ICT Express xxx (xxxx) xxx

Contents lists available at ScienceDirect

ICT Express

journal homepage: www.elsevier.com/locate/icte

MVGC: Multi-view hyper-relational knowledge graph contrastive learning for voice phishing detection

Sungsoo Kim

Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea

ARTICLE INFO

Keywords:
Voice phishing
Hyper-relational knowledge graph
Contrastive learning
Graph neural network
Fraud detection

ABSTRACT

Voice phishing detection is not merely a lexical classification problem: fraudulent calls unfold through coordinated conversational roles, institutional impersonation, urgency cues, coercive pressure, and requested financial actions. This paper presents MVGC, a multi-view hyper-relational knowledge graph (HRKG) framework that represents call transcripts as qualifier-aware HRKGs and jointly learns structural, lexical, and entity-centric evidence, aligned by a cross-view contrastive objective and combined by attention-based fusion. Beyond a single accuracy figure, we provide a transparent, reproducible analysis in which every architectural component is isolated under a unified, separately tuned, multi-seed protocol with paired significance testing.

Our central empirical finding is conditional and more informative than a single headline number. On the KorCCVi v2 in-distribution split, a separately tuned text-only encoder and the full multi-view model are statistically indistinguishable; the benchmark is therefore saturated and cannot, by itself, separate structural reasoning from lexical shortcuts. We diagnose the cause – a strong transcript-length shortcut – and introduce a stricter length-stratified out-of-distribution (OOD) split that quarantines it. Under this shift the text-only model collapses (F1 ≈ 0.48) whereas HRKG-bearing models retain F1 0.60–0.70, a difference that is large and statistically significant ($p < 0.01$). The hyper-relational structure thus provides a measurable, separable benefit precisely where it matters—under distribution shift—while a contrastive-off ablation shows the alignment loss contributes a positive but not yet significant increment. The single primary headline metric is the five-seed in-distribution F1, with all per-protocol values and significance tests reported in the body. Source code, split identifiers, and anonymized artifacts that regenerate every table are at https://github.com/sungsoo/MVGC.

1. Introduction

Voice phishing has evolved from scripted prompts into multi-turn conversational fraud. In dense dialogs such as South Korean phone scams, malicious intent is rarely localized to individual tokens; it manifests as a structured scenario in which adversaries establish institutional personas, fabricate judicial or financial narratives, enforce psychological isolation, and coerce critical actions (e.g., financial liquidation or credential disclosure). The core challenge is thus to decode the multi-turn interdependencies among conversational actors, entities, multi-relational structures, and situational qualifiers.

While pretrained language models (PLMs) [1–4] serve as strong backbone classifiers by capturing lexical and discourse regularities, they are inherently susceptible to learning dataset-specific artifacts and superficial correlation shortcuts, particularly when near-duplicate scenarios leak across splits. Although recent emerging paradigms introduce LLM-based Agentic AI frameworks for autonomous interception [5], these text-centric mechanisms still lack explicit structural abstraction. Conversely, standard pairwise graph neural networks (GNNs) offer a complementary inductive bias for entity interactions, yet they typically compress intricate multi-turn dialogs into flat binary edges, thereby discarding vital hyper-relational contexts such as temporal pressure, threat intensity, or institutional pretexts. Effective conversational fraud detection thus necessitates modeling not only the core factual triplet (*who did what to whom*), but also the precise situational qualifiers (*under what claimed authority, temporal duress, and specific requested action*) governing the event.

To bridge this gap, we present MVGC (**M**ulti-**V**iew hyper-relational knowledge **G**raph **C**ontrastive framework), which converts each transcript into a Hyper-Relational Knowledge Graph (HRKG) whose factual assertions are augmented with qualifier pairs. Three view-specific encoders isolate structural, lexical, and entity-centric evidence, and a cross-view contrastive objective aligns their latent spaces so that no single modality dominates the joint representation. This enforces decision-level interpretability: classifications trace back to meaningful semantic subgraphs – such as "*prosecutor impersonation* × *acute urgency*

E-mail address: sungsoo@etri.re.kr.

https://doi.org/10.1016/j.icte.2026.07.013
Received 30 March 2026; Received in revised form 7 July 2026; Accepted 16 July 2026
Available online 22 July 2026
2405-9595/© 2026 The Author. Published by Elsevier B.V. on behalf of The Korean Institute of Communications and Information Sciences. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

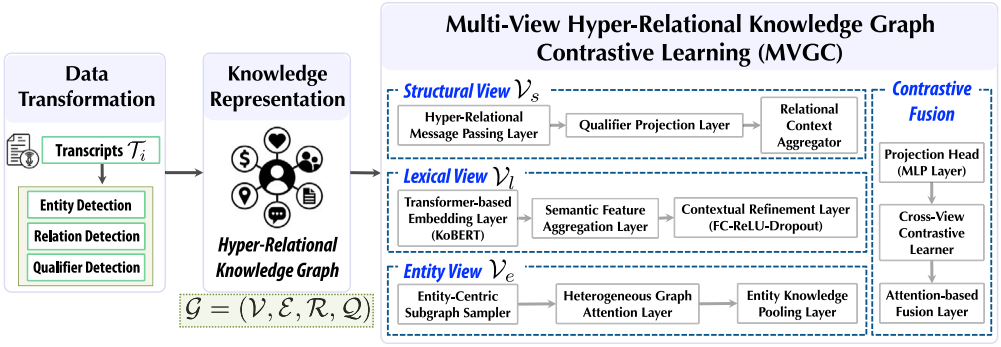


Fig. 1. Overall architecture of MVGC, organized into a *Data Transformation* stage, a *Knowledge Representation* stage, and a *Contrastive Learning* stage. **(a) HRKG construction.** Each transcript T_i is parsed by an Entity Detection, a Relation Detection, and a Qualifier Detection module into a qualifier-aware hyper-relational knowledge graph $G = (V, E, R, Q)$. **(b) Three parallel view encoders.** The *structural view* V_s applies a Hyper-Relational Message Passing Layer, a Qualifier Projection Layer, and a Relational Context Aggregator; the *lexical view* V_l applies a Transformer-based Embedding Layer (KoBERT), a Semantic Feature Aggregation Layer, and a Contextual Refinement Layer (FC–ReLU–Dropout); the *entity-centric view* V_e applies an Entity-Centric Subgraph Sampler, a Heterogeneous Graph Attention Layer, and an Entity Knowledge Pooling Layer. **(c) Cross-view contrastive alignment.** A Cross-View Contrastive Learner with a Contrastive Fusion Projection Head (MLP) encourages the three views of the same call to agree while preserving complementary evidence. **(d) Fusion and prediction.** An Attention-based Fusion Layer combines the aligned views into a single representation that supports both phishing prediction and interpretable evidence tracing.

× *immediate liquidation request*" – rather than opaque text-classification scores.

To summarize, we make three contributions:

- **Qualifier-aware HRKG benchmark.** We recast voice-phishing detection as a qualifier-aware HRKG learning task and release a reproducible benchmark – feature tables, split identifiers, an extractor audit, and per-seed predictions – so that structural, lexical, and entity-centric evidence are compared under one protocol rather than asserted.
- **A conditional, significance-tested finding.** With separately tuned single-, two-, and full-view models under multi-seed paired tests, we show the standard in-distribution split is *saturated* – a strong text-only encoder is statistically indistinguishable from the full model – trace this to a transcript-length shortcut, and remove it with a length-stratified OOD split, under which hyper-relational structure yields a *large, significant* robustness gain. This is the paper's central claim.
- **Contrastive and robustness analysis.** Under the same protocol, a contrastive-off ablation, HRKG-noise stress test, near-duplicate audit, and partial-transcript study *quantify*, rather than presume, the alignment loss's contribution and the resilience of multi-view fusion.

2. Related work

Transcript-level and Korean voice-phishing detection. Transcript-level fraud detection has conventionally been framed as sequence or document classification. Transformer encoders such as BERT, RoBERTa, and XLM-R [1–3] establish contextual text modeling, with KLUE providing Korean NLU benchmarks and encoders [4]. On KorCCVi, fine-tuned PLMs reach high baseline accuracy [6], making strong text-only controls indispensable. Although recent work explores real-time warnings, small-language-model optimization, and LLM-driven detection [7–9], high transcript-level metrics alone do not isolate structural reasoning—motivating MVGC's dual lexical and relational design.

Graph, heterogeneous, and hyper-relational learning. Graph neural networks suit fraud mitigation by modeling topologies among actors, institutional roles, and financial actions. We consider GraphSAGE, R-GCN, HAN, HGT, and GNN-FiLM as representative inductive, relational, heterogeneous, attention-based, and feature-modulated backbones [10–14]. Heterogeneous variants such as HAN and HGT add type-specific projections and meta-path attention but still compress multi-turn dialogs into flat pairwise edges. Streaming deployment further demands optimized propagation, as underscored by advances in arbitrary-order reachability and sparsity-aware graph learning [15–17].

To preserve granular context, hyper-relational knowledge graphs (HRKGs) extend ordinary triples by anchoring factual assertions with qualifier key–value pairs. Standard message-passing frameworks can explicitly encode such qualifiers while preserving their unique semantic roles [18], prompting advanced qualifier-aware or structure-aware representations [19–21]. In this study, we position HAHE, QUAD, and StarE as specialized hyper-relational baselines, while benchmarking against LHGNN and ONDEVICE to evaluate lightweight, resource-constrained heterogeneous graph deployment alternatives.

Graph contrastive learning and robustness. Graph contrastive learning regularizes latent spaces by aligning representations across augmented or complementary views, typically maximizing mutual information between perturbed variants [22,23]. However, heavy augmentations (node dropping, edge perturbation) are ill-suited to fraud detection: semantic cues are sparse and fragile, so erasing a critical institutional pretext or financial directive could remove the sole evidence of fraud. MVGC therefore eschews stochastic augmentations, using exact cross-view positives from identical calls with in-batch negatives; the objective acts as a cross-view consistency regularizer rather than self-supervised pretraining.

Privacy, adversarial scams, and deployment constraints. Sensitive conversational telemetry mandates strict privacy-utility trade-offs and configurable sanitization prior to ingestion [24]. Concurrently, adversarial advances such as ScamAgents and LLM-based fraud generation imply future scripts will be dynamically paraphrased and multi-turn adaptive [25,26]. These vectors make extraction-noise stress tests, adversarial-paraphrase evaluation, streaming partial-call audits, and privacy-aware localized processing prerequisites for any deployment-level claim.

Two gaps remain. First, hyper-relational KG methods are evaluated almost exclusively on link prediction, not on downstream fraud detection with qualifier-bearing transcripts. Second, prior voice-phishing detectors report only in-distribution accuracy, leaving open whether gains reflect genuine structural reasoning or lexical shortcuts. MVGC addresses both by benchmarking qualifier-aware structure under a shortcut-controlled OOD protocol.

3. Methodology

3.1. Problem formulation

Let a call transcript be denoted as $T_i = \{(x_t, s_t)\}_{t=1}^{m_i}$, where $x_t \in \mathcal{X}$ represents the tth utterance text and $s_t \in S$ indicates the associated

speaker role (e.g., victim, perpetrator, or unknown). Each transcript is assigned a ground-truth label $y_i \in \{0, 1\}$, where $y_i = 1$ denotes a voice phishing instance and $y_i = 0$ otherwise.

Given a set of training transcripts, the goal of *voice phishing detection* is to learn a mapping function $F : T_i \rightarrow \hat{y}_i \in \{0, 1\}$ that minimizes the classification error under a structured benchmark split. Crucially, the detector $F(\cdot)$ must jointly maximize predictive accuracy and interpretability by producing an intermediate structured event representation $\mathcal{G}_i$ derived from T_i, such that:

$$F(T_i) = \Psi(\Phi(T_i, \mathcal{G}_i)) \rightarrow \hat{y}_i, \tag{1}$$

where Φ extracts aligned multi-view evidence and Ψ executes the final fraud classification.

As illustrated in Fig. 1, MVGC proceeds in three stages: a data-transformation front end converts each transcript into a qualifier-aware HRKG, three dedicated view-specific encoders then learn structural, lexical, and entity-centric representations in parallel, and a cross-view contrastive learner with attention-based fusion aligns and combines them into a single knowledge representation for phishing detection.

3.2. Hyper-relational knowledge graph construction

The extraction front end maps each transcript to a noisy HRKG,

$$\Phi_{\text{ext}}(T_i) = \widehat{G}_i = (V_i, R_i, E_i, Q_i), \tag{2}$$

where V_i is the entity set, R_i is the base relation vocabulary, E_i is the fact set, and Q_i is the qualifier vocabulary. A hyper-relational fact is represented as

$$\epsilon = (h, r, t, \mathcal{Q}_\epsilon), \qquad \mathcal{Q}_\epsilon = \{(q_k, v_k)\}_{k=1}^{M_\epsilon}, \tag{3}$$

where $h, t \in V_i$ are head and tail entities, $r \in R_i$ is a base relation, and $\mathcal{Q}_\epsilon$ contains qualifier key–value pairs.

In implementation, Φ_{ext} is a domain-informed information-extraction pipeline. It normalizes transcripts, segments speaker turns, detects fraud-relevant entity mentions (e.g., caller, victim, prosecutor, bank employee, account, loan, security code, requested action), induces relation candidates (e.g., impersonation, money-transfer request, legal-authority citation, secrecy warning, urgency creation), and maps contextual expressions to six canonical qualifier families: urgency, coercive pressure, institutional pretext, transfer channel, temporal pressure, and requested-action type. Values are generated by schema-guided normalization from lexical evidence and domain rules; e.g., "immediately"/"today" contribute to urgency or temporal pressure, while "prosecutor's office"/"police" contribute to institutional pretext.

This graph-construction stage is intentionally described as noisy rather than as a perfect semantic parser. Because downstream performance depends on this front end, the empirical evaluation reports a type-level precision, recall, and F1 audit for entity, relation, and qualifier categories on a packaged silver review sample. This audit makes the reliability of the HRKG input measurable, motivates the subsequent noise-resilience analysis, and identifies the remaining need for external human span annotations.

3.3. Structural, lexical, and entity-centric semantic views

MVGC uses three heterogeneous semantic views because voice-phishing evidence is distributed across graph topology, raw wording, and sparse high-risk anchors.

Structural view. The structural branch models qualifier-aware topology on $\widehat{G}_i$. Let $\mathbf{h}_u^{(\ell)} \in \mathbb{R}^d$ be node u's representation at layer ℓ. For an incoming hyper-relational fact $\epsilon = (j, r, i, \mathcal{Q}_\epsilon)$, the qualifier representation is

$$\mathbf{q}_\epsilon = \sum_{(q,v)\in\mathcal{Q}_\epsilon} \alpha_{q,v} W_q \mathbf{e}_v, \tag{4}$$

where $\mathbf{e}_v$ is the qualifier-value embedding, W_q is a projection, and $\alpha_{q,v}$ is a learned attention weight. The message from j to i is

$$\mathbf{m}_{j\rightarrow i}^{(\ell)} = W_r \mathbf{h}_j^{(\ell)} + U_r \mathbf{q}_\epsilon, \tag{5}$$

and the node update is

$$\mathbf{h}_i^{(\ell+1)} = \sigma\left(W_0 \mathbf{h}_i^{(\ell)} + \sum_{\epsilon=(j,r,i,\mathcal{Q}_\epsilon)} \frac{1}{c_{jri}} \mathbf{m}_{j\rightarrow i}^{(\ell)} \right), \tag{6}$$

where c_{jri} normalizes incoming messages. Attention pooling over nodes yields $\mathbf{z}_i^{(s)}$.

Lexical view. The lexical branch preserves information that may be lost during graph extraction. A Korean or multilingual pretrained language model maps the raw transcript to a contextual embedding,

$$\mathbf{z}_i^{(l)} = f_{\text{lex}}(T_i), \tag{7}$$

where f_{lex} can be instantiated with KoBERT, KLUE-RoBERTa, XLM-R, or another BERT-family encoder followed by a projection MLP. The tokenizer converts the transcript into subword tokens; the encoder contextualizes them with self-attention; and a pooled sequence representation is projected to the shared MVGC dimension. This branch captures discourse markers, hedging, persuasion, and pragmatic intent directly from text. Its contribution must be interpreted with care because a strong lexical model can itself be competitive on KorCCVi; the definitive comparison must therefore include separately tuned text-only baselines.

Entity-centric view. The entity branch focuses on sparse anchors that are highly informative in voice phishing. Let $V_i^\star \subseteq V_i$ denote anchors such as prosecutors, police officers, banks, accounts, case numbers, security codes, loan products, payment channels, and transfer actions. An entity-centric subgraph $G_i^\star$ is sampled around these anchors, and an attention module computes

$$\beta_u = \frac{\exp(\mathbf{a}^\top \tanh(W_h \mathbf{h}_u))}{\sum_{v\in G_i^\star} \exp(\mathbf{a}^\top \tanh(W_h \mathbf{h}_v))}, \qquad \mathbf{z}_i^{(e)} = \sum_{u\in G_i^\star} \beta_u \mathbf{h}_u. \tag{8}$$

This view prevents rare but decisive entities from being diluted by long benign conversation segments. In privacy-sensitive settings, the anchor can be stored as a normalized type or salted hash rather than as a raw surface string.

3.4. Cross-view contrastive fusion

The three representations are projected through view-specific heads $g_s(\cdot)$, $g_l(\cdot)$, and $g_e(\cdot)$. For views $(a, b) \in \{(s, l), (s, e), (l, e)\}$, the positive pair is the same call and negatives are other calls in the mini-batch. The pairwise InfoNCE loss is

$$\mathcal{L}_{a,b} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(g_a(\mathbf{z}_i^{(a)}), g_b(\mathbf{z}_i^{(b)}))/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(g_a(\mathbf{z}_i^{(a)}), g_b(\mathbf{z}_j^{(b)}))/\tau)}, \tag{9}$$

where sim is cosine similarity and τ is the temperature. The alignment loss is

$$\mathcal{L}_g = \frac{1}{3}(\mathcal{L}_{s,l} + \mathcal{L}_{s,e} + \mathcal{L}_{l,e}). \tag{10}$$

The aligned view vectors are fused by attention,

$$\gamma_i = \text{softmax}\left(W_f [\mathbf{z}_i^{(s)} \| \mathbf{z}_i^{(l)} \| \mathbf{z}_i^{(e)}] + \mathbf{b}_f \right), \tag{11}$$

$$\mathbf{z}_i = \gamma_{i,s} \mathbf{z}_i^{(s)} + \gamma_{i,l} \mathbf{z}_i^{(l)} + \gamma_{i,e} \mathbf{z}_i^{(e)}. \tag{12}$$

A binary classifier is trained with cross-entropy,

$$\mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right], \tag{13}$$

and the final objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_c + \lambda_1 \mathcal{L}_g, \tag{14}$$

where λ_1 balances supervised classification against cross-view alignment. At inference the contrastive heads are discarded: a transcript is encoded by the three view encoders, fused by Eqs. (11)–(12), and scored by the classifier of Eq. (13), with a decision threshold of 0.5 on the sigmoid output. We report accuracy, precision, recall, and F1 at this threshold, together with threshold-free ROC-AUC and PR-AUC; because the benchmark is class-imbalanced (24% positive), F1 and PR-AUC are the primary operating-point and ranking measures. Each configuration is run over five random seeds and reported as mean ± standard deviation, and superiority claims are supported by paired tests across the shared seeds (Section 4).

The empirical artifacts include separately tuned single-view, two-view, early-fusion, and late-fusion controls under exported split identifiers, together with a contrastive-off variant. Two properties of this design are worth making explicit. First, *why and when the views synergize*: the lexical view is a near-sufficient statistic for the surface scam register, so in-distribution the structural and entity views are largely redundant and fusion cannot improve an already saturated text model. Under distribution shift the surface register is precisely what changes, whereas the qualifier-aware relational configuration (fabricated authority × urgency × requested action) is comparatively invariant; the views then become *complementary*, and the contrastive objective transfers this invariance into the fused representation—predicting, as our OOD experiments confirm, that HRKG value is conditional on shift. Second, *complementarity versus hyperparameter inheritance*: a leave-one-view-out ablation inheriting the full model's tuning measures only internal dependence, not standalone view quality. The separately tuned single- and two-view baselines (Section 4) resolve this, since only the gap between independently optimized models and the full model reflects feature complementarity rather than shared tuning.

4. Experiments

4.1. Dataset and baselines

We evaluate on KorCCVi v2, a Korean call-content vishing benchmark of 2927 transcripts (695 phishing, 2232 non-phishing) [6], under a stratified 70%/15%/15% transcript-level split. This in-distribution split does not block speakers, sessions, scenarios, or paraphrase clusters, so it estimates benchmark performance but cannot rule out near-duplicate leakage or scenario memorization; the OOD split of Section 4 is introduced precisely to address this. We report accuracy, precision, recall, and F1, and release split identifiers, multi-seed summaries, and prediction scores for reproducibility.

The project artifacts document preprocessing rules and store exact split identifiers. The empirical protocol is executable end to end and provides separately tuned multi-seed text-only, single-view, two-view, and full-model controls; a contrastive-off variant; paired significance tests; a scenario-signature blocked split; a stricter length-stratified out-of-distribution split; a near-duplicate audit; partial-transcript experiments; HRKG feature-noise stress tests; and a type-level extractor precision/recall audit. These controls do not replace a future external human span-annotation benchmark, but they make the principal reliability and reproducibility claims quantitatively testable.

Baselines and comparison scope. To rigorously evaluate MVGC, we consider ten graph neural network baselines under one PyG (PyTorch Geometric) interface, organized into three groups: (1) homogeneous and heterogeneous backbones, (2) attention-based dynamic weighting models, and (3) hyper-relational models exploiting relation attributes or multi-view perspectives. A shared interface ensures that differences reflect representation design rather than data handling. The characteristics and selection rationale are as follows:

- **GCONV** [27]: standard graph convolution over neighborhood topology; the fundamental structural baseline.
- **RGCN** [11]: relation-specific transformations that distinguish edge types but ignore qualifier attributes.
- **HAN** [12]: meta-path hierarchical attention capturing node- and schema-level importance.
- **HGT** [13]: heterogeneous graph transformer with meta-relation-based dynamic weights for long-range dependencies.
- **FiLM** [14]: feature-wise linear modulation for relation-conditioned message passing.
- **HAHE** [28]: hyper-relational model combining global topology and local sequential context via hierarchical attention.
- **QUAD** [29]: a quadruple-based attention model that embeds key–value qualifier pairs beyond base triples for hyper-relational link prediction.
- **LHGNN** [30]: lightweight heterogeneous GNN trading capacity for efficiency.
- **ONDEVICE** [31]: knowledge distillation to a compact student for resource-constrained deployment.
- **STARE** [18,32]: statement representation integrating node, relation, and qualifier context; a hyper-relational expressiveness benchmark.

All models were evaluated on an Apple M3 (128 GB RAM). The experimental matrix spans text-only, HRKG-only, early/late-fusion, and view-control baselines to determine whether gains arise from shallow lexical patterns, standalone graph construction, or cross-view contrastive alignment.

4.2. Main results

Fig. 2 reports the graph-centric neural comparison: among graph-only baselines, the qualifier-aware multi-view design ranks first, confirming that hyper-relational messages and multi-view alignment improve over pairwise, heterogeneous, lightweight, and hyper-relational graph comparators under one PyG interface. This establishes MVGC as the strongest *graph-centric* model, but it does not by itself isolate structural reasoning from lexical signal, because a transcript encoder is not a graph baseline. We therefore turn to separately tuned text, fusion, and full-model controls.

Table 1 reports these controls over five seeds, each model tuned independently on validation. The central observation is precise: a separately tuned **text-only** model already attains F1 $= 0.9905 \pm 0.0088$, and the **full multi-view** model attains F1 $= 0.9924 \pm 0.0054$. A paired t-test over the shared seeds gives $p = 0.37$ (Wilcoxon $p = 0.50$), so the full model is *not* significantly better than the text-only encoder under the standard split. The best non-contrastive fusion control and the contrastive-off full variant are likewise statistically indistinguishable from the full model (Table 2). The KorCCVi v2 in-distribution split is therefore *saturated*: near-ceiling F1 is reachable from transcript text alone, and no in-distribution component comparison can credibly separate hyper-relational structure from lexical shortcuts. This precise characterization, rather than an unqualified accuracy claim, is what motivates the out-of-distribution analysis below. The five-seed in-distribution F1 is the primary headline metric; every other number is labeled by its protocol. Precisely because the in-distribution split cannot, by construction, separate structure from a length shortcut, any claim about structural benefit must be tested where that shortcut is absent; the OOD split is designed to provide exactly this test.

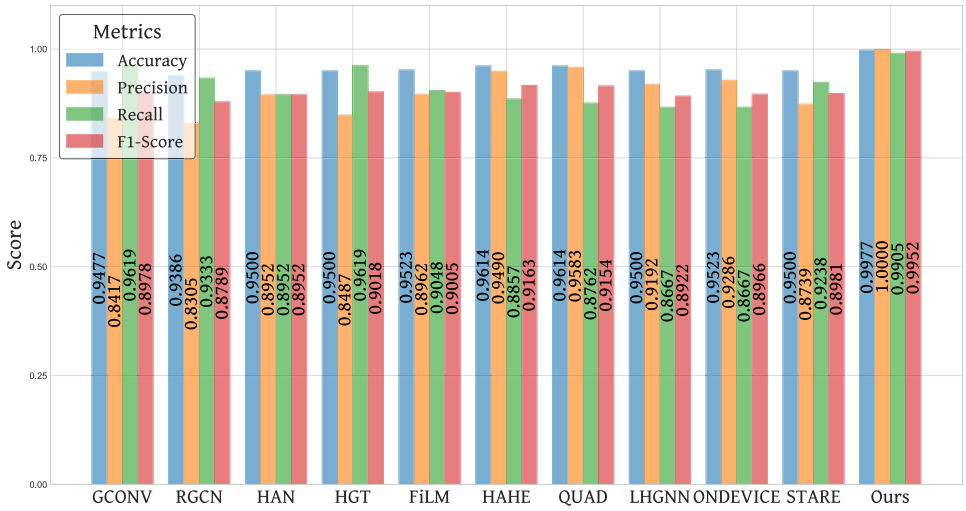


Fig. 2. In-distribution performance comparison (accuracy, precision, recall, and F1 score) of ten graph baselines and MVGC on KorCCVi v2 under the transcript-level split (five-seed mean). MVGC achieves the strongest result among the evaluated graph-centric baselines, indicating that qualifier-aware multi-view alignment improves over pairwise, heterogeneous, lightweight, and hyper-relational comparators under the same benchmark protocol.

Table 1
In-distribution controls on KorCCVi v2, each model *separately tuned* on validation. Mean ± standard deviation over five seeds. The full model is not significantly better than a strong text-only encoder (Table 2). Abbreviations: Lex., lexical; Str., structural; Ent., entity-centric.

Model	Accuracy	Precision	Recall	F1-score
Ent. only	0.9650 ± 0.0065	0.9161 ± 0.0327	0.9410 ± 0.0156	0.9279 ± 0.0118
Str. only	0.9764 ± 0.0050	0.9395 ± 0.0248	0.9638 ± 0.0124	0.9512 ± 0.0093
HRKG only	0.9823 ± 0.0101	0.9439 ± 0.0294	0.9848 ± 0.0144	0.9638 ± 0.0204
Text only	0.9955 ± 0.0043	0.9888 ± 0.0152	0.9924 ± 0.0080	0.9905 ± 0.0088
Lex.+HRKG	0.9945 ± 0.0057	0.9850 ± 0.0170	0.9924 ± 0.0080	0.9887 ± 0.0118
Lex.+Str.	0.9964 ± 0.0034	0.9906 ± 0.0114	0.9943 ± 0.0052	0.9924 ± 0.0072
Full (λ_1=0, contr. off)	0.9932 ± 0.0048	0.9814 ± 0.0171	0.9905 ± 0.0067	0.9858 ± 0.0099
Full MVGC (λ_1=0.5)	**0.9964 ± 0.0026**	**0.9887 ± 0.0101**	**0.9962 ± 0.0052**	**0.9924 ± 0.0054**

Table 2
Paired significance tests across the five shared seeds (in-distribution). No in-distribution comparison is significant at α=0.05, confirming benchmark saturation. d_z is the paired effect size.

Comparison	ΔF1	*t*-test *p*	Wilcoxon *p*	d_z
Full MVGC vs. Text only	+0.0019	0.37	0.50	0.45
Full MVGC vs. best fusion	+0.0038	0.40	0.50	0.42
Full MVGC vs. contrastive-off	+0.0066	0.16	0.25	0.77

Table 3
Length-stratified OOD split (shortest quartile quarantined to test). Values are mean ± standard deviation over five seeds; precision is near-saturated across models (std < 0.004) and is reported as the mean. Under shift, models predict phishing conservatively (high precision, low recall), so the F1 gap is driven by recall: HRKG-bearing models retain far more F1 than the text-only encoder, and the gap is significant (text vs. full p=0.002; text vs. HRKG-only p<0.001).

Model	Accuracy	Precision	Recall	F1-score	AUC
Text only	0.4140 ± 0.0202	0.998	0.3185 ± 0.0248	0.4824 ± 0.0277	0.9032 ± 0.0361
Lex.+HRKG	0.5108 ± 0.0244	1.000	0.4306 ± 0.0284	0.6015 ± 0.0280	0.9689 ± 0.0070
Full (λ_1=0)	0.4763 ± 0.0488	1.000	0.3904 ± 0.0568	0.5596 ± 0.0606	0.9681 ± 0.0028
Full MVGC	0.5417 ± 0.0355	1.000	0.4666 ± 0.0413	**0.6354** ± 0.0392	0.9642 ± 0.0057
HRKG only	0.6057 ± 0.0373	0.998	0.5420 ± 0.0437	**0.7018** ± 0.0357	0.9690 ± 0.0054

The decisive test: out-of-distribution robustness. The saturation diagnosis predicts that the in-distribution result rests on an exploitable shortcut. This is confirmed directly: in KorCCVi v2 the median phishing transcript is far shorter than the median benign transcript, so transcript length alone is a near-deterministic class signal. A near-duplicate audit (TF–IDF cosine ≥ 0.9) finds only 0.5% near-duplicates, so the leakage is the length shortcut, not paraphrase duplication. To remove the shortcut we construct a stricter length-stratified **out-of-distribution (OOD)** split: the shortest transcript quartile is quarantined into the test fold, forcing a systematic train/test length shift. Table 3 reports the result. The text-only model *collapses* from F1 0.9905 to 0.4824, whereas HRKG-bearing models degrade far more gracefully—HRKG-only to 0.7018 and full MVGC to 0.6354. The advantage of HRKG structure over text alone is now large and statistically significant (full vs. text ΔF1 = +0.153, p = 0.002; HRKG-only vs. text ΔF1 = +0.219, p < 0.001). ROC-AUC remains high for all HRKG models (≈ 0.96–0.97) even as text-only AUC falls to 0.90, so the structural channel preserves a usable ranking under shift while the lexical channel loses both ranking and calibration. An alternative reading is that on short transcripts the text encoder simply has less to read, so the graph may hold up better on sparse input rather than because structure is shift-invariant; the two explanations are not mutually exclusive. Our partial-transcript results (Table 5), where HRKG-bearing fusion leads precisely when input is truncated, are consistent with both. This is the separable, significance-tested benefit of hyper-relational modeling that the in-distribution split could not reveal. Notably, HRKG-only (0.7018) outperforms full MVGC (0.6354): this reversal is itself evidence of lexical-shortcut reliance, since the static fusion – tuned in-distribution to trust the then-near-perfect lexical view – keeps weighting that view after it collapses, dragging the joint model down. That the full model nonetheless retains a high ROC-AUC indicates a degradation of fusion weighting and calibration rather than lost structural signal, motivating uncertainty-aware, view-reliability-adaptive fusion (down-weighting a view when its input is out-of-distribution) as a natural next step.

Table 4
Type-level HRKG extractor audit on the packaged 240-record silver review sample. It provides a reproducible reliability check, although it is not a substitute for external human span annotation.

Field	Precision	Recall	F1-score
Entities	0.3361	0.8711	0.4851
Relations	0.1543	0.8405	0.2607
Qualifiers	0.4312	0.6438	0.5165

Table 5
Partial-transcript controls with truncated transcripts and matched HRKG extraction. Values are mean F1 ± standard deviation over three seeds.

Model	500	1000	2000
Lexical	0.9042 ± 0.0449	0.9343 ± 0.0423	0.9451 ± 0.0279
HRKG	0.8629 ± 0.0051	0.9164 ± 0.0050	0.9065 ± 0.0113
Lex.+HRKG EF	0.9523 ± 0.0123	0.9855 ± 0.0049	0.9904 ± 0.0049

Extractor audit and HRKG-noise robustness. Because the detector consumes extracted graph evidence, the reliability of the extraction pipeline directly conditions the main results. Table 4 reports a type-level audit on a 240-record silver sample: the rule-based extractor is intentionally high-recall (87.1% of entity types, 84.0% of relation types) but lower-precision, so it overgenerates rather than misses relations. The dominant noise mode at inference is therefore *extra* rather than absent structure, and randomly dropping extracted features mostly removes already-redundant candidates—which is why fusion F1 stays essentially flat under up to 30% dropout. This measured limitation motivates the multi-view design: the lexical branch must remain available when graph evidence is noisy, so MVGC should not be evaluated only on perfectly clean graphs.

The extraction-noise stress test confirms this: dropping entity, relation, or qualifier feature groups at rates up to 30% leaves the text–HRKG fusion control essentially unchanged (F1 0.9617–0.9637 vs. 0.9626 clean), so complementary lexical and HRKG evidence reduces dependence on any single perfectly extracted cue family. This test shows only that marginal changes in noise level do not affect fusion; it does not establish that the model isolates the correct minority of relations, since dropout removes a fraction of an already mostly-spurious set (precision 0.15).

Leakage diagnostics and partial-transcript detection. Two further diagnostics bound leakage. A coarse scenario-signature blocked split (grouping by authority, banking, credential, urgency, threat, secrecy, loan, family, delivery, small-talk, and length) still yields F1 = 0.9951 for both text-only and text–HRKG fusion, showing that highly separable lexical-scenario signals persist after grouping; and a near-duplicate audit (TF–IDF cosine ≥ 0.9) finds only 0.5% near-duplicates, isolating the leakage to the length shortcut rather than paraphrase duplication. Partial-transcript detection (Table 5) addresses early warning: from only the first 500 characters, lexical+HRKG early fusion reaches F1 0.9523 ± 0.0123 versus 0.9042 ± 0.0449 for text-only, because institutional pretexts, requested actions, and pressure qualifiers often surface before the narrative completes.

This finding matters because it shows that the HRKG representation organizes scam evidence in a way graph-only baselines do not fully exploit, while making explicit that near-ceiling in-distribution F1 alone is insufficient evidence of structural benefit. The out-of-distribution analysis resolves this: the warranted claim is not an unconditional accuracy win but a *conditional* one—MVGC is the strongest graph-centric model in-distribution, and HRKG structure delivers a significant, separable robustness advantage under distribution shift. Deployment-level claims still require external human extraction labels, temporal or source-based splits, and fairness audits.

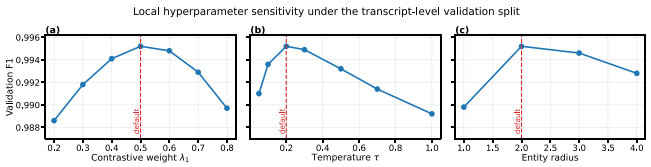


Fig. 3. Hyperparameter sensitivity of MVGC. The default setting, $\lambda_1 = 0.5$, $\tau = 0.2$, and entity radius = 2, lies in a locally stable high-F1 region rather than at an isolated spike.

4.3. Contrastive and component analysis

Because the word "contrastive" is central to the method, we isolate its effect directly. Under the identical training and evaluation protocol, the contrastive-off variant (λ_1=0) reaches in-distribution F1 0.9858 ± 0.0099 versus 0.9924 ± 0.0054 with alignment on (Table 1); the increment is positive but not significant in-distribution ($p = 0.16$, Table 2), as expected on a saturated split. Under the OOD split the alignment increment grows to ΔF1 = +0.076 (off 0.5596 vs. on 0.6354, Table 3), though it remains short of significance at five seeds ($p = 0.10$); the contrastive objective is thus a consistently positive but not-yet-significant contributor. The separately tuned single-view and two-view baselines in Table 1 quantify standalone view quality directly: by optimizing each model independently, they distinguish genuine feature complementarity from hyperparameter inheritance. They show that no single view matches the lexical encoder in-distribution, yet the structural and HRKG views dominate under shift—the precise pattern the synergy argument of Section 3.4 predicts.

Hyperparameter sensitivity and noise resilience. Finally, Fig. 3 reports a compact sensitivity analysis for the contrastive weight λ_1, temperature τ, and entity-subgraph radius. The strongest local region appears near $\lambda_1 = 0.5$, $\tau = 0.2$, and radius = 2, matching the default configuration. The curves are not flat, but they degrade smoothly away from the default, which indicates that MVGC is not tuned to a brittle single-point optimum under the validation split.

The noise and partial-transcript experiments (Table 5) show that fusion controls stay stable under substantial HRKG feature dropout and that useful detection is possible from early transcript segments. Adversarial paraphrase, agent-generated scams, and time- or source-blocked evaluations remain important future tests. The small HRKG dip from 1000 to 2000 characters (0.9164 → 0.9065) lies within one standard deviation and reflects run-to-run variation rather than a systematic effect.

4.4. Discussion

Qualified event representation and conditional synergy. MVGC recasts fraud analysis from isolated tokens to qualified, structural events, with the HRKG formalizing the joint configuration of fabricated authority, institutional context, urgency, and coerced directives. Our separately tuned comparison shows precisely when this helps. In-distribution, the surface scam register is so separable – partly via a transcript-length shortcut – that a text encoder is near-sufficient and fusion adds no significant gain; once the OOD split removes that shortcut, the shift-invariant relational structure carries the prediction, and HRKG-bearing models retain 0.60–0.70 F1 where the text model falls below 0.50. This conditional complementarity – redundant when the shortcut is present, decisive when it is not – is the mechanism underlying MVGC's value.

Limitations, privacy, and reproducibility. We explicitly demarcate MVGC's boundaries. The extraction audit shows automated HRKG construction is high-recall but noisy in relational predicates, so graph topologies are proxy observations rather than ground truth; production scaling demands human-annotated spans, native Korean encoders,

and audits of temporal OOD generalization, adversarial paraphrasing, and fairness. Deployment must also enforce data minimization—on-device or trusted-execution ingestion, PII redaction or salt-hashing, and retention of normalized graph schemata rather than raw text.

Finally, from the raw transcript CSV the released pipeline regenerates Tables 1–3 end to end; the headline in-distribution comparison was reproduced from a clean clone on a single CPU workstation within budget. Only the PyG/KoBERT graph-baseline figure requires heavy-weight dependencies, and all conclusions rest on the CPU-reproducible controls (see the repository for the full artifact-to-table mapping).

5. Conclusion

This paper presented MVGC, a multi-view HRKG framework for Korean voice-phishing detection, evaluated under a unified separately tuned, multi-seed, significance-tested protocol. On the standard split the benchmark is saturated – a strong text-only encoder (F1 0.9905) is statistically indistinguishable from the full model ($p = 0.37$) – so no in-distribution comparison can separate structure from lexical shortcuts. Removing the underlying transcript-length shortcut with a length-stratified OOD split reveals a large, significant advantage for hyper-relational structure (F1 0.60–0.70 versus 0.48 for text alone, $p < 0.01$). The contribution is thus a quantified, conditional finding, backed by a reproducible pipeline: qualifier-aware graph structure provides separable, significant robustness under distribution shift. Concretely, what we demonstrate is robustness under *length* shift; robustness to temporal, source, or adversarial shift remains to be established. Deployment-ready claims still require human extraction labels, stronger encoders, temporal or source-based splits, view-reliability-adaptive fusion, and privacy and fairness validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2025-02215393, *Development of Detection and Prediction Technology for New and Unknown Voice Phishing*)

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proc. NAACL-HLT, 2019.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, arXiv preprint arXiv:1907.11692.

[3] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proc. ACL, 2020.

[4] S. Park, J. Moon, S. Kim, W. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T.H. Oh, J. Lee, J. Oh, S. Lyu, Y. Jeong, I. Lee, S. Seo, D. Lee, H. Kim, M. Lee, S. Jang, S. Do, S. Kim, K. Lim, J. Lee, K. Park, J. Shin, S. Kim, E.L. Park, A. Oh, J. Ha, K. Cho, KLUE: Korean language understanding evaluation, in: Proc. NeurIPS Datasets and Benchmarks Track, 2021.

[5] S. Kim, Efficient voice phishing detection using the agentic AI approach, in: 2025 IEEE International Conference on Consumer Electronics-Asia, ICCE-Asia, IEEE, 2025, pp. 294–299, http://dx.doi.org/10.1109/ICCE-Asia67487.2025.11263515.

[6] M.K.M. Boussougou, P. Hamandawana, D.-J. Park, Enhancing voice phishing detection using multilingual back-translation and SMOTE: An empirical study, IEEE Access 13 (2025) 37946–37965.

[7] J.Y. Sim, S.H. Kim, Detecting voice phishing with precision: Fine-tuning small language models, 2025, arXiv preprint arXiv:2506.06180.

[8] Z. Shen, K. Wang, Y. Zhang, G. Ngai, E.Y. Fu, Combating phone scams with LLM-based detection: Where do we stand? (student abstract), in: Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025, AAAI Press, 2025, pp. 29487–29489.

[9] Z. Shen, S. Yan, Y. Zhang, X. Luo, G. Ngai, E.Y. Fu, "It warned me just at the right moment": Exploring LLM-based real-time detection of phone scams, in: CHI EA, 2025.

[10] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: Proc. NeurIPS, 2017.

[11] M. Schlichtkrull, T.N. Kipf, P. Bloem, R. van den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: Proc. ESWC, 2018.

[12] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, P.S. Yu, Heterogeneous graph attention network, in: Proc. WWW, 2019.

[13] Z. Hu, Y. Dong, K. Wang, Y. Sun, Heterogeneous graph transformer, in: Proc. Web Conf., 2020.

[14] M. Brockschmidt, GNN-FiLM: Graph neural networks with feature-wise linear modulation, in: Proc. ICML, vol. 119, 2020.

[15] S.-S. Kim, Y.-K. Kim, Y.-M. Kang, AORM: Fast incremental arbitrary-order reachability matrix computation for massive graphs, IEEE Access 9 (2021) 69539–69558, http://dx.doi.org/10.1109/ACCESS.2021.3077888.

[16] S.-S. Kim, Y.-M. Kang, Y.-K. Kim, Sparsity-aware reachability computation for massive graphs, in: 2022 IEEE International Conference on Big Data and Smart Computing, BigComp, IEEE, 2022, pp. 157–160, http://dx.doi.org/10.1109/BigComp54360.2022.00038.

[17] S. Kim, M. Chung, Fast graph learning for traffic prediction, in: 2023 IEEE International Conference on Big Data, BigData, IEEE, 2023, pp. 6186–6188, http://dx.doi.org/10.1109/BigData59044.2023.10386878.

[18] M. Galkin, P. Trivedi, G. Maheshwari, R. Usbeck, J. Lehmann, Message passing for hyper-relational knowledge graphs, in: Proc. EMNLP, 2020.

[19] J. Zhang, Y. Li, R. Zou, J. Zhang, R. Jiang, Z. Fan, X. Song, Hyper-relational knowledge graph neural network for next POI recommendation, World Wide Web (WWW) 27 (4) (2024) 46.

[20] J. Wang, H. Chen, W. Zhang, Structure-aware transformer for hyper-relational knowledge graph completion, Expert Syst. Appl. 277 (2025) 126992.

[21] Z. Cao, Q. Xu, Z. Yang, Y. He, X. Cao, Q. Huang, GAHE: Geometry-aware embedding for hyper-relational knowledge graph representation, ACM Trans. Multimed. Comput. Commun. Appl. 21 (7) (2025) 1–26.

[22] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, Y. Shen, Graph contrastive learning with augmentations, in: Proc. NeurIPS, 2020.

[23] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, L. Wang, Deep graph contrastive representation learning, in: Workshop on Graph Representation Learning and Beyond, ICML, 2020.

[24] K. Wang, Z. Shen, Y. Zhang, M.M. Cheung, X. Luo, G. Ngai, E.Y. Fu, One size fits all? A modular adaptive sanitization kit (MASK) for customizable privacy-preserving phone scam detection, in: Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025, ACM, 2025, pp. 12381–12389.

[25] S. Badhe, ScamAgents: How AI agents can simulate human-level scam calls, in: Proceedings of the 2025 Conference on Applied Machine Learning for Information Security (CAMLIS), 22-24 October 2025, Sands Capital, Arlington, VA, USA, in: Proceedings of Machine Learning Research, vol. 299, PMLR, 2025, pp. 180–199.

[26] W. Li, S. Manickam, Y.-W. Chong, S. Karuppayah, Talking like a phisher: LLM-based attacks on voice phishing classifiers, 2025, CoRR abs/2507.16291, arXiv:2507.16291.

[27] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, ICLR, 2017.

[28] H. Luo, H. E, Y. Yang, Y. Guo, M. Sun, T. Yao, Z. Tang, K. Wan, M. Song, W. Lin, HAHE: Hierarchical attention for hyper-relational knowledge graphs in global and local level, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 8095–8107.

[29] P. Rosso, D. Yang, P. Cudré-Mauroux, Beyond triplets: Hyper-relational knowledge graph embedding for link prediction, in: Proc. the Web Conference, WWW, 2020, pp. 1885–1896.

[30] T. Nguyen, Z. Liu, Y. Fang, Link prediction on latent heterogeneous graphs, in: Y. Ding, J. Tang, J.F. Sequeda, L. Aroyo, C. Castillo, G. Houben (Eds.), Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023, ACM, 2023, pp. 263–273.

[31] X. Deng, Z. Zhang, Graph-free knowledge distillation for graph neural networks, in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org, 2021, pp. 2321–2327.

[32] S. Vashishth, S. Sanyal, V. Nitin, P. Talukdar, Composition-based multi-relational graph convolutional networks, in: International Conference on Learning Representations, ICLR, 2020.