpus contains **12,886** fraudulent and legitimate phone-call dialogues with adaptive scammer strategies, dynamic victim responses, and psychologically grounded termination patterns. Each dialogue is annotated with **96 fine-grained labels** covering entities, persuasion tactics, and psychological states, supporting interpretable analysis beyond binary fraud detection. Experiments on real-world Vietnamese vishing data show that models trained on ViScamDial achieve **85.2% accuracy** and **94.3% AUROC**, outperforming translated synthetic baselines by a large margin. Combining ViScamDial with limited real data further improves accuracy to **94.5%** with a **96.4% scam detection rate**. Cross-category evaluation also demonstrates strong generalization to unseen scam types. These results suggest that psychological realism is a key factor in effective synthetic fraud data generation.

1 Introduction

Voice phishing (vishing) has emerged as one of the most pervasive and sophisticated forms of social engineering worldwide. By exploiting human cognitive biases and psychological vulnerabilities, attackers manipulate victims into disclosing sensitive information, leading to substantial financial losses globally (Federal Bureau of Investigation, 2026). Unlike traditional phishing, vishing exploits real-time conversational dynamics: attackers adapt their persuasion tactics based on victim responses, leveraging established psychological principles such as authority, urgency, and reciprocity (Cialdini, 2006). The rapid expansion of digital banking in Vietnam has made the country a prime target, with scammers increasingly impersonating police, bank officials, and government agencies to manipulate victims into divulging sensitive information.

Automated vishing detection using Natural Language Processing (NLP) offers a scalable defense, enabling real-time analysis of transcribed conversations to flag fraudulent patterns. However, developing robust detection models remains severely hindered by **data scarcity**. Real vishing recordings are exceptionally difficult to obtain at scale due to strict privacy regulations (e.g., GDPR, Vietnam's Personal Data Protection Decree 13/2023), legal liability concerns, and the presence of sensitive Personal Identifiable Information (PII). This challenge is particularly acute for low-resource languages: while several English and Chinese vishing datasets exist (Ma et al., 2025; BothBosu, 2024), **no publicly available annotated corpus covers Vietnamese vishing scenarios with the granularity necessary for training fine-grained detection models**.

Recent work has turned to **synthetic data generation** using Large Language Models (LLMs) to circumvent privacy constraints while scaling dataset creation (Finch and Choi, 2024; Mehri et al., 2022). However, existing synthetic dialogue generation approaches face three critical limitations. First, they produce conversations with **limited strategic variability**: generated scammers follow repetitive persuasion patterns rather than dynamically adapting tactics in response to victim resistance (Campagna et al., 2020). Second, they fail to model **victim suspicion and cognitive states**—a key factor determining whether scam attempts succeed or fail (Stajano and Wilson, 2011). Third, they lack **domain-specific grounding**: synthetic dialogues omit culturally and procedurally accurate details (e.g., Vietnamese banking verification protocols,

Table 1: **Comparison of vishing and related fraud detection datasets.** Email and SMS datasets are included for context given the scarcity of dedicated vishing corpora.

Dataset	Year	Lang	Size	Type	Public	Modality		Conv.	Turns	Label Type			
						Text	Audio			Bin.	Multi	Intent	Token
Vishing & Conversational Scam Datasets													
KorCCViD (Kim, 2023)	2023	KR	2,927	Real	✓	✓	✗	✓	✓	✓	✓	✗	✗
TeleAntiFraud-28k (Ma et al., 2025)	2024	ZH	28.5K	Synthetic	✓	✓	✓	✓	✓	✓	✓	✗	✗
Scam-Dialogue (BothBosu, 2024)	2024	EN	1,600	Synthetic	✓	✓	✗	✓	✓	✓	✓	✗	✗
Multi-Agent Scam (BothBosu, 2025)	2025	EN	1,600	Synthetic	✓	✓	✗	✓	✓	✓	✓	✗	✗
BYU Scam Call (BYU-PCCL, 2025)	2025	EN	1K+	Real	✓	✓	✗	✓	✓	✓	✓	✗	✗
Related Fraud Detection Datasets (Email & SMS)													
Enron Spam (Metsis et al., 2006)	2006	EN	33.7K	Email	✓	✓	✗	✗	✗	✓	✗	✗	✗
SpamAssassin (Apache Software Foundation, 2003)	2003	EN	6K	Email	✓	✓	✗	✗	✗	✓	✗	✗	✗
SMS Spam (Almeida et al., 2011)	2011	EN	5.5K	SMS	✓	✓	✗	✗	✗	✓	✗	✗	✗
ViScamDial (Ours)	**2026**	**VN**	**12,886***	**Synthetic†**	✓	✓	✗	✓	✓	✓	✓	✓	✓

Type: Real = real-world recordings; Synthetic = LLM/rule-based generation; Email/SMS = non-conversational fraud detection. **Label:** Bin. = Binary; Multi = Multi-class; Intent = Turn-level tactic; Token = Entity/tactic tagging. * = Base dialogues; expandable via augmentation. † = Theory-grounded multi-agent generation with psychological principles (ours); see (Ma et al., 2025) for TeleAntiFraud methodology.

police identification procedures) that situate conversations in realistic contexts (Kim et al., 2021).

We argue that generating realistic vishing data requires **explicit integration of behavioral psychology theories** that govern persuasion and deception. To this end, we propose a **theory-driven adversarial multi-agent framework** in which a *Mastermind* agent coordinates *Scammer* and *Listener* (victim) agents, ensuring dialogues exhibit: (1) adaptive scammer strategies guided by Cialdini's six principles of influence (Cialdini, 2006), (2) dynamic victim responses shaped by suspicion levels and cognitive biases, and (3) scenario-specific fraud patterns prevalent in Vietnam. Crucially, we introduce a **Suspicion Score mechanism** that models when victims terminate conversations, preventing unrealistically long scam dialogues and ensuring ecological validity.

Using this framework, we construct **ViScamDial**, the first large-scale Vietnamese vishing corpus with theory-grounded multi-layer annotations. Our contributions are:

- A **theory-grounded multi-agent generation framework** that integrates nine behavioral psychology principles with adversarial simulation, alongside a novel *Suspicion Score* mechanism for modeling realistic victim disengagement behaviors.
- **ViScamDial**: 12,886 Vietnamese dialogues across four fraud scenarios, annotated with 96 labels at conversation, and token levels spanning entity types, persuasion tactics, and victim psychological states.
- **A fine-grained annotation taxonomy** explicitly linking dialogue behaviors to psychological persuasion principles, enabling interpretable analysis of scammer strategies and victim responses.
- **Extensive experiments** on multiple detection tasks demonstrating that theory-grounded synthetic dialogues provide effective supervision for training vishing detection models.

2 Related Work

Vishing Detection Methods. Automated scam detection has evolved from handcrafted features (Sahingoz et al., 2019) to deep learning architectures (Halgaš et al., 2019) and transformer-based models (Devlin et al., 2019). Recent work explores Large Language Models (LLMs) for analyzing social engineering tactics (Park et al., 2025), though high computational costs limit real-time deployment (Uddin et al., 2026). Despite architectural advances, a critical bottleneck remains the scarcity of behaviorally realistic training data; models trained on template-based synthetic datasets fail to generalize to adaptive real-world vishing strategies (Campagna et al., 2020). Most existing approaches focus on coarse-grained classification without modeling the fine-grained persuasion tactics or psychological states of victims, motivating our focus on theory-grounded dataset construction.

Existing Datasets for Scam and Vishing Detection. As summarized in Table 1, publicly

available annotated datasets remain scarce, especially for low-resource languages like Vietnamese and fine-grained token-level annotations. Real-world datasets include KorCCViD (Kim, 2023) (Korean) and BYU Scam Call (BYU-PCCL, 2025) (English), but they suffer from privacy restrictions or a lack of peer-reviewed validation. Synthetic alternatives like TeleAntiFraud-28k (Ma et al., 2025), Scam-Dialogue (BothBosu, 2024), and Multi-Agent Scam (BothBosu, 2025) provide multi-class dialogue labels but lack explicit integration of psychological theories of deception. Classic fraud corpora such as Enron Spam (Metsis et al., 2006) and SMS Spam Collection (Almeida et al., 2011) offer only binary labels and lack conversational structure, while general dialogue datasets like DailyDialog (Li et al., 2017) focus exclusively on benign interactions.

Psychological Foundations of Social Engineering. Social engineering attacks exploit human cognitive biases and behavioral patterns rather than pure technical vulnerabilities (Mitnick, 2002; Hadnagy, 2010). While standard datasets rely on coarse annotations, psychological frameworks suggest that scams are highly interactive, adaptive processes driven by behavioral manipulation. In this work, we address these gaps by mapping fine-grained token-level tactical actions directly to established behavioral models. Due to space constraints, we provide an extensive, detailed synthesis of these psychological and theoretical foundations in Appendix A.

3 The ViScamDial Methodology

3.1 Data Generation Framework

Existing LLM dialogue generation pipelines often collapse into static, template-driven interactions where agents execute linear scripts without context-aware tactical adaptation (Park et al., 2025), failing to capture the adversarial dynamics of real-world social engineering where fraudsters calibrate psychological pressure based on victim feedback.

To address this, we design an adaptive multi-agent pipeline with three core constraints: (i) asymmetric information access between agents, (ii) macro-level tactical planning via an isolated strategist, and (iii) continuous monitoring of the victim's cognitive state. As illustrated in Figure 1, three LLM agents cooperate: a **Mastermind** planner, a **Caller** adversary, and a **Listener** target. The pipeline samples and deduplicates caller-listener occupation pairs, instantiates asymmetric persona cards (the Listener holds the full profile; the Caller receives only leaked information), and uses the Mastermind to generate scenario trajectories before a Dialogue Controller drives multi-turn interaction with validation, suspicion tracking, and seed injection. This operationalizes Interpersonal Deception Theory (Buller and Burgoon, 1996), framing fraud as a co-adaptive exchange rather than a static transmission process.

Asymmetric Persona Initialization. Conversations are initialized with disjoint persona cards to mirror the informational imbalance of real-world vishing vectors. The *Listener card* maintains a complete ground-truth profile, mapping socioeconomic attributes (derived from hierarchical occupation sampling) to hidden psychological traits that govern vulnerability. Conversely, the *Caller card* contains only information a scammer could realistically harvest via data breaches. High-risk authentication targets are strictly excluded from the Caller's initial state and designated as extraction objectives, forcing the adversary to elicit sensitive information through behavioral persuasion rather than direct recall, operationalizing Information Manipulation Theory (McCornack, 1992).

Mastermind Macro-Planning. Before the dialogue exchange begins, an isolated **Mastermind** agent receives both occupations to perform trust-reasoning alignment, establishing a natural intersection and constructing a believable pretext. The Mastermind maps out a macroscopic four-phase attack trajectory:

- **S1 (Approach):** Establish credibility using known victim information and a tailored pretext.
- **S2 (Manipulation):** Induce heavy psychological pressure through cues of authority, urgency, or fear.
- **S3 (Action Request):** Deploy high-risk extraction queries targeting specific authentication data.
- **S4 (Closure):** Handle lingering victim objections or naturally terminate the connection.

Crucially, the Mastermind coordinates this macro-strategy but does not participate in the turn-by-turn text generation, and neither agent can inspect the other's hidden prompt constraints or active psychological attributes.

Cognitive State Tracking and Suspicion Dynamics. To model the evolving cognitive state of the

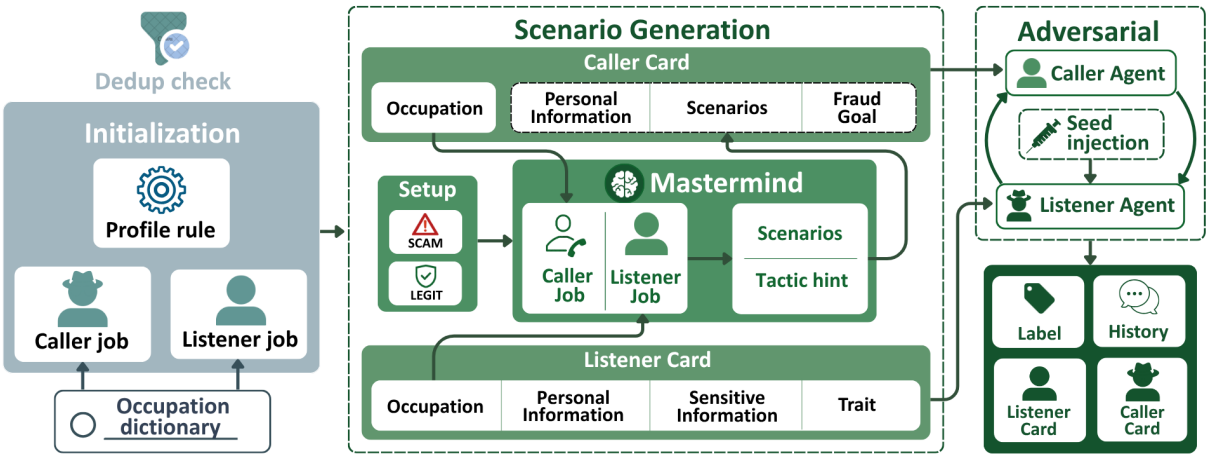

Figure 1: Overview of the ViScamDial generation pipeline.

target under the influence of deceptive pressure, the pipeline maintains a continuous execution-time *Suspicion Score* $S(t) \in [0, 1]$. Grounded in Dual-Process Theory (Kahneman, 2011), this metric quantifies the target's cognitive transitions between emotional compliance (System 1) and analytical resistance (System 2). At each turn t, the suspicion score evolves dynamically based on the interplay between the caller's tactic signal severity, the listener's internal suspicion thresholds ($\bar{\theta}$), and the direct risk associated with the listener's realized action:

$$S(t) = \text{clip}(S(t-1) + \Delta S(t),\, 0,\, 1) \qquad (1)$$

Aggressive persuasion tactics or premature credential requests spike the score toward analytical System 2 resistance, while compliant behaviors or smooth de-escalations temporarily decrease or stabilize it (Workman, 2008).

Termination Hooks and Verification Probes. The dialogue execution is driven by a decentralized controller that terminates the stream when any canonical terminal boundary is breached: successful target data extraction (SCAM_SUCCESS), persistent target refusals (SCAM_REFUSE), or explicit scam identification (SCAM_EXPOSE). To enrich the corpus with realistic negotiating dynamics and break the monotony of binary (pure success/fail) trajectories, the framework injects probabilistic verification probes at intermediate suspicion levels ($S(t) > 0.45$). These probes force the dialogue to simulate realistic partial-resistance loops where victims actively challenge the fraudster or attempt to consult external resources before reaching a final outcome (Jagatic et al., 2007).

Due to space limitations, the granular algorithmic formalization of the dialogue controller, explicit mathematical criteria for the suspicion delta $\Delta S(t)$, exact numerical constants, prompt schemas, and the validation retry architecture are fully detailed in Appendix C.

3.2 Annotation Process

Two-level Annotation Framework ViScamDial employs a two-level annotation architecture to decouple local linguistic cues from global conversational outcomes. At the token level, annotators identify contiguous text spans within individual turns and assign them taxonomy labels, capturing the granular dynamics of the interaction. Simultaneously, at the conversation level, document-wide metadata is recorded to capture the attack outcome, victim psychological state sequences, and quality indicators. This dual-layer design enables the model to learn both the specific markers of manipulation and the overall trajectory of the fraud attempt.

Theory-Grounded Taxonomy The annotation taxonomy maps linguistic patterns to psychological mechanisms across two dimensions: *Entity labels* capture attack structure (fake personas and disclosed victim information), while *Tactic labels* identify psychological manipulation strategies. Labels are grounded in Cialdini's influence principles (Cialdini, 2006) and Speech Act Theory (Austin, 1962; Searle, 1969). PII labels are filtered to include only legally recognized, directly exploitable data for real-world security applications.

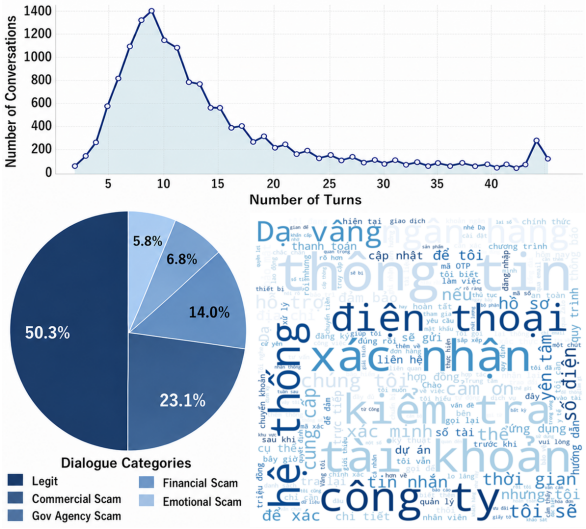

Figure 2: Overview of the ViScamDial corpus.

3.3 Dataset Characteristics

ViScamDial comprises 12,886 multi-turn phone dialogues with a near-balanced distribution of fraudulent (49.7%) and legitimate (50.3%) conversations, totaling 163K turns. The dataset captures realistic long-context scam interactions spanning government, commercial, financial, and social-engineering scenarios, with rich annotations for persuasion dynamics, suspicion evolution, compliance trajectories, disclosure behavior, demographic and psychographic vulnerabilities, and attacker personalization. Sensitive information disclosure occurs in 32.6% of dialogues overall (53.6% of scam vs. 11.9% of legitimate calls), typically around turn 8, reflecting delayed trust-building and multi-stage escalation. Crucially, legitimate dialogues also contain sensitive disclosures (e.g., customers verifying identity with genuine bank or service agents), preventing models from exploiting disclosure presence as a shortcut and forcing them to learn manipulation patterns rather than surface cues. Scam objectives include money transfer, OTP extraction, credential theft, and malware installation, enabling research on conversational manipulation, social engineering, and adaptive scam detection. Figure 2 summarizes the corpus statistics.

4 Experimental Setup

4.1 Research Questions

The experimental evaluation is organized around two main questions concerning the quality and utility of theory-grounded synthetic vishing data.

1. **Does theory-grounded multi-agent generation provide stronger training data than prompt-based and adversarial synthetic baselines?** ViScamDial is compared with three synthetic baselines: Vanilla, TeleAntiFraud-28k, and BothBosu. All models are evaluated on held-out real Vietnamese vishing conversations under both binary and multi-class classification settings.
2. **To what extent can synthetic data substitute for or complement scarce real-world vishing data?** Three training configurations are considered: real-only training, synthetic-only training, and real-synthetic joint training. This comparison quantifies the performance gap between real and synthetic supervision, as well as the complementary effect of combining the two data sources.

In addition, ablation studies examine cross-category generalization, robustness to consecutive-turn filtering, and the effect of different embedding backbones.

4.2 Datasets

The experiments draw on private real-world Vietnamese vishing data, existing synthetic fraud datasets, and the proposed ViScamDial. Table 2 summarises key statistics.

Table 2: Datasets used in experiments. †Private; not publicly released.

Dataset	Source	#Dial.	Lang.
RealTrain†	Real calls	417	VI
RealTest†	Real calls	614	VI
Vanilla	Prompt-LLM	3,497	VI
TeleAntiFraud	Multi-agent	14,364	ZH→VI
BothBosu	Multi-agent	1,600	EN→VI
ViScamDial (Ours)	Theory-grounded	12,886	VI

RealTrain / RealTest are anonymized collections of real Vietnamese call interactions; all PII is removed per internal privacy policy and neither set is publicly released. RealTrain is used only for training (Experiment 3); RealTest is held out exclusively for evaluation across all experiments.

Vanilla is a prompt-based synthetic baseline generated with GPT-4o-mini conditioned on distributional statistics from RealTrain (fraud-goal distribution, conversation length, linguistic markers), giving it privileged access to real-world patterns.

TeleAntiFraud-28k and **BothBosu** are existing synthetic fraud datasets in Chinese and English respectively, translated and culturally localized into

Vietnamese via a unified pipeline (see Appendix E).

ViScamDial is the proposed dataset, constructed through theory-grounded multi-agent simulation without access to RealTrain statistics.

All datasets are normalized to a shared dialogue structure and label space. Each training set is used in full to preserve its original scale. All results are reported on RealTest.

4.3 Our Proposed Model

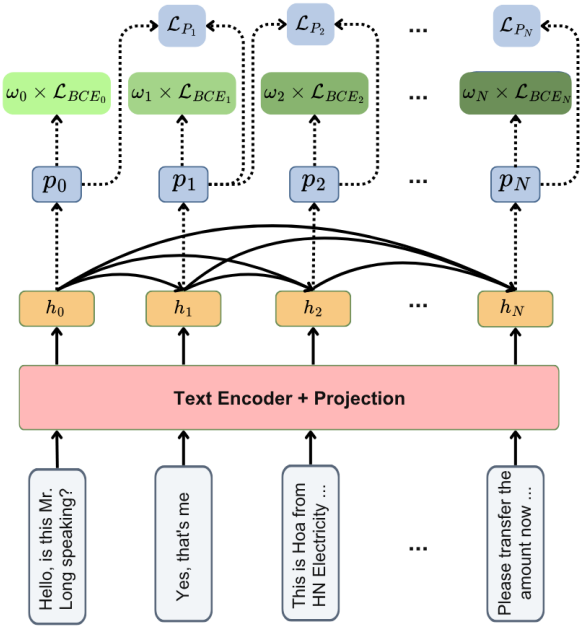

Figure 3: Architecture of ViVOD in binary streaming detection mode.

We propose **ViVOD** (**Vi**etnamese **V**ishing **O**nline **D**etector), a novel framework designed to treat vishing detection as a streaming dialogue classification task, as illustrated in Figure 3. Given a dialogue $D = \{x_1, \ldots, x_T\}$, the model updates the scam probability after each turn rather than waiting until the complete conversation is observed. Each turn x_t is encoded using HaLong-embedding (Ngo, 2024), yielding $e_t \in \mathbb{R}^{d_e}$. The resulting turn representations are linearly projected and contextualized with causal cross-turn attention to produce a per-turn scam probability $p_t \in (0, 1)$.

The training objective is designed to balance early detection with false-alarm control. Let $\tilde{t} = t/N$ denote the normalized turn position. The dialogue is divided into two phases.

Evidence Gathering. For early turns, where $\tilde{t} < 0.5$, a soft-label binary cross-entropy loss is used. The target increases linearly with the amount of observed evidence $\hat{y}_t = \min(2\tilde{t}, 1) \cdot \mathbf{1}[\text{scam}]$. The corresponding loss is

$$\ell_t^{(1)} = \omega_f \left[-\hat{y}_t \log p_t - (1 - \hat{y}_t) \log(1 - p_t) \right], \quad (2)$$

where ω_f controls the contribution of early-turn supervision.

Commitment. For later turns, where $\tilde{t} \geq 0.5$, a hard-label binary cross-entropy loss is applied with a time-dependent weight $\omega_t = 1 + (\omega_{\max} - 1)\text{clip}(2\tilde{t} - 1, 0, 1)$. The loss is defined as

$$\ell_t^{(2)} = \omega_t \cdot \begin{cases} -\log p_t, & \text{if scam,} \\ -\log(1 - p_t), & \text{if legit.} \end{cases} \quad (3)$$

This phase places stronger penalties on late errors, encouraging the model to commit once sufficient conversational evidence has been observed.

The per-turn objective is given by $\ell_t = \ell_t^{(1)}$ for $\tilde{t} < 0.5$ and $\ell_t = \ell_t^{(2)}$ otherwise. The main loss over a mini-batch $\mathcal{B}$ is

$$\mathcal{L}_{\text{M}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} c_b \cdot \frac{1}{N_b} \sum_{t=1}^{N_b} \ell_t, \quad (4)$$

where c_b is an optional class weight and N_b is the number of turns in dialogue b.

A temporal patience regularizer is further introduced to discourage abrupt confidence drops in scam conversations:

$$\mathcal{L}_{\text{P}} = \frac{1}{|\mathcal{B}|} \sum_{b: y_b = 1} \frac{1}{N_b - 1} \sum_{t=1}^{N_b - 1} \max\left(0, p_t - p_{t+1}\right). \quad (5)$$

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{M}} + \lambda_p \mathcal{L}_{\text{P}}, \quad (6)$$

where λ_p controls the strength of the patience regularizer.

The model is trained with AdamW, cosine learning-rate scheduling, and linear warm-up. The encoder is unfrozen after a fixed epoch to stabilize optimization. Evaluation on RealTest reports Accuracy, F1, Detection Rate (DR), False Alarm (FA), AUROC, and Expected Calibration Error (ECE).

5 Evaluation

5.1 Comparison with Synthetic Baselines

We first examine whether theory-grounded multi-agent generation provides stronger training signal

than prompt-based and adversarial synthetic alternatives.

Table 3: Comparison when training on synthetic datasets, evaluated on RealTest.

Binary Classification						
Training Dataset	Acc	F1	DR	FA	ECE	AUROC
Vanilla	76.2	71.6	65.9	15.2	9.8	84.2
TeleAntiFraud	66.3	49.4	36.2	8.7	29.3	76.5
BothBosu	59.4	23.4	13.6	**2.4**	35.4	72.6
ViScamDial	**85.2**	**81.7**	**73.1**	4.8	**8.7**	**94.3**
Multi-class Classification						
Training Dataset	Acc	F1	Pre	Recall		
Vanilla	**66.9**	**64.0**	**62.9**	**65.2**		
TeleAntiFraud	32.6	35.6	42.6	30.6		
BothBosu	30.1	36.4	41.6	32.4		
ViScamDial	55.2	55.7	58.4	53.3		

Table 3 compares training on different synthetic datasets, evaluated on RealTest. For binary classification, ViScamDial achieves the best overall performance (85.2% Acc, 81.7% F1, 73.1% DR, 8.7% ECE, 94.3% AUROC), consistently outperforming translated baselines TeleAntiFraud (66.3% Acc / 49.4% F1) and BothBosu (59.4% Acc / 23.4% F1). Translated datasets obtain lower false alarm rates (BothBosu: 2.4%), but at the cost of detection performance (13.6% DR). For multi-class classification, Vanilla performs best (66.9% Acc / 64.0% F1), with ViScamDial second (55.2% Acc / 55.7% F1), still substantially above translated baselines. This gap arises because Vanilla conditions prompts directly on scam categories, injecting category-specific lexical and procedural cues that benefit fine-grained discrimination, whereas ViScamDial prioritizes psychologically grounded interaction dynamics, letting categories emerge implicitly from conversational trajectories. Overall, explicit category conditioning aids class separability, while ViScamDial is more effective for robust binary scam detection, confirming that theory-grounded generation better captures the linguistic and behavioral patterns of real Vietnamese vishing calls than translated adversarial data.

5.2 Comparison with Real Data

We next investigate to what extent synthetic data can substitute for or complement scarce real-world supervision. Table 4 illustrates the impact of different training datasets on model performance. Due to the limited size and high collection difficulty of authentic data, training solely on RealTrain yields suboptimal binary performance (71.5% Acc, 55.9% F1) and suffers from a severely low detection rate (39.8% DR), despite maintaining high precision. Conversely, the synthetic ViScamDial dataset provides a larger sample volume, significantly improving overall robustness (85.2% Acc, 81.7% F1). The best results are achieved by combining both datasets, leading to a substantial performance leap in both binary (94.5% Acc, 94.4% F1) and multi-class (78.6% Acc, 65.9% F1) configurations. This demonstrates a strong complementary effect: real data anchors precision and mitigates false alarms, while synthetic data expands behavioral coverage to dramatically boost recall and detection capabilities, suggesting that the two sources are complementary rather than substitutable.

Table 4: Comparison when training on real, synthetic, and combined datasets, evaluated on RealTest.

Training	Binary				Multi-class			
	Acc	F1	DR	FA	Acc	F1	Pre	Recall
RealTrain	71.5	55.9	39.8	**2.1**	75.5	61.1	**96.4**	44.7
ViScamDial	85.2	81.7	73.1	4.8	55.2	55.7	58.4	**53.3**
Combine	**94.5**	**94.4**	**96.4**	7.2	**78.6**	**65.9**	89.1	52.3

5.3 Ablation Studies

Table 5: Backbone comparison on the binary classification task.

Backbone	Params	Acc	F1	DR	FA	AUROC	ECE
dangvantuan-embedding	135M	78.8	78.2	78.9	21.2	86.9	**8.6**
BKAI-embedding	135M	79.2	79.1	**91.0**	30.7	92.3	13.7
RoBERTa	500M	81.8	**81.7**	88.5	23.9	92.3	9.7
HaLong-Embedding	270M	**85.2**	**81.7**	73.1	**4.8**	**94.3**	8.7

Embedding backbone comparison. Table 5 compares multiple embedding backbones on the binary vishing detection task. HaLong-Embedding (Ngo, 2024) achieves the best overall performance among the baselines, securing the top Accuracy (85.2%) and a high F1-score (81.7%) at the default threshold. Furthermore, HaLong-Embedding demonstrates the strongest threshold-independent discriminative capability, achieving the highest AUROC (94.3%) and outperforming the heavier RoBERTa baseline (Liu et al., 2019) (92.3% AUROC). Notably, HaLong-Embedding achieves an exceptionally low false alarm rate (4.8%) compared to RoBERTa (23.9%) and BKAI-embedding (BKAI Foundation Models, 2023) (30.7%), while requiring nearly half the parameters of RoBERTa (270M vs. 500M). Although smaller backbones like dangvantuan-embedding (Tuan, 2019) feature a slightly lower

calibration error (8.6% ECE), they suffer from limited detection coverage or higher false alarms. Consequently, HaLong-Embedding offers the most optimal trade-off between discriminative power, false alarm control, and computational efficiency, making it our default backbone.

Table 6: Cross-category generalization (train on 3, test on held-out 4th).

Held-out Category	Acc	F1	DR	FA	AUROC	ECE
InnerTest						
Gov Agency Scam	97.3	95.8	99.6	10.5	**99.6**	1.6
Commercial Scam	91.7	88.0	96.4	10.2	98.5	7.6
Financial Scam	**97.2**	**95.3**	**100.0**	13.1	99.4	2.4
Social/Emotional Scam	93.9	91.0	96.8	**11.4**	98.7	4.5
RealTest						
Gov Agency Scam	85.0	83.6	96.8	53.1	92.7	6.6
Commercial Scam	**86.0**	**84.0**	96.8	48.7	93.1	6.3
Financial Scam	85.7	83.4	95.7	**46.3**	93.8	**4.9**
Social/Emotional Scam	85.3	83.2	**96.1**	48.1	**93.2**	5.8

Cross-category generalization. Table 6 evaluates the model's zero-shot cross-category generalization using a leave-one-out strategy. On InnerTest, the model demonstrates high robustness with an average accuracy of 95.0% and an outstanding detection rate of 98.2% DR, indicating effective semantic knowledge transfer to unseen categories. However, transitioning to the tougher RealTest domain leads to a notable performance drop. While the model preserves its strong detection capabilities (96.4% average DR), it suffers from a major surge in false alarms (49.1% average FA). This suggests that while cross-category transfer successfully identifies scam patterns, distinguishing them from the subtle nuances of harmless real-world dialogues remains a critical challenge.

Robustness to consecutive-turn filtering. Production fraud detection systems require stable alerts to avoid user fatigue from isolated noisy predictions. Figure 4 shows performance when requiring N consecutive turns above threshold before triggering an alert. Single-turn thresholding (N=1) achieves 99.6% detection rate but 59.7% false alarm rate, impractical for deployment. Increasing to N=3-4 reduces false alarms to 25-32% while maintaining 83-91% detection, with alert delay increasing to 2.6 turns. These results indicate ViScamDial-trained models produce temporally coherent predictions, enabling practical deployment configurations that trade latency for reduced false alarms.

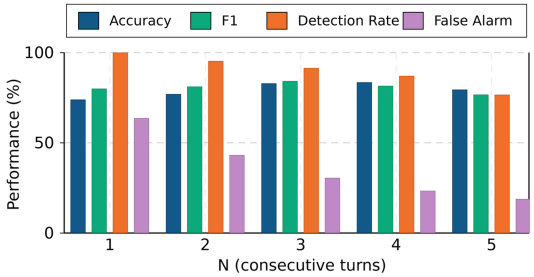

Figure 4: Consecutive-turn filtering trade-offs

5.4 Discussion

Our results suggest two mechanisms underlying ViScamDial's effectiveness. First, psychological realism drives distributional alignment: explicitly modeling authority escalation, suspicion dynamics, and victim resistance produces dialogues whose behavioral signatures more closely match real Vietnamese vishing than translated or prompt-conditioned alternatives, explaining the large gap over BothBosu and TeleAntiFraud. Second, real and synthetic data are complementary rather than substitutable: RealTrain alone lacks behavioral diversity (39.8% DR), while ViScamDial alone lacks precise distributional anchoring, and their combination resolves both failure modes simultaneously. The main open challenge is false alarm control under cross-category transfer: DR generalizes well ($\sim$96% average) but FA spikes to $\sim$49% on RealTest, indicating that category-specific legitimate patterns—not scam recognition—are the primary confound, which we leave to future work.

6 Conclusion

We introduced ViScamDial, the first large-scale Vietnamese vishing corpus constructed using theory-grounded multi-agent generation integrating nine behavioral psychology principles. The dataset comprises 12,886 dialogues with 96 fine-grained annotations. Experiments demonstrate ViScamDial substantially outperforms existing synthetic baselines (85.2% accuracy, 94.3% AUROC), while combining real and synthetic data yields a peak performance of 94.5% accuracy and a robust 96.4% scam detection rate. Cross-category generalization (85.5% average accuracy on unseen categories) confirms models learn transferable behavioral patterns, establishing a practical pathway for robust detection in data-scarce settings.

Limitations

First, while psychologically grounded, ViScamDial is generated synthetically and may not capture all unpredictable behavioral nuances present in real-world vishing attacks, and LLM-based generation may introduce subtle linguistic artifacts. While our empirical results demonstrate strong downstream utility, a comprehensive human evaluation involving cybersecurity and fraud-domain experts is required to definitively quantify psychological realism. Given the scale of the dataset, such validation is beyond the scope of this initial resource paper and remains an active area of ongoing work.

Second, generating adversarial multi-agent dialogues at scale is resource-intensive, making granular ablation of individual pipeline components (e.g., the Suspicion Score mechanism, asymmetric persona cards, or Mastermind planning phase) cost-prohibitive. We note, however, that downstream performance on RealTest—where ViScamDial substantially outperforms Vanilla, which also uses LLMs but lacks these components—provides indirect evidence that the full pipeline contributes meaningfully beyond simple prompt-based generation. Modular ablation is left to future work.

Third, evaluation relies on a small proprietary test set (RealTest) that cannot be released due to telecom privacy regulations and the presence of sensitive personal information. We emphasize that ViScamDial itself and all training code are fully released, so the dataset contribution is reproducible; only the evaluation benchmark is private. Future work will explore sanitized proxy benchmarks to facilitate community comparison.

Fourth, to our knowledge, no prior work addresses vishing detection as a streaming per-turn classification task; existing methods either encode the full transcript as a single input or apply LLMs in zero-shot or few-shot settings over complete dialogues. Because no established baseline exists for this formulation, we propose ViVOD as an initial streaming detector to demonstrate ViScamDial's utility, rather than as a claim to state-of-the-art performance. A direct comparison with full-transcript encoders (e.g., fine-tuned PhoBERT on concatenated turns) and LLM-based classifiers (e.g., Gemini zero-shot) would quantify the latency–accuracy trade-off inherent to streaming detection and is left for future work.

Finally, experiments are Vietnamese-only, and generalization to other low-resource languages remains unexplored. Additionally, although we provide a 96-label token-level taxonomy, evaluations prioritize conversation-level classification; token-level tasks such as real-time tactic extraction are left for future work, though the annotations establish a foundation for such research.

Ethical Considerations

The in-house real-world dataset was compiled through voluntary user fraud reporting, with explicit consent, and underwent rigorous anonymization to remove all personally identifiable information (PII). Similarly, all synthetic personas, organizations, and generated scenarios in ViScamDial are strictly fictitious. We recognize that LLM-generated data may inadvertently encode or amplify societal biases, such as stereotyping specific age groups or occupations as disproportionately susceptible to manipulation. While our framework employs hierarchical, randomized demographic sampling to mitigate these effects, downstream users should remain mindful of potential residual biases when training classification models. Because high-quality synthetic vishing scripts present a potential dual-use risk (i.e., exploitation by malicious actors to refine social engineering tactics), access to the ViScamDial dataset will not be public. Instead, it will be rigorously gated and restricted to verified academic and industry researchers under a strict Data Use Agreement (DUA). Finally, we emphasize that detection models trained on this corpus are intended to augment—not replace—human judgment in real-world fraud investigation workflows, ensuring responsible human-in-the-loop oversight.

References

Tiago A. Almeida, José María G. Hidalgo, and Akebo Yamakami. 2011. Contributions to the study of sms spam filtering: new collection and results. In *Proceedings of the 11th ACM Symposium on Document Engineering*, DocEng '11, page 259–262, New York, NY, USA. Association for Computing Machinery.

Apache Software Foundation. 2003. Apache spamassassin public corpus. Online dataset. Accessed May 2026.

J. L. Austin. 1962. *How to Do Things with Words*. Oxford University Press, Oxford, UK.

BKAI Foundation Models. 2023. Vietnamese bi-encoder: Vietnamese sentence embedding model. Hugging Face model repository. Accessed May 2026.

BothBosu. 2024. Scam dialogue: Synthetic dataset for scam detection. Hugging Face dataset repository. Accessed May 2026.

BothBosu. 2025. Multi-agent scam conversation. Hugging Face dataset repository. Accessed May 2026.

David B. Buller and Judee K. Burgoon. 1996. Interpersonal deception theory. *Communication Theory*, 6(3):203–242.

BYU-PCCL. 2025. Scam call identification dataset. GitHub repository. Accessed May 2026.

Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica Lam. 2020. Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 122–132, Online. Association for Computational Linguistics.

Robert B. Cialdini. 2006. *Influence: The Psychology of Persuasion*, revised edition. Harper Business, New York, NY.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Federal Bureau of Investigation. 2026. 2025 internet crime report. Technical report, Federal Bureau of Investigation.

James D. Finch and Jinho D. Choi. 2024. Diverse and effective synthetic data generation for adaptable zero-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12527–12544, Miami, Florida, USA. Association for Computational Linguistics.

Christopher Hadnagy. 2010. *Social Engineering: The Art of Human Hacking*. Wiley.

Lukáš Halgaš, Ioannis Agrafiotis, and Jason R. C. Nurse. 2019. Catching the phish: Detecting phishing attacks using recurrent neural networks (rnns). page 219–233, Berlin, Heidelberg. Springer-Verlag.

Tom N. Jagatic, Nathaniel A. Johnson, Markus Jakobsson, and Filippo Menczer. 2007. Social phishing. *Commun. ACM*, 50(10):94–100.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Desok Kim. 2023. Text classification of voice phishing transcripts. GitHub repository. Accessed May 2026.

Sungdong Kim, Minsuk Chang, and Sang-Woo Lee. 2021. NeuralWOZ: Learning to collect task-oriented dialogue via model-based simulation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3704–3717, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Zhiming Ma, Peidong Wang, Minhua Huang, Jingpeng Wang, Kai Wu, Xiangzhao Lv, Yachun Pang, Yin Yang, Wenjie Tang, and Yuchen Kang. 2025. Teleantifraud-28k: An audio-text slow-thinking dataset for telecom fraud detection. *arXiv preprint arXiv:2503.24115*.

Steven McCornack. 1992. Information manipulation theory. *Communication Monographs - COMMUN MONOGR*, 59:1–16.

Shikib Mehri, Yasemin Altun, and Maxine Eskenazi. 2022. LAD: Language models as data for zero-shot dialog. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 595–604, Edinburgh, UK. Association for Computational Linguistics.

Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. 2006. Spam filtering with naive bayes – which naive bayes? In *Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS 2006)*.

Kevin D. Mitnick. 2002. *The Art of Deception: Controlling the Human Element of Security*. Wiley.

Hieu Ngo. 2024. Halongembedding: A vietnamese text embedding model. Hugging Face model repository. Accessed May 2026.

Hyunghee Park, Jiwon Lee, Sanghyun Han, and Haewon Byun. 2025. Enhanced voice phishing detection using an llm-based framework for data augmentation and classification. *IEEE Access*, 13:152530–152545.

Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, and Banu Diri. 2019. Machine learning based phishing detection from urls. *Expert Systems with Applications*, 117:345–357.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Frank Stajano and Paul Wilson. 2011. Understanding scam victims: seven principles for systems security. *Commun. ACM*, 54(3):70–75.

Dang Van Tuan. 2019. vietnamese-embedding: Vietnamese sentence embedding model. Hugging Face model repository. Accessed May 2026.

Mohammad Amaz Uddin, Md Mahiuddin, and Iqbal H. Sarker. 2026. An explainable transformer-based model for phishing email detection: A large language model approach. *Computer Networks*, 277:112061.

Michael Workman. 2008. Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security. *JASIST*, 59:662–674.

Supplementary Material

A Extended Related Work: Comprehensive Survey of Theoretical and Legal Foundations

To provide a comprehensive background for the design of our theory-grounded multi-agent framework and fine-grained annotation layers, this appendix details the psychological, cognitive, and communication theories underpinning social engineering and vishing attacks.

A.1 Principles of Persuasion and Behavioral Exploitation

The bedrock of social engineering modeling relies heavily on Robert Cialdini's foundational principles of influence (Cialdini, 2006/2007). Attackers systematically deploy specific compliance-gaining tactics that bypass critical evaluation:

- **Authority:** Attackers impersonate law enforcement, bank officials, or government agents, leveraging institutional trust and fear of legal consequences to force obedience (Ferreira & Teles, 2015).
- **Scarcity and Urgency:** By creating artificial time constraints (e.g., "your account will be locked within 10 minutes"), attackers induce anxiety, forcing rapid decision-making (Workman, 2008).
- **Reciprocity:** Offering unexpected assistance or a fraudulent benefit creates an implicit psychological obligation for the victim to cooperate.
- **Commitment and Consistency:** Attackers guide victims through minor initial concessions (e.g., verifying a phone number) before escalating to high-stakes demands like monetary transfers.
- **Liking and Social Proof:** Establishing false rapport or invoking fake collective behaviors diminishes the victim's skepticism (Jagatic et al., 2007).

A.2 Cognitive Processing and Dual-Process Models

To understand why otherwise vigilant individuals fall prey to voice scams, social engineering research integrates two major cognitive frameworks:

- **Dual-Process Theory (Kahneman, 2011):** Human cognition operates via two modes: *System 1* (fast, emotional, automatic, and heuristic-driven) and *System 2* (slow, analytical, deliberate, and resource-intensive). Vishing scripts are specifically engineered to provoke strong affect—such as panic, acute fear, or greed—which floods the victim's cognitive load. This effectively suppresses analytical System 2 processing and forces reliance on heuristic System 1 compliance.
- **Elaboration Likelihood Model (Petty & Cacioppo, 1986):** This model posits that persuasive messages are processed via either the *central route* (deep scrutiny of arguments) or the *peripheral route* (reliance on surface cues like source credibility or emotional tone). Scammers construct high-pressure environments to ensure victims remain strictly on the peripheral route, reacting entirely to authority markers and situational stress rather than the factual logic of the conversation.

A.3 Interactive Deception and Communication Frameworks

Unlike static phishing emails, vishing is a highly dynamic, turn-by-turn interactive discourse. This necessitates modeling via adaptive communication theories:

- **Interpersonal Deception Theory (IDT) (Buller & Burgoon, 1996):** IDT frames deception as a fluid, co-adaptive exchange. Attackers do not follow rigid scripts; instead, they actively monitor the victim's responses for signs of suspicion, hesitation, or fear, and dynamically adjust their persuasion tactics in real time to maintain credibility.
- **Information Manipulation Theory (IMT) (McCornack, 1992):** IMT demonstrates how deceptive messages deliberately manipulate the four Gricean maxims of cooperative conversation (Grice, 1975): *Quantity* (withholding crucial information), *Quality* (presenting outright falsehoods), *Relation* (introducing irrelevant distractions), and *Manner* (using ambiguous or highly vague phrasing). Scammers systematically violate these maxims to build compelling but entirely fraudulent narratives.
- **Speech Act Theory (Austin, 1962; Searle, 1969):** This framework treats utterances not merely as carriers of information, but as actions intended to produce specific illocutionary forces and perlocutionary effects (e.g., demanding, threatening, or promising). Our fine-

grained token-level annotations capture these tactical speech actions directly as they unfold in dialogue.

B Appendix: Detailed Token-Level Taxonomy

The ViScamDial corpus employs a sophisticated two-layer annotation schema. While Layer 1 focuses on conversation-level outcomes, Layer 2 provides fine-grained token-level annotations across two primary dimensions: Entity labels (covering both fabricated scammer entities and victim-disclosed information) and Persuasion Tactic labels. In total, the taxonomy comprises 96 distinct labels.

B.1 Caller Entities (Group A)

Group A labels identify entities fabricated or impersonated by the caller to establish legitimacy. This group is further divided into Personas, Organizations, and Artifacts.

Table 7: Taxonomy for Caller Personas and Organizations.

Tag	Description
A1: Personas	
PERSONA_AUTHORITY	High-ranking official with coercive power.
PERSONA_STAFF	Service employee (e.g., CS agent, delivery).
PERSONA_EXPERT	Professional consultant (e.g., Broker, IT).
PERSONA_CONTACT	Fake acquaintance, friend, or relative.
A2: Government Organizations	
ORG_POLICE	Police, security forces, or criminal investigators.
ORG_TAX_CUSTOMS	Tax authorities, customs, or fire safety.
ORG_COURT	Courts, prosecutors, or execution agencies.
ORG_LOCAL_GOV	Local administrative offices (commune/district).
ORG_HEALTH_PUB	Public hospitals, Ministry of Health, CDC.
ORG_EDU_PUB	Public schools, Ministry of Education.
ORG_GOV_OTHER	Social security, post office, other state agencies.
A2: Commercial Organizations	
ORG_UTILITY	Electricity, water, or gas providers.
ORG_ECOM	E-commerce platforms (e.g., Shopee, Lazada).
ORG_LOGISTICS	Courier and shipping services.
ORG_RETAIL	Supermarkets or retail chains.
ORG_TELECOM	Mobile network operators.
ORG_TECH	Technology companies (e.g., Apple, Microsoft).
ORG_MEDIA	Social media or news outlets.
ORG_BANK	Commercial banks.
ORG_FINTECH	Digital wallets or Fintech apps.
ORG_INVEST	Fake investment platforms or crypto exchanges.
ORG_INSURANCE	Insurance companies.
ORG_REALESTATE	Real estate agencies or brokers.
A2: Service & Community Organizations	
ORG_HOSPITALITY	Hotels, travel agencies, or restaurants.
ORG_TRAVEL	Airlines or tour operators.
ORG_HEALTH_PRI	Private clinics or pharmacies.
ORG_EDU_PRI	Private centers or online courses.
ORG_RECRUITMENT	Job portals or recruitment agencies.
ORG_CONSULTING	Law firms or financial consultants.
ORG_ENTERTAIN	Entertainment apps or gaming platforms.
ORG_GAMBLING	Online casinos or betting sites.
ORG_CHARITY	Fake charity funds or relief organizations.
ORG_ASSOC	Professional associations or clubs.
ORG_INTL	International bodies (e.g., WHO, UN).
ORG_COMMUNITY	Online community groups (e.g., Zalo/FB).

B.1.1 Personas and Organizations

Personas (A1) define the fake identity, while Organizations (A2) define the entity the caller claims to represent.

B.1.2 Caller Artifacts (A3)

Artifacts are fake documents or tools used to legitimize the scam.

Table 8: Taxonomy for Caller Artifacts.

Tag	Description
ARREST_WARRANT	Fake arrest or summons warrant.
ASSET_FREEZE	Notice of account or asset freezing.
CASE_ID	Official-looking case or file numbers.
BADGE_CARD	Digital image of an ID or badge.
OFFICIAL_DOC	Generic official government documents.
ADMIN_NOTICE	Administrative alerts or warnings.
SERVICE_ID	Fake service or transaction IDs.
LINK / QR	Phishing links or malicious QR codes.
APP	Fake application installations.
REMOTE_TOOL	Screen-sharing or remote control software.
ALERT_SCREEN	Fake system alerts or error pop-ups.
DEEPFAKE_AV	AI-generated audio or video clips.
DEBT_AMOUNT	Specified amount of alleged debt.
PENALTY_AMOUNT	Fine or penalty amount.
REFUND_AMOUNT	Promised refund or reward amount.
RETURN_AMOUNT	Amount to be returned.
GIFT_AMOUNT	Value of a fake gift or prize.
INVEST_DOC	Fake investment portfolios or contracts.
PROFILE	Fabricated professional profiles.
PARCEL_CONTENT	Description of a fake "restricted" package.
TESTIMONY	Fake witness or victim testimonials.
S_NOTICE	Specific legal or administrative notices.

B.2 Victim-Disclosed Information (Group A')

Group A' tags identify sensitive information leaked by the victim during the call.

Table 9: Taxonomy for Victim-Disclosed Information.

Tag	Description
A'1: Identity	
VIC_FULLNAME	Legal full name of the victim.
VIC_ID_DOC	ID numbers (Passport, National ID).
VIC_ADDRESS	Home or workplace address.
VIC_EMPLOYMENT	Job title, company, or department.
VIC_CONTACT	Secondary phone numbers or emails.
VIC_TAX_SOCIAL_ID	Tax ID or Social Security numbers.
VIC_MINOR_INFO	Information about children or dependents.
A'2: Authentication Targets	
VIC_OTP	One-Time Passwords or verification codes.
VIC_CREDENTIAL	Passwords, PINs, or security answers.
VIC_SIM_INFO	SIM ICCID or PUK codes.
VIC_BIOMETRIC	Face ID, fingerprints, or voice prints.
A'3: Financial Information	
VIC_BANK_ACCOUNT	Bank account numbers and bank names.
VIC_CARD_DATA	Credit card numbers, CVV, or expiry dates.
VIC_BALANCE	Account balance or total assets.
VIC_ASSET	Investment portfolios, crypto wallets.
VIC_BEHAVIOR	Transaction habits or banking preferences.
A'4: Sensitive Context	
VIC_HEALTH_DATA	Medical history or health records.
VIC_LOCATION	Real-time location or movement history.
VIC_ONLINE_ACT	Browsing history or app usage.
VIC_FAMILY_MEM	Details about family members.

B.3 Persuasion Tactics (Group B)

Group B labels mark the shortest spans of text that exhibit psychological manipulation tactics.

Table 10: Taxonomy for Persuasion Tactics.

Tag	Description
B1: Cialdini's Principles	
AUTHORITY	Leveraging status or titles to force compliance.
SCARCITY	Creating fake deadlines or limited opportunities.
SOCIAL_PROOF	Citing others who have already complied.
LIKING	Building rapport or fake emotional bonds.
RECIPROCITY	Creating a sense of obligation through "help".
COMMITMENT	Using previous small agreements to force larger ones.
B2: Threats and Isolation	
THREAT_LEGAL	Threatening arrest, lawsuits, or prosecution.
THREAT_FINANCIAL	Threatening account freezes or asset loss.
THREAT_SOCIAL	Threatening to expose secrets to family/peers.
ISOLATION	Demanding secrecy and cutting off support.
OVERWHELM	Flooding the victim with data to paralyze logic.
B3: Cognitive Deception	
HELP	Posing as a "savior" to help the victim.
TRUST_BUILD	Revealing leaked info to prove legitimacy.
CONFESSION	Eliciting a "confession" to create guilt.
REFRAME	Redefining a harmful act as a "protective" one.
B4: Direct Exploitation	
INFO_REQUEST	Direct request for sensitive data.
ACTION_REQUEST	Direct request to transfer money or install apps.
DEFLECT_EXT	Redirecting suspicion to fake official channels.
DEFLECT_MIN	Normalizing unusual requests as "standard".

C Detailed Generation Framework

This appendix provides a complete specification of the ViScamDial generation pipeline: the three-agent architecture and full prompt skeletons, the mathematical formulation of the Suspicion Score with all numerical constants, the persona sampling logic, termination dynamics, and output validation procedures.

C.1 Three-Agent Architecture

C.1.1 The Mastermind Agent (Scenario Director)

The Mastermind does not participate in the dialogue but acts as a centralized strategist. It receives the caller's real occupation, the listener's occupation, the leaked-information dictionary (`known_info`), and a pre-sampled `fraud_goal`. Its primary contribution is the **Intersection Logic**: it identifies a natural point of contact between the two occupations and constructs a believable pretext. For example, if the caller's real occupation is "delivery driver" and the listener is a "housewife," the Mastermind may produce a pretext around a "misdelivered high-value package."

The Mastermind is constrained to produce a JSON object containing:

- `caller_role_job`: the role the caller will impersonate, chosen through an explicit three-step trust-reasoning procedure (candidate professional neighbors → trust-aligned role → sanity check). The prompt explicitly forbids defaulting to high-authority roles when no occupational basis supports them.
- `caller_org`: a real organization, or the literal string `"individual"` when no organization fits, to prevent fabricated organizations.
- `scenario_context`: a one-to-two sentence justification of the caller-listener intersection.
- `scam_path`: a four-phase attack trajectory (S1-S4, Table 11).
- `legit_path`: a parallel three-phase benign trajectory (L1-L3) used when the target outcome is `LEGIT`.

Crucially, the Mastermind does *not* receive the listener's personality traits or the target outcome, operationalizing the asymmetric-information design from Interpersonal Deception Theory: the strategist plans an attack but cannot foresee whether it will succeed.

Phase	Objective
S1: Approach	Establish credibility using known victim info (*weapons*) and a believable pretext.
S2: Manipulation	Induce psychological pressure (fear, urgency, or greed).
S3: Action Request	Request a specific high-risk action (e.g., reading an OTP, transferring money).
S4: Closure	Handle objections or naturally terminate the call after success/failure.

Table 11: The Four-Phase Attack Trajectory defined by the Mastermind.

C.1.2 The Caller Agent (Adversary)

The Caller prompt embeds:

1. The active Mastermind plan and phase sequence.
2. The `known_info` dictionary, framed as "weapons" to be deployed early for credibility.
3. A goal-conditioned tactic guide drawn from `FRAUD_GOAL_TACTICS`, providing per-goal procedural recipes (e.g., for `otp_extraction`: build a security-verification pretext, walk the listener through opening their banking app, request the verification code).
4. A **Dynamic Honorific Module**: Vietnamese first-person pronouns are selected based on the age gap between the Caller's claimed age and the Listener's true age, with an explicit override to authoritative pronouns when de-

ploying AUTHORITY or THREAT_* tactics.

5. Hard rules forbidding (i) `end_call=true` before turn 4 except on explicit exposure, and (ii) OTP/PIN/password requests in phases S1 and S2 (extraction is constrained to S3).

Output is a JSON object containing `utterance`, `current_phase`, `tactic_hint`, and `end_call`.

C.1.3 The Listener Agent (Victim)

The Listener prompt embeds:

1. The full Listener card (Sec. C.2) including all A1-A4 fields, so the agent "knows" its own ground-truth sensitive information.
2. Sampled personality traits with full behavioral descriptions (Table 13).
3. A high-priority compliance-pressure block, injected only for traits with low resistance (e.g., *Naive*, *Authority-susceptible*, *Tech-illiterate*).
4. A tech-literacy modifier flagging unfamiliarity with OTP/2FA terminology when `tech_savvy=False`.
5. Symmetric honorific rules to the Caller, with one-directional pronoun drift allowed under pressure but no oscillation.
6. Goal-specific disclosure templates (Table 12) that constrain the format of the `disclosed_info` field when full compliance occurs.

Output is a JSON object containing `utterance`, `listener_action` (one of six canonical actions, Table 15), `disclosed_info`, `suspicion_delta`, and `end_call`.

Fraud goal	Disclosure template
otp_extraction	`{"VIC_OTP": <ground-truth value from card>}`
credential_theft	`{"VIC_CREDENTIAL": <ground-truth value from card>}`
money_transfer	`{"MONEY_TRANSFERRED": <amount agreed to transfer>}`
malware_install	`{"APP_INSTALLED": <app name agreed to install>}`

Table 12: Per-goal disclosure templates injected into the Listener prompt. The Listener is required to use these exact keys when emitting `disclosed_info` on full compliance, enabling deterministic success detection.

C.2 Persona Construction and Sampling

To balance occupational diversity with realistic socioeconomic distributions, the framework employs **hierarchical occupation sampling**:

1. **Group sampling:** uniformly sample one of 8 socioeconomic groups (G1-G8) from `occupations_grouped_8G.json`.
2. **Occupation sampling:** uniformly sample one job title within the selected group.
3. **Demographic inference:** `tech_savvy` is set True for G2/G5/G6, False for G8, and randomly assigned otherwise; `income_tier` is sampled from {mid, high} for G1-G5, {low, low_mid} for G8, and defaults to `mid` for others. Income tier conditions the synthetic `VIC_BALANCE` and `VIC_ASSET` ranges to ensure financial attributes are plausible for the listener's profession.

Caller and Listener occupations are sampled independently using this hierarchical procedure.

C.2.1 The 13 Personality Traits

The Listener pool contains 13 traits. Each dialogue samples k traits with $k = 1$ at probability 0.4 and $k = 2$ at probability 0.6. Each trait carries both a behavioral description (injected into the Listener prompt) and a numerical suspicion threshold θ_p governing the Suspicion Score update (Eq. 8).

C.2.2 Known-Info Sampling (Caller Weapons)

The `known_info` dictionary handed to the Caller is sampled from the Listener card as follows: `VIC_FULLNAME` and `VIC_CONTACT` are mandatory (a scammer must know who they are calling); 1-3 additional fields are sampled uniformly from the union of A1 (identity), A3 (financial), and A4 (sensitive context). The A2 authentication targets (`VIC_OTP`, `VIC_CREDENTIAL`, `VIC_SIM_INFO`, `VIC_BIOMETRIC`) are *never* placed in `known_info`, enforcing the asymmetric-information design: these must be extracted through dialogue, not recalled.

C.3 Suspicion Score: Mathematical Formulation

To model the evolving cognitive state of the victim, we implement a multi-layer suspicion scoring mechanism. The suspicion score $S(t)$ at turn t is updated as:

$$S(t) = \text{clip}(S(t-1) + \Delta S(t),\, 0,\, 1) \qquad (7)$$

The delta $\Delta S(t)$ is computed from the interplay between the scammer's tactic signal, the victim's

Trait	θ_p	Behavioral description
Naive / gullible	0.35	Trusts confident strangers; defaults to "it's probably fine" over suspicion.
Conservative	0.18	Fears trouble; complies quickly to "keep things safe."
Anxious / worried	0.20	Easily destabilized by fear framings.
Compassionate	0.25	Soft-hearted toward urgency narratives (illness, accidents).
Authority-susceptible	0.30	Defers to claimed officials (police, banks, courts).
Tech-illiterate	0.40	Cannot distinguish fake apps, links, or OTP requests.
Status-conscious	0.28	Susceptible to "VIP" or exclusive-opportunity framings.
Worried-about-children	0.22	Loses control on family-emergency framings.
Skeptical-but-polite	0.08	Suspicious but unwilling to hang up; lingers in the call.
Greed-baitable	0.28	Drawn to promotions, refunds, high-yield investments.
Cautious / vigilant	0.05	Pauses before unusual requests; verifies actively.
Verification-seeking	0.10	Always asks for confirmation before disclosing.
Quick-tempered	0.12	Hangs up or shouts back when pushed past tolerance.

Table 13: The 13 Listener personality traits and their suspicion thresholds θ_p. Lower θ_p corresponds to more skeptical victims who flag manipulation tactics earlier.

personality threshold, and the victim's action:

$$\Delta S(t) = \begin{cases} (\text{signal} - \bar{\theta}) \cdot \text{sens} + \alpha_{\text{action}}, & \text{if signal} > \bar{\theta} \\ -0.02 + \alpha_{\text{action}}, & \text{otherwise} \end{cases} \tag{8}$$

where:

- signal is the goal-conditioned persuasion signal used by the runtime controller. For scam dialogues, when the active fraud goal is available, the controller uses a goal-level prior signal: `otp_extraction`=0.40, `credential_theft`=0.35, `money_transfer`=0.30, and `malware_install`=0.25. When no active fraud goal is provided, the signal is parsed from the Caller's `tactic_hint` by substring matching against the tactic-strength table (Table 14).
- $\bar{\theta} = \frac{1}{|T|} \sum_{p \in T} \theta_p$ is the trait-averaged suspicion threshold for the Listener's sampled traits T (Table 13).
- sens is the trait-averaged sensitivity coefficient.
- α_{action} is the anchor value for the Listener's action on turn t (Table 15).

The -0.02 passive decay applied when the signal is below threshold prevents suspicion from monotonically increasing in quiet conversational turns. The runtime controller updates `suspicion_score` online after each Listener turn using the active fraud-goal prior when available, the Listener action anchor, and the Listener's trait-averaged threshold. The logged per-turn `suspicion_delta` and final `suspicion_score` are stored with each dialogue.

Tactic	Signal strength
REQUEST_OTP	0.40
REQUEST_PASSWORD	0.35
REQUEST_TRANSFER	0.30
SCREEN_SHARE	0.28
INSTALL_APP	0.25
THREAT_LEGAL	0.20
URGENCY	0.15
IDENTITY_VERIFY	0.12
AUTHORITY_CLAIM	0.10

Table 14: Caller tactic signal strengths. Higher values indicate tactics more likely to trigger victim suspicion when detected.

Listener action	α_{action}
CONTINUE	$+0.00$
PARTIAL_COMPLIANCE	-0.05
FULL_COMPLIANCE	-0.15
REFUSAL	$+0.20$
INTERRUPTED_VICTIM_HANGUP	$+0.50$
EXPOSURE	$+0.80$

Table 15: Listener action anchor values α_{action}. Compliant actions reduce suspicion (consistent with cognitive dissonance reduction); resistance and exposure raise it sharply.

C.4 Outcome Distribution Sampling

The target outcome for each dialogue is pre-sampled from a fixed distribution: `LEGIT` (50%), `SCAM_SUCCESS` (40%), `SCAM_REFUSE` (5%), `SCAM_EXPOSE` (5%). Realized outcome frequencies in the released corpus (Section 3.3 of main paper: 50.3% / 40% / 5% / 4.7%)

differ slightly because target outcomes are *aspirational*: the actual termination is determined by the realized interaction. The framework logs both `target_outcome` and `actual_end_reason` for analysis. For scam dialogues, the fraud goal is sampled independently from `money_transfer` (35%), `otp_extraction` (30%), `credential_theft` (20%), `malware_install` (15%).

C.5 Termination Conditions

A dialogue terminates when any of the following fires:

1. **Sensitive disclosure check.** The Listener's `disclosed_info` contains a value matching one of the Listener card's A2 authentication targets (`VIC_OTP`, `VIC_CREDENTIAL`, `VIC_SIM_INFO`) via fuzzy substring match, or contains the synthetic keys `MONEY_TRANSFERRED` or `APP_INSTALLED`. → `SCAM_SUCCESS`.
2. Listener action is `FULL_COMPLIANCE` → `SCAM_SUCCESS` (or `LEGIT_COMPLETED` for benign dialogues).
3. Two consecutive `REFUSAL` actions → `SCAM_REFUSE`.
4. Listener action is `INTERRUPTED_VICTIM_HANGUP` → `SCAM_REFUSE`.
5. Listener action is `EXPOSURE` → `SCAM_EXPOSE`.
6. Caller emits `end_call=true`, overridden if turn < 4 (preventing degenerate two-turn dialogues).
7. At suspicion > 0.9 in late phases (S3 or S4), a stochastic exposure trigger fires with probability 0.3 → `SCAM_EXPOSE`.
8. Maximum turn budget of 40 reached → `MAX_TURNS`.

C.6 Seed Injection (Verification Probes)

To increase behavioral diversity beyond purely successful or failed trajectories, a probabilistic verification probe is injected into the dialogue history (as a special role `listener_seed`, not generated by the Listener LLM) when all of the following hold: (i) turn index ≥ 5, (ii) suspicion score > 0.45, (iii) no probe has been injected previously, and (iv) a Bernoulli draw with $p = 0.4$ succeeds. The probe is one of three templated utterances: asking for an official hotline, asking for the caller's number, or claiming to consult a family member. This mechanism produces partial-resistance trajectories where victims challenge or hesitate mid-call before either complying or refusing, corresponding to the "Seed injection" block in Figure 1.

C.7 Output Validation and Retry Loop

Both Caller and Listener outputs are validated per turn against four conditions: (i) JSON parses successfully (otherwise a brace-extraction fallback attempts to recover); (ii) `utterance` is non-empty, has ≥ 10 characters for the Caller or ≥ 5 for the Listener, and is not in a blocklist of placeholder strings that occasionally leak from the prompt; (iii) for the Caller, `current_phase` must be a single valid phase ID from the active plan; (iv) for the Listener, `listener_action` must be one of the six canonical actions. On validation failure, the turn is retried up to 3 times with an explicit error hint appended to the user prompt. After three consecutive failures, the dialogue is aborted and excluded from the released dataset, ensuring all retained conversations meet structural quality bars.

C.8 Dialogue Classification (Labeling Agent)

After termination, a final classification call assigns one of four scam category labels (*Government Impersonation*, *Commercial Impersonation*, *Digital/Technical Financial Fraud*, *Social/Emotional Fraud*) or *Legit*. The classifier receives the Caller's `role_job`, `caller_org`, fraud goal, pretext, and scenario context, and is prompted with explicit priority rules: (1) impersonating state agencies → Government; (2) impersonating commercial services → Commercial; (3) primarily digital/crypto means → Digital; (4) emotional exploitation → Social. The Labeling Agent runs at `temperature`=0.1 to minimize stochastic class assignment. Legitimate dialogues bypass the classifier entirely and are assigned "Legit" directly. The fine-grained token-level annotations of the 96-label taxonomy described in the main text are derived in a downstream post-processing pass that maps logged `tactic_hint` and `listener_action` fields to taxonomy entries.

C.9 Deduplication and Checkpointing

To prevent duplicate (Caller occupation, Listener occupation) pairs across batches and across parallel workers, the framework maintains a checkpoint file storing the set of completed pairs. File access

is protected by `filelock` to support concurrent execution. On batch start, completed pairs are excluded from sampling; pairs are committed to the checkpoint immediately upon dialogue start ("optimistic locking") and rolled back only on fatal generation failure. This prevents race conditions and duplicate work and corresponds to the "Dedup check" block in Figure 1.

D Dataset Statistics

D.1 Corpus Overview

Table 16: Corpus overview and dialogue statistics.

Statistic	Value
Total dialogues	12,886
Scam	6,405 (49.7%)
Legitimate	6,481 (50.3%)
Turns per dialogue	
Mean / Median	12.66 / 10.0
Min / Max	2 / 41
LLM tokens consumed (all models)	
Mean / Median	56,200 / 45,673
Min / Max	7,143 / 443,288
Dialogues with disclosure	4,200 (32.6%)
First disclosure turn (mean / median)	8.2 / 8.0

D.2 Label Distribution

Table 17: Dialogue label distribution across scam categories.

Category	Count	%
Legitimate	6,481	50.3
Commercial Scam	2,976	23.1
Gov Agency Scam	1,802	14.0
Financial Scam	881	6.8
Social/Emotional Scam	746	5.8
Total	**12,886**	**100.0**

D.3 Listener Demographics

Table 18: Listener demographic distribution.

Attribute	Value	Count	%
Gender	Female	6,614	51.3
	Male	6,272	48.7
Age group	30-44	4,608	35.8
	45-59	4,474	34.7
	<30	2,067	16.0
	60+	1,737	13.5
City tier	Urban	6,521	50.6
	Rural	3,225	25.0
	Suburban	3,140	24.4
Tech-savvy	No	7,929	61.5
	Yes	4,957	38.5

D.4 Listener Occupation Groups

Table 19: Listener occupation group distribution (abbreviated labels).

Occupation Group	Count	%
Skilled labor, agriculture & military	5,533	42.9
Healthcare, education & science	1,710	13.3
Leadership & senior management	1,348	10.5
Service, sales & gig workers	1,261	9.8
Office & administrative support	1,255	9.7
ICT, data & engineering	1,142	8.9
Legal, judicial & regulatory	371	2.9
Finance, accounting & investment	266	2.1
Total	**12,886**	**100.0**

D.5 Scam Tactics and Fraud Goals

Table 20: Primary fraud objectives in scam dialogues (top atomic goals; compound goals excluded).

Fraud Goal	Count	% of Scam
Money transfer	2,246	35.1
OTP extraction	1,922	30.0
Credential theft	1,229	19.2
Malware installation	884	13.8
Other / compound	124	1.9
Total scam	**6,405**	**100.0**

Table 21: Top personal information weapons deployed by scammers (*key_weapons* field).

Weapon type	Freq.	% of Scam
Full name (VIC_FULLNAME)	6,404	99.98
Contact info (VIC_CONTACT)	5,082	79.3
Employment (VIC_EMPLOYMENT)	1,294	20.2
Location (VIC_LOCATION)	1,218	19.0
Address (VIC_ADDRESS)	1,208	18.9
Bank account (VIC_BANK_ACCOUNT)	984	15.4
Tax/Social ID (VIC_TAX_SOCIAL_ID)	925	14.4
ID document (VIC_ID_DOC)	910	14.2
Balance (VIC_BALANCE)	709	11.1
Card data (VIC_CARD_DATA)	672	10.5
Asset info (VIC_ASSET)	563	8.8

Table 22: Victim action distribution across all listeners turns.

Action	Turns	%
CONTINUE	62,882	80.6
PARTIAL COMPLIANCE	4,061	5.2
REFUSAL	8,308	10.7
FULL COMPLIANCE	2,102	2.7
INTERRUPTED / HANGUP	530	0.7
EXPOSURE	74	0.1
Total	**77,957**	**100.0**

D.6 Disclosure Analysis

Table 23: Disclosure rate by scenario.

Scenario	Discl.	Discl. %
Commercial Scam	1,448	48.7
Gov Agency Scam	987	54.8
Social/Emotional Scam	516	69.2
Financial Scam	480	54.5
All scam	3,431	53.6
Legitimate	769	11.9
All dialogues	4,200	32.6

D.7 Generation Cost

Table 24: Token usage and cost by model. Costs reflect actual API spend during dataset generation; per-dialogue figures are averages.

Model	In (M)	Out (M)	In/dial	Out/dial
Claude Haiku 4.5	21.1	15.3	1,641	1,186
Gemini 2.5 Flash	89.0	6.6	6,909	511
GPT-4o-mini	107.5	2.9	8,343	224
Claude Sonnet 4.5	253.9	5.8	19,702	447
DeepSeek-V4 Pro	97.5	17.6	7,564	1,363
DeepSeek-V3.2	52.2	4.3	4,051	333
Gemini 2.5 Flash Lite	49.0	1.6	3,806	122

Total cost: $255.58 ($146.78 input + $108.80 output)
Per-dialogue: mean $0.020 / median $0.018 (min $0.007, max $0.091)

E Cross-Lingual Data Processing Pipeline for Translated Baselines

To ensure fair comparison between ViScamDial and existing non-Vietnamese synthetic vishing datasets (TeleAntiFraud-28k in Chinese, BothBosu in English), we built a unified, multi-stage translation and localization pipeline. This appendix documents the pipeline, model selection rationale, and quality control procedures.

E.1 Pipeline Overview

The pipeline consists of four stages: (1) raw data ingestion and record normalization, (2) machine translation to Vietnamese, (3) cultural localization, and (4) translation status verification. Each baseline dataset is processed through identical stages to produce comparable Vietnamese versions while preserving original dialogue structure, role labels, and scam category annotations.

E.2 Stage 1: Ingestion and Normalization

Raw datasets are first parsed into a unified schema with fields `_id`, `dialogue` (list of role-content turns), `label`, and optional metadata (`thinking`, `result`). For TeleAntiFraud-28k, we separate the original Chinese content from any pre-existing `vi_translation` field, producing two aligned files (Chinese and partially-translated Vietnamese) that share sequential IDs to maintain one-to-one record correspondence. Records lacking the `vi_translation.status = "done"` flag retain original Chinese content as placeholders, ensuring ID alignment is preserved for downstream stages.

E.3 Stage 2: Translation Engines

We evaluated five translation engines spanning open-source NMT, commercial LLMs, and open LLM APIs:

- **VietAI/envit5-translation** (open NMT, seq2seq, 5-beam search, max length 512). Used for English→Vietnamese baselines; long dialogues are split by speaker turn (`Innocent:/Suspect:`) to bypass the 512-token input limit, then reassembled with post-processed role labels.
- **Google Gemini** (`gemini-3-flash`) via the official `google.generativeai` SDK, with rotating API-key support and exponential backoff on rate-limit errors (initial delay 30s, max 3 retries).

- **OpenAI GPT-4o-mini** via OpenRouter (`temperature=0`, single-shot prompting).
- **xAI Grok** (`grok-4-latest`, `temperature=0.3`) via the xAI-compatible OpenAI client.
- **Open LLM APIs** on OpenRouter (Llama 3.3 70B, Gemma 3, StepFun) for localization-only passes on already-translated Chinese content.

For the final translated baselines reported in the main paper, we use Gemini for the primary translation pass and reserve VietAI/envit5 for fallback on samples exceeding LLM rate limits. This matches the OPUS-MT / EraX-Translator / Hirashiba-MT-ZH-VI pipeline described in Section 4, with Gemini 2.5 Pro post-editing.

E.4 Stage 3: Prompt Design and Cultural Localization

All LLM-based translation calls share a common prompt template enforcing three properties: (i) **conversational naturalness** (use of Vietnamese sentence-final particles and discourse markers; subject elision when contextually permitted), (ii) **role-appropriate register** (scammer: serious, threatening, or coaxing; victim: anxious, confused, or skeptical), and (iii) **terminology localization** via fixed mappings:

- Role labels: English speaker tags (`Innocent:`, `Suspect:`) are mapped to canonical Vietnamese equivalents denoting "call recipient" and "scammer".
- Institutions: *Social Security Administration* is mapped to either the Vietnamese Ministry of Public Security (when context implies coercion) or the Social Insurance Agency; *Federal Trade Commission* is mapped to the Vietnamese Trade Commission.
- Identifiers: *social security number / SSN* is mapped to the Vietnamese citizen identification number (CCCD).
- Titles: *Officer* is mapped to Vietnamese equivalents for "official" or "inspector"; English honorifics (*ma'am/sir*) are mapped to context-dependent Vietnamese second-person pronouns.

For Chinese→Vietnamese (TeleAntiFraud-28k), an additional **cultural localization** pass converts China-specific entities to Vietnamese equivalents while preserving dialogue semantics: Chinese personal names → Vietnamese names; Chinese cities (e.g., Beijing, Shanghai) → Vietnamese cities (e.g., Hanoi, Ho Chi Minh City); Chinese banks and payment apps (e.g., Alipay, WeChat Pay, Taobao) → Vietnamese equivalents (e.g., MoMo, ZaloPay, Shopee); and currency conversion (RMB → VND at ∼3,500x). Localization enforces consistent honorific and pronoun selection throughout each dialogue, prohibiting mid-conversation register shifts.

E.5 Stage 4: Post-Processing and Verification

Translated outputs undergo regex-based post-processing to (i) strip language prefixes (`vi:`), (ii) standardize role labels across translation engines (e.g., correcting hallucinated speaker-tag variants back to canonical labels), (iii) normalize whitespace, and (iv) insert line breaks before each speaker label. A separate verification script audits translation completeness by checking three signals per record: dialogue Vietnamese-character ratio, presence of Vietnamese diacritic marks, and divergence from the original Chinese/English text. Records are classified as *fully translated* (all three signals), *partially translated* (1-2 signals), or *not translated* (no signals). Only fully-translated records enter the downstream training pipeline; partial records are re-queued for re-translation.

E.6 Resumability and Cost Control

All translation scripts implement checkpoint-based resumability: per-record results are flushed to disk immediately after generation, with a `processed_ids` set tracked separately. On resumption, scripts skip already-processed IDs, enabling fault-tolerant long-running jobs. Multi-key rotation is implemented for Gemini to distribute load across free-tier quotas. Inter-request delays (1s default, 30s on rate-limit) prevent quota exhaustion.

E.7 Quality Control

To prevent translation artifacts from confounding the baseline comparison in Section 5.1, we apply three filters: (i) length filter (source dialogues $<$50 characters are discarded as degenerate), (ii) Vietnamese-character ratio check ($\geq 20\%$ on output), and (iii) round-trip sanity check on a random 5% sample, where a separate LLM judges whether the translation preserves the original scam intent. Records failing any filter are excluded from baseline training data.

E.8 Limitations of Translated Baselines

We acknowledge that even with cultural localization, translated TeleAntiFraud-28k and BothBosu datasets cannot fully recover Vietnam-specific procedural details (e.g., authentic banking verification flows, MoMo OTP wording, local police identification protocols) that are present in natively-generated ViScamDial dialogues. This residual distributional gap partially contributes to performance differences observed in Table 2, and motivates our framing of the comparison as evidence that *native-language theory-grounded generation* outperforms *translated synthetic data*, rather than isolating the contribution of the theory-grounded framework alone. We further address this concern via the natively-Vietnamese Vanilla baseline (Section 5.1), which controls for translation effects.

F Implementation Details

F.1 Architecture

Each turn x_t is tokenized and encoded independently via HaLong-embedding (Ngo, 2024) (frozen initially), with a maximum token length of 144 per turn and up to 20 turns per dialogue. The CLS representation is projected to a hidden dimension of $d = 256$ via a linear layer, then contextualized by a causal cross-turn attention module with 8 heads and dropout 0.2. For binary detection, a two-layer MLP head ($256 \rightarrow 128 \rightarrow 1$) with GELU activation produces per-turn logits.

F.2 Training Procedure

Training follows a two-stage curriculum. In *Stage 1* (binary detection), the encoder is frozen for the first 3 epochs to stabilize the randomly-initialized heads, then unfrozen for full fine-tuning with gradient checkpointing. In *Stage 2* (multiclass scenario), the encoder is similarly frozen for 5 epochs before unfreezing. Both stages use AdamW with cosine learning-rate decay and linear warm-up over 10% of total steps, early stopping on validation F1 with patience 5 (Stage 1) and 8 (Stage 2). For data augmentation, we apply truncated sub-window sampling: $k = 3$ random contiguous sub-windows per dialogue (minimum 3 turns), generating additional training examples at varying dialogue lengths.

Table 25: Training hyperparameters for Stage 1 (binary) and Stage 2 (multiclass).

Hyperparameter	Stage 1	Stage 2
Learning rate	2×10^{-5}	3×10^{-4}
Batch size	8	8
Gradient accumulation	8	8
Effective batch size	64	64
Weight decay	10^{-2}	10^{-2}
Gradient clip	1.0	1.0
Warmup ratio	0.1	0.1
Max epochs	50	50
Unfreeze epoch	3	5
Early stopping patience	5	8

F.3 Hyperparameters

Table 26: Loss hyperparameters for the hybrid detection loss (Stage 1).

Hyperparameter	Symbol	Value
Max ramp weight	$\omega_{\max}$	3.0
Floor weight	ω_f	0.1
Patience weight	λ_p	0.5
Class weight (benign)	c	1.0

Table 27: Loss hyperparameters for the multiclass contrastive loss (Stage 2).

Hyperparameter	Symbol	Value
Contrastive temperature	τ	0.07
SupCon weight	α	0.5

F.4 Training Splits and Reproducibility

All training runs use a fixed random seed of 42. For each training dataset (ViScamDial, Vanilla, TeleAntiFraud, BothBosu, RealTrain, or their combinations), we apply a stratified 80/20 train/validation split that preserves the binary scam/legit label distribution (and the 4-way scenario distribution in Stage 2). The validation split drives early stopping on F1 with patience 5 (Stage 1) and 8 (Stage 2). Augmented sub-windows (Sec. F.2) are generated only from training dialogues; validation dialogues are never sub-sampled, preventing within-dialogue leakage between the two splits. Final evaluation in all experiments (Tables 2–5) is performed exclusively on the held-out RealTest set, which is never seen during training or hyperparameter tuning.

G Qualitative Analysis: Cross-Turn Attention and Detection Dynamics

Figures 5 and 6 visualize per-turn scam probability and the causal cross-turn attention heatmap for two representative dialogue pairs drawn from the ViScamDial test set. Heatmap entry (i, j) with $i > j$ denotes the attention weight that query turn T_i places on key turn T_j. Both figures show scam (top) and harmless (bottom), with the heatmap on the left and the $P(\text{scam})$ trajectory on the right. All dialogue excerpts below are translated from Vietnamese.

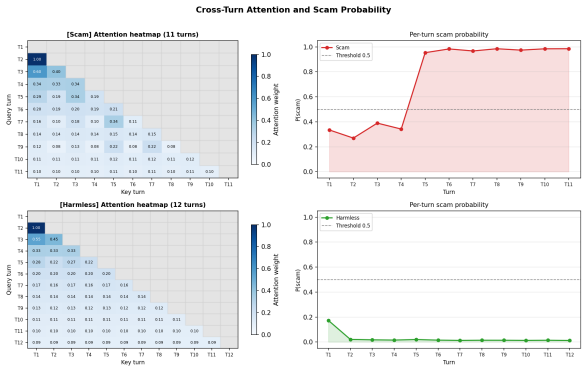

Figure 5: Cross-turn attention and per-turn scam probability for Case 1. Scam: delivery-pretext malware installation (11 turns). Harmless: equipment sales cold call (12 turns). All dialogues originally in Vietnamese.

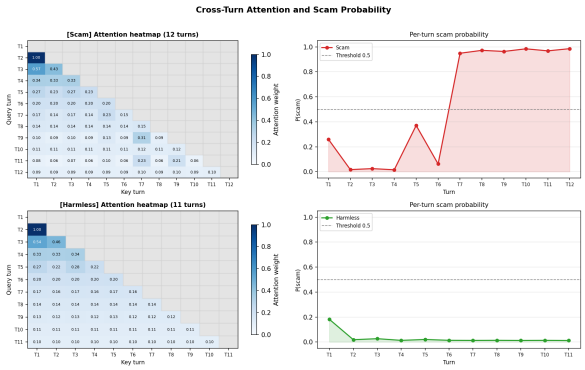

Figure 6: Cross-turn attention and per-turn scam probability for Case 2. Scam: fake HR salary-verification leading to credential theft (12 turns). Harmless: food order clarification call (11 turns). All dialogues originally in Vietnamese.

Case 1: Single-turn trigger after a low-confidence opening (Figure 5). The scammer impersonates a supplier contacting a military-unit manager to confirm a coffee delivery, with the ultimate goal of inducing the victim to install a malicious application disguised as an official military warehouse tool. For the first four turns, $P(\text{scam})$ stays in the range 0.27–0.39, reflecting an ambiguous opening in which the caller deploys the victim's full name and a plausible business pretext (`PERSONA_STAFF` + `PROFILE`). The probability spikes sharply at T5 ($0.34 \rightarrow 0.95$), the turn where the scammer first requests an app installation—a direct `ACTION_REQUEST` co-occurring with fabricated `ORG_GOV_OTHER` legitimacy:

> *(Translated) "To check the order details and confirm them in the military warehouse system, please go to the Play Store and search for an application called 'Military Warehouse', or I can send you the link via Zalo. This is the official application for military units to track and confirm supplied goods."*

The attention row for T5 distributes weight notably toward both T5 (0.34) and T3 (0.34), the earlier turn where the scammer introduced a 24-hour confirmation deadline, indicating the model contextualizes the install request against the prior urgency signal. From T5 onward, confidence stabilizes above 0.95. The harmless dialogue—a routine equipment sales call—collapses to near zero at T2 ($0.17 \rightarrow 0.02$) and remains flat throughout.

Case 2: Oscillating confidence with late commit (Figure 6). The scammer impersonates an HR employee requesting the victim to confirm ATM card digits under the pretext of a monthly salary transfer. Unlike Case 1, $P(\text{scam})$ oscillates for the first six turns (0.26, 0.02, 0.03, 0.02, 0.37, 0.06): T5 produces a local spike when the scammer first explicitly requests card digits, but recedes when the victim again seeks verification at T6. The model commits definitively at T7 ($0.06 \rightarrow 0.95$), the turn that combines an explicit financial threat with a reissued credential request (`THREAT_FINANCIAL` + `INFO_REQUEST`):

> *(Translated) "This information verification is a mandatory procedure to protect your rights. If you do not provide it immediately, your salary this month will be put on hold. Please confirm the last four digits of your ATM card ending in 4699."*

The attention heatmap confirms that T9 places disproportionate weight on T7 (0.31), showing the model treats T7 as a persistent context anchor for

all subsequent turns rather than distributing attention uniformly. The harmless dialogue replicates the rapid-decay pattern of Case 1.

Cross-case observations. Both cases exhibit the same structural pattern: a low-confidence evidence-gathering phase followed by a sharp probability jump at a single high-salience turn that combines authority or urgency framing with a direct extraction request. This is consistent with the two-phase training objective (Section 4.3), which applies soft supervision in early turns and increasingly penalizes missed commits in later ones. The attention mechanism reinforces this behavior by routing subsequent queries back to the trigger turn as a shared anchor, producing temporally coherent predictions. In both harmless cases, the absence of any such trigger causes $P(\text{scam})$ to stabilize near zero within two turns, demonstrating that the model responds to manipulation signals rather than conversational length.

H Real Data Collection and Statistics

Data source and collection. RealTrain and RealTest were collected from production fraud-reporting workflows provided by a major telecommunications and cybersecurity firm in Vietnam. The dataset consists of Vietnamese vishing conversations recorded in a real-world fraud detection environment, where users voluntarily reported suspicious phone calls to a customer support system for investigation purposes. For each reported case, trained analysts reviewed the conversation content together with relevant metadata such as originating phone numbers, call regions, timestamps, and operational information necessary for fraud analysis. Data collection and processing were conducted with user consent and in compliance with internal privacy and data protection regulations. All recordings were anonymized prior to research use by removing personally identifiable information (PII), including personal names, phone numbers, addresses, banking information, and other sensitive identifiers.

Dataset statistics. Table 28 summarizes the composition of RealTrain and RealTest, including the number of scam and legitimate conversations, average dialogue length, and the distribution of fraud categories.

Table 28: Statistics of the real-world Vietnamese vishing dataset.

Statistic	RealTrain	RealTest
Total conversations	417	614
Scam conversations	179	279
Legitimate conversations	238	335
Average turns per dialogue	10.20	12.02
Gov Agency Scam	56	88
Commercial Scam	61	112
Financial Scam	39	55
Social/Emotional Scam	23	24

Annotation protocol. All conversations were manually annotated by professional analysts with experience in fraud investigation and telecommunications security. Each conversation was labeled for binary classification (scam vs. legitimate) and fine-grained fraud category. Annotation disagreements were resolved through expert discussion and adjudication by a senior reviewer.