CVNSS4.0 IR-Based Reference IDs as Semantic Coordinates for Vietnamese Digital Infrastructure: A Precomputed Vector Registry Approach

Dai-Long Ngo-Hoang*
* Vietnam National University, Ho Chi Minh City, Vietnam (VNUHCM)
Email: nhdlong@vnuhcm.edu.vn

Abstract:

Modern natural-language processing pipelines usually transform text into tokens, tokens into vectors, and vectors into downstream representations. Nevertheless, not every tokenization strategy has the same computational or semantic role. Byte-pair encoding (BPE) and related subword tokenizers are effective for language-model prediction, but their token identifiers are model-internal and do not constitute stable semantic references. This paper proposes an IEEE-style conceptual architecture in which Vietnamese expressions are first normalized through a CVNSS4.0 intermediate representation (IR), then mapped to registry-stable reference identifiers, and finally associated with precomputed semantic vectors. The key analogy is geographic coordinates: a place name may vary, but a coordinate in a reference frame enables consistent localization. Likewise, a Vietnamese concept such as "traceability" may be mapped to a stable identifier, e.g., ID 30588, which functions as a discrete semantic coordinate. A precomputed embedding attached to this identifier functions as a continuous coordinate in semantic space. The resulting architecture separates the expensive phase of vector generation from the lightweight phase of vector usage. It enables O(1) embedding lookup, persistent vector indexing, compact QR/NFC/RFID payloads, auditable blockchain metadata, semantic GIS attributes, chipless RFID encoding, and low-latency edge-AI inference.

Index Terms:

CVNSS4.0, intermediate representation, reference ID, semantic registry, embedding lookup, Vietnamese NLP, vector search, RFID, GIS, blockchain, edge AI.

I. INTRODUCTION

The dominant computational view of language in current artificial-intelligence systems is token-centric: an input string is segmented into tokens, the tokens are mapped to vectors, and the vectors are processed by a neural architecture. In large language models, this pipeline is typically implemented through BPE, WordPiece, SentencePiece, or related subword tokenizers. These methods are highly effective for open-vocabulary modeling and next-token prediction; however, the token identifiers they produce are usually internal to a specific tokenizer vocabulary and model version. They should not be treated as permanent identifiers for concepts, assets, records, or physical objects.

This paper develops a different but complementary proposition: Vietnamese can be mapped into a controlled semantic addressing layer through a CVNSS4.0 intermediate representation and a registry of stable reference IDs. In such a system, a Vietnamese lexical unit, phrase, concept, command, product descriptor, GIS attribute, or traceability term receives a registry-managed identifier. Once that identifier is stable, a semantic vector can be precomputed and stored. Runtime systems no longer need to re-tokenize and re-encode the expression through a large model. Instead, they may simply perform an ID-to-vector lookup.

The conceptual contribution is best understood through a geographic analogy. A place may have many names, spellings, historical variants, and abbreviations. Yet, once it is assigned a coordinate in a known geodetic reference frame, a machine can locate it consistently. Similarly, a Vietnamese concept may have multiple orthographic and contextual forms. Once normalized through CVNSS4.0 IR and mapped to a stable reference ID, the concept becomes addressable by machines across applications. The ID is a discrete coordinate; the embedding vector is a continuous coordinate in semantic space.

II. CONCEPTUAL BACKGROUND

A. BPE as Statistical Segmentation

BPE and other subword tokenizers are designed to reduce vocabulary sparsity and to improve language-model coverage. They learn frequent character or symbol merges from a corpus and form a vocabulary that reflects statistical patterns in the training data. This is powerful for neural translation and language generation, especially for rare words and morphologically diverse inputs.

However, a BPE token ID is not a public semantic coordinate. Its meaning is bound to a tokenizer vocabulary, a training corpus, and a model release. The same integer in another tokenizer may refer to a completely different string fragment.

Therefore, although BPE is deterministic within one tokenizer, it is not suitable as a cross-platform identifier for registries, QR payloads, RFID tags, blockchain metadata, or GIS objects.

B. Reference IDs as Controlled Semantic Addresses

A reference-ID approach starts from an opposite design objective. Instead of learning identifiers as a statistical byproduct, it explicitly assigns identifiers to controlled lexical or conceptual entries. Each entry may store the original Vietnamese expression, the CVNSS4.0 IR form, the domain, the concept type, version information, binary and hexadecimal encodings, and one or more precomputed vectors.

In this architecture, the registry is analogous to a semantic reference frame. An ID is not merely an integer. It is a durable address that can be audited, transmitted, indexed, and reused across devices. It can also be mapped to physical encodings such as QR, NFC, RFID memory blocks, or chipless RFID notch patterns.

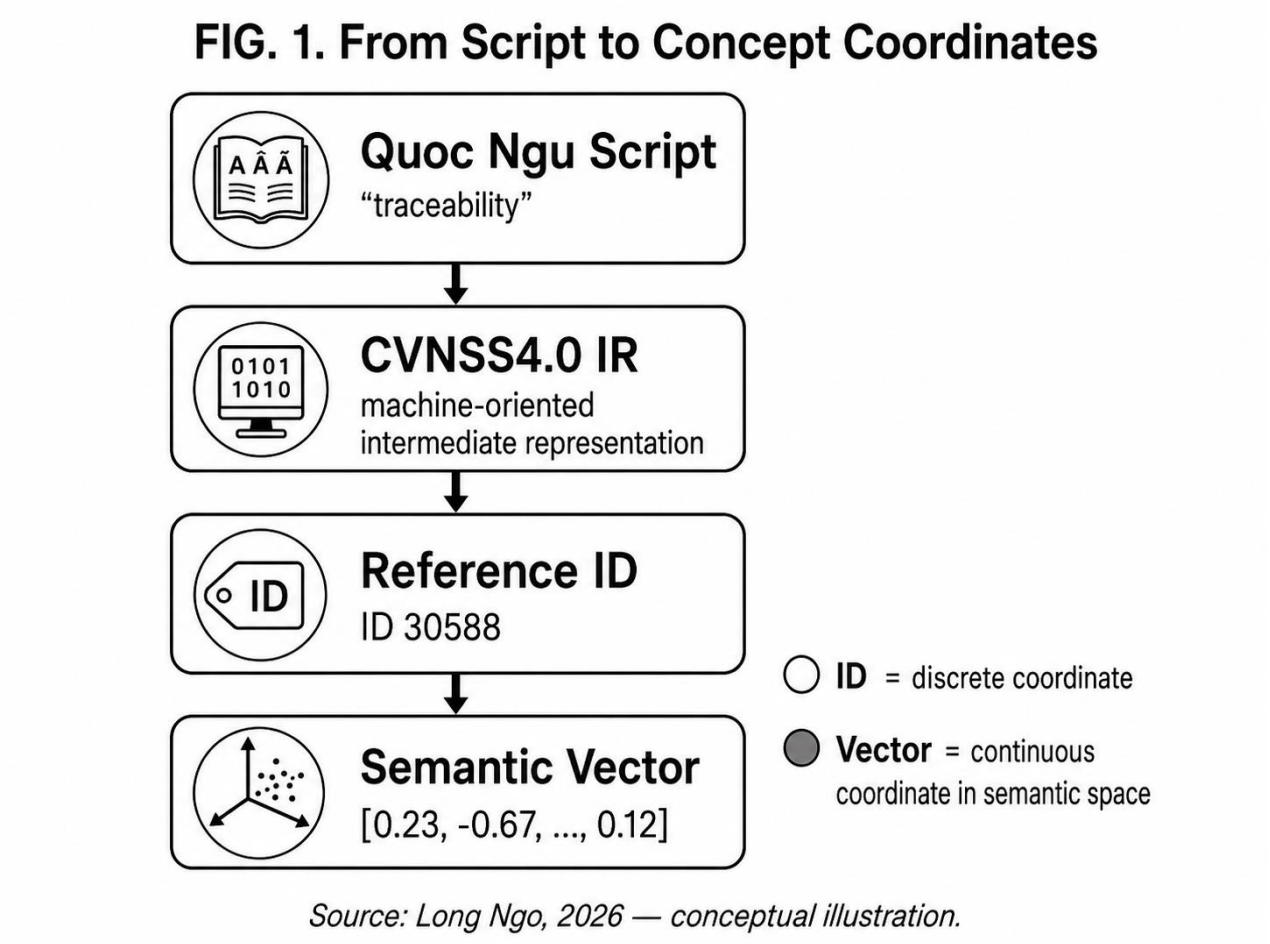

Fig. 1. From script to concept coordinates: Quoc Ngu script is normalized through CVNSS4.0 IR, mapped to a stable reference ID, and associated with a semantic vector. Source: Long Ngo, 2026 — conceptual illustration.

III. PROPOSED ARCHITECTURE

The proposed architecture consists of seven layers: input acquisition, script normalization, CVNSS4.0 IR conversion, segmentation and registry matching, reference-ID assignment, vector-table lookup, and application-level computation. The input may be ordinary Vietnamese text, speech-to-text output, QR payloads, RFID records, blockchain metadata, or GIS attributes. The normalization stage applies Unicode and orthographic normalization before passing the expression into CVNSS4.0 IR.

The CVNSS4.0 IR layer is not intended to replace Quoc Ngu in human communication. It functions as a machine-oriented intermediate representation, analogous to a map projection. It makes the written form more suitable for deterministic matching, ASCII-friendly transmission, embedded systems, and payload-limited environments. After IR conversion, segmentation and registry matching determine whether a phrase, word, syllable, character, or raw payload should be mapped to an entry.

The core relation may be written as follows:

$$x \rightarrow IR(x) \rightarrow ID(x) \rightarrow E[ID(x)]$$

where x denotes the original expression, IR(x) denotes its CVNSS4.0 intermediate representation, ID(x) denotes the registry-stable identifier, and E[ID(x)] denotes the precomputed vector retrieved from the embedding table.

The lookup operation separates expensive model inference from lightweight runtime processing. A large embedding model may be used offline to produce a high-quality vector for each registry entry. At runtime, the system only receives IDs and retrieves their corresponding vectors.

TABLE I
COMPARISON BETWEEN BPE/SUBWORD TOKENIZATION AND CVNSS4.0 IR-BASED REFERENCE IDS

Criterion	BPE / Subword Tokenization	CVNSS4.0 IR -> Reference ID	Strategic Assessment
Primary nature	Statistical segmentation of strings.	Controlled mapping from Vietnamese IR to stable concept IDs.	BPE supports LLM prediction; reference IDs support data interoperability.
Processing unit	Subword, byte, or frequent symbol fragment.	Phrase, word, syllable, character, or raw fallback.	Reference IDs preserve domain concepts more directly.
Semantic commitment	No guaranteed independent meaning.	Metadata-bound: Quoc Ngu, CVNSS, domain, type, binary, hex, vector.	Suitable for audit, GIS, QR, RFID, and blockchain.
New terms	Strong, because unknown words can be split.	Requires fallback tiers and registry extension.	A four-tier design prevents data loss.
Runtime vector generation	Often requires model-dependent encoding.	Can use precomputed ID-to-vector lookup.	Lookup is lightweight and edge-friendly.
Version stability	Bound to tokenizer/model versions.	Managed by a registry and semantic reference frame.	This is the key difference for infrastructure use.
Best use cases	LLMs, translation, text generation, contextual inference.	Traceability, IoT, RFID, chipless encoding, blockchain, GIS, edge AI.	The systems are complementary rather than exclusive.

IV. LOOKUP-BASED VECTOR COMPUTATION

The central acceleration mechanism is embedding lookup. Suppose a registry contains N entries and each entry is assigned a vector of dimensionality d. The embedding table can be represented as a matrix E in R^(N x d). Given an ID i, the vector representation is retrieved by a gather operation:

$$v_i = E[i]$$

For a sequence of IDs S = {i1, i2, ..., ik}, a lightweight document or payload representation may be computed through an average, a weighted average, or an EmbeddingBag-like aggregation:

$$v_S = (1 / k) \Sigma E[i_j], \; j = 1,...,k$$

This representation can be used for classification, nearest-neighbour search, clustering, rule scoring, semantic filtering, or alert generation. When the embedding table is fixed for a registry version, downstream vector indexes such as FAISS, HNSW, or other approximate-nearest-neighbour structures can be prebuilt and cached.

Algorithm 1: Runtime Reference-ID Vectorization

```
Input: Vietnamese text or compact payload x
Output: Vector representation v_S
1: Normalize x using Unicode and orthographic rules
2: Convert x to CVNSS4.0 IR
3: Segment IR into phrase/word/syllable candidates
4: Match candidates against the semantic registry
5: If a phrase is found, emit its Reference ID
6: Else fallback to word, syllable, character, or raw encoding
7: Retrieve vectors E[i] for each emitted ID i
8: Aggregate vectors by mean, weighted mean, or task-specific pooling
9: Return v_S to classification, search, GIS, RFID, blockchain, or IoT modules
```

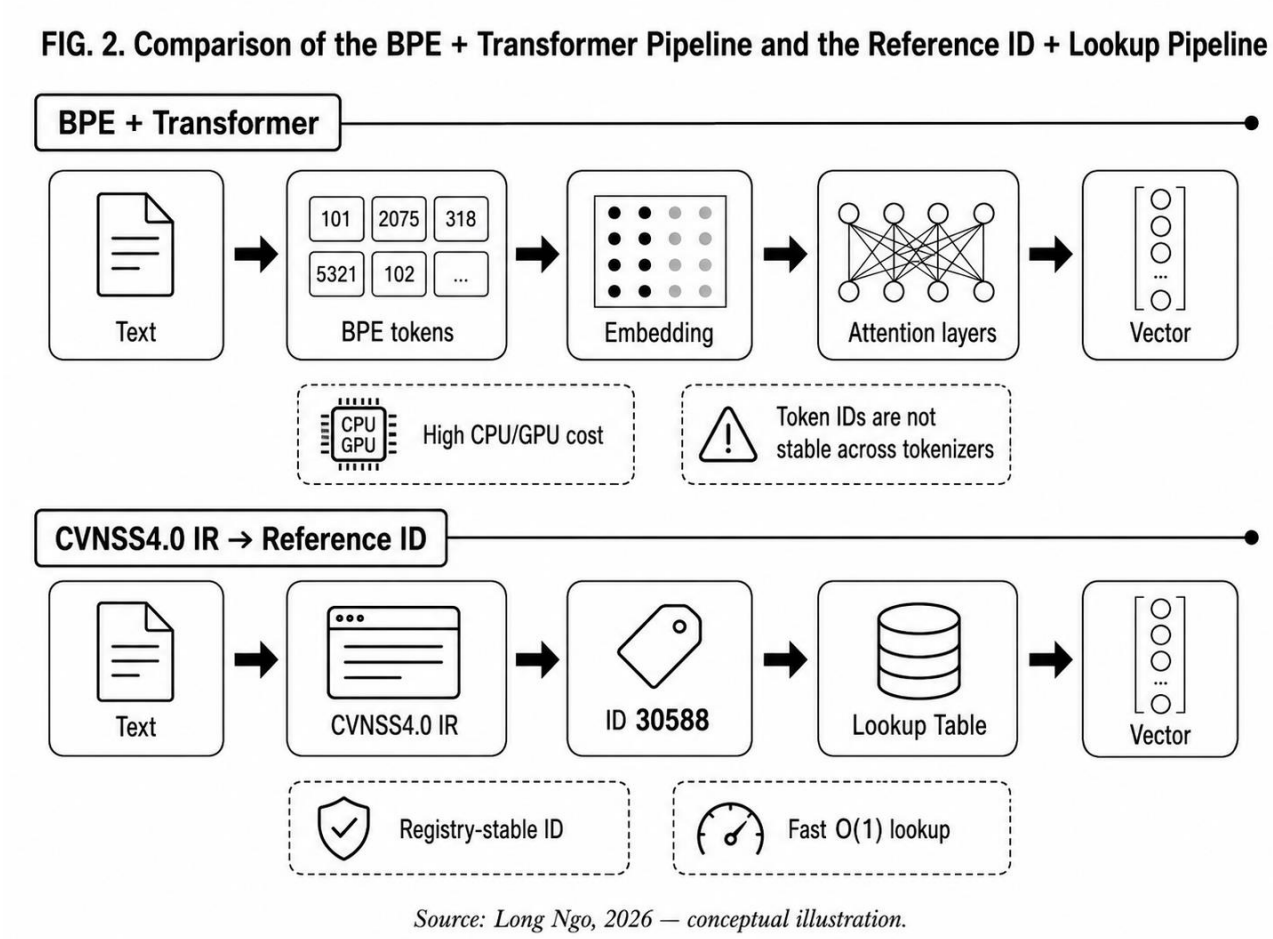

Fig. 2. Comparison between a BPE + Transformer pipeline and a CVNSS4.0 IR -> Reference ID lookup pipeline. Source: Long Ngo, 2026 — conceptual illustration.

V. SEMANTIC REGISTRY AS A REFERENCE FRAME

A registry-stable ID is only meaningful if it belongs to a well-defined reference frame. In geographic information systems, a coordinate requires a datum and projection; otherwise, the number may be ambiguous. In a semantic infrastructure, an ID requires namespace, version, lexical form, concept type, domain, vector model, vector dimensionality, and governance rules.

Therefore, the registry is more than an ID table. It is an auditable semantic reference frame. Each entry should specify whether it represents a surface word, a phrase, a concept, a device, an action, a product class, a GIS attribute, or a domain-specific relationship. The registry should also define whether an ID is active, deprecated, merged, split, or superseded.

A single ID may hold multiple vectors when different application domains require different semantic views. For example, the ID representing traceability may have a general-language vector, a supply-chain vector, an agricultural vector, and a legal-document vector. These vectors should not be mixed without recording their model, corpus, dimension, and normalization protocol.

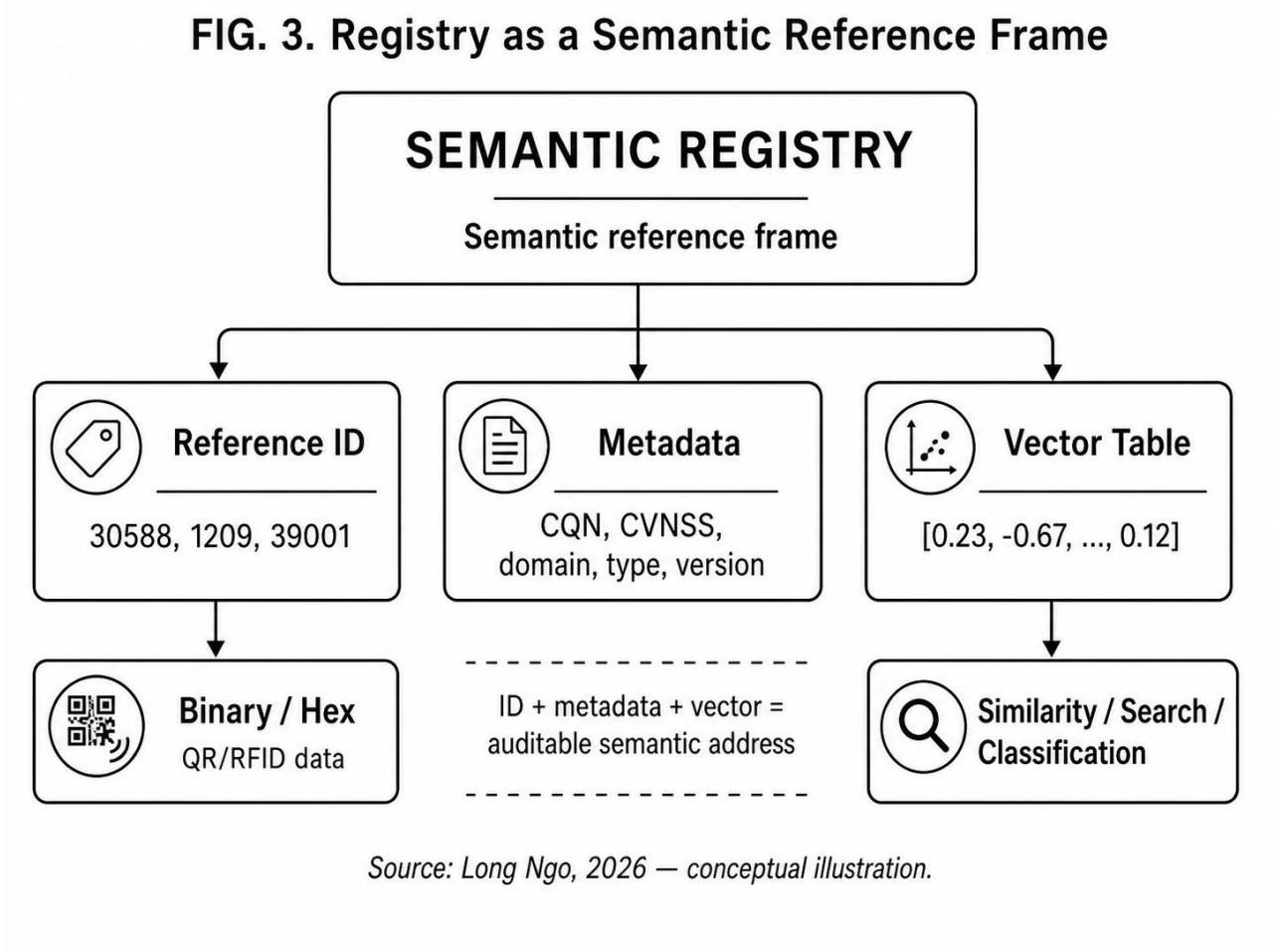

Fig. 3. Registry as a semantic reference frame integrating IDs, metadata, vector tables, QR/RFID payloads, and similarity/search/classification operations. Source: Long Ngo, 2026 — conceptual illustration.

VI. APPLICATION SCENARIOS

A. Traceability, QR/NFC/RFID, and VeChip

In a traceability scenario, a product need not store a long natural-language description. Instead, it may store a compact list of IDs such as [1201, 39001, 30588, 41012, 50021], where each ID points to a controlled registry entry: organic fish, aquaponics, traceability, Ben Tre, and quality inspection, respectively. This payload can be encoded into QR, NFC, RFID memory, or blockchain metadata.

A reader device decodes the payload, queries the local registry, retrieves vectors, and aggregates them into a product-level semantic representation. The system may then compare that representation with safety standards, certification profiles, risk categories, or consumer-facing disclosure templates. In a VeChain + RFID deployment, the registry ID layer functions as the semantic bridge between physical tags and blockchain audit trails.

B. GIS and Semantic Spatial Search

GIS objects already combine identifiers, geometry, and attributes. A reference-ID system can enrich this structure by turning attributes into controlled semantic coordinates. A parcel may be associated with IDs representing paddy land, mild salinity, canal proximity, erosion risk, or aquaponics conversion suitability. By aggregating the vectors of those IDs, the parcel receives a semantic vector in addition to its spatial geometry.

This allows semantic spatial search. A user may query for areas similar to a target condition—such as paddy land affected by mild salinity, near irrigation canals, and suitable for aquaponics—while the system retrieves candidate parcels by vector similarity and then intersects them with spatial constraints.

C. IoT and Lightweight Vietnamese Commands

Many IoT commands do not require a full language model. Commands such as “turn on pump 2”, “check pH”, or “open the irrigation valve for 15 minutes” can be represented by controlled IDs: ACTION_ON, DEVICE_PUMP, NUMBER_2,

SENSOR_PH, or VALVE_OPEN. A rule engine or a small classifier may operate over these IDs and vectors on an edge device, even without cloud connectivity.

The key advantage is that the user-facing language remains natural, while the machine-facing representation is compact, stable, and auditable. This is especially relevant for agricultural IoT, aquaponics monitoring, rural edge devices, and low-bandwidth contexts.

D. Chipless RFID and Physical Encoding

A stable reference ID can be converted into binary or hexadecimal form and then mapped to physical encodings such as frequency notches in chipless RFID. For example, ID 30588 may correspond to the binary pattern 0111011101111100, which can be mapped to a notch pattern in a radio-frequency scattering spectrum. This physicalization of semantic IDs is difficult to achieve with model-internal BPE tokens, but natural for registry-stable IDs.

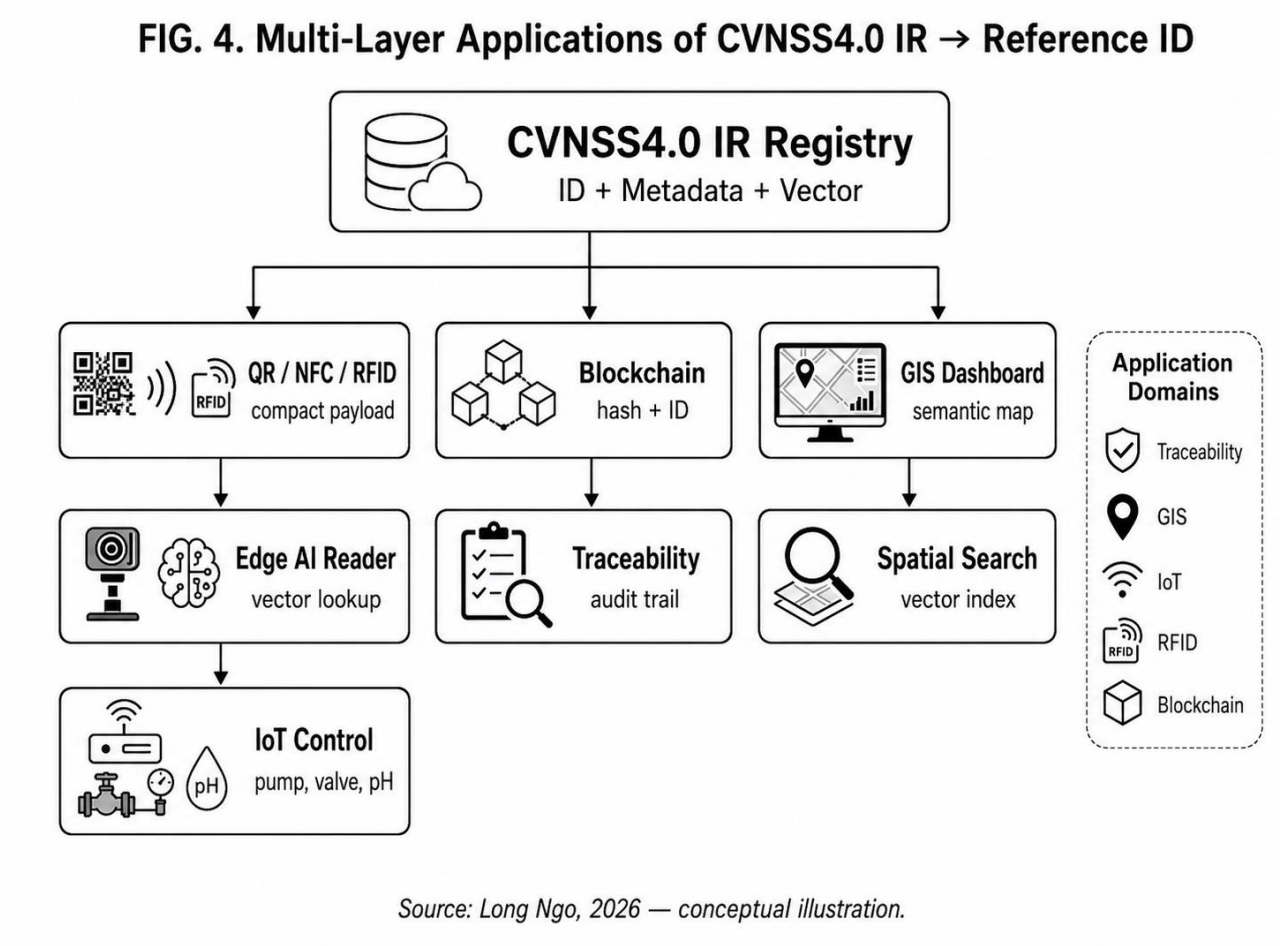

Fig. 4. Multi-layer applications of CVNSS4.0 IR -> Reference ID across QR/NFC/RFID, blockchain, GIS, edge-AI readers, traceability, spatial search, and IoT control. Source: Long Ngo, 2026 — conceptual illustration.

VII. DISCUSSION

The proposed architecture does not attempt to replace BPE or Transformer-based models. Instead, it assigns them a different role. BPE and Transformer models remain appropriate for open-ended language understanding, generation, translation, summarization, and contextual reasoning. The reference-ID system is better suited for infrastructure tasks requiring stable identifiers, compact payloads, deterministic decoding, and cross-device interoperability.

The principal engineering challenge is registry governance. Vietnamese contains homographs, polysemous words, domain-specific terms, administrative variations, and new technical expressions. A robust registry must distinguish surface word IDs from concept IDs. For example, the Vietnamese word for “road”, “sugar”, or “line” may require multiple concept entries depending on domain. CVNSS4.0 IR normalizes the written form, but semantic disambiguation requires domain metadata and contextual matching.

Another challenge is vector versioning. A vector is meaningful only relative to the embedding model that produced it. Just as spatial coordinates require a datum, semantic vectors require a model name, training corpus, dimension, normalization rule, and

registry version. A professional deployment should avoid mixing incompatible vector spaces without transformation or metadata control.

VIII. CONCLUSION

This paper presented a conceptual IEEE-style framework for treating CVNSS4.0 IR-based reference IDs as semantic coordinates for Vietnamese digital infrastructure. The central claim is that a stable ID can serve as a discrete semantic coordinate, while a precomputed embedding vector attached to that ID can serve as a continuous coordinate in semantic space. This architecture enables the system to separate expensive vector generation from lightweight runtime vector use.

Compared with a pure BPE + Transformer pipeline, a registry-based reference-ID approach enables compact payloads, deterministic lookup, auditable semantics, persistent vector indexes, and edge-friendly computation. Its strongest applications are not general text generation but infrastructure-level interoperability: traceability, QR/NFC/RFID systems, chipless RFID encoding, blockchain audit trails, GIS semantic search, IoT control, and low-latency edge AI.

The broader implication is that Vietnamese digital infrastructure can benefit from a semantic reference frame analogous to geographic coordinates. Quoc Ngu remains the human-facing script; CVNSS4.0 IR functions as the machine-oriented projection; reference IDs become stable semantic addresses; vector tables enable fast computation; and physical-digital systems become capable of sharing a common semantic coordinate system.

REFERENCES

[1] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in Proc. 54th Annu. Meeting Assoc. Comput. Linguistics, Berlin, Germany, 2016, pp. 1715–1725.

[2] T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” in Proc. EMNLP: System Demonstrations, 2018, pp. 66–71.

[3] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv:1301.3781, 2013.

[5] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in Proc. EMNLP, 2014, pp. 1532–1543.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[7] D. Q. Nguyen and A. T. Nguyen, “PhoBERT: Pre-trained language models for Vietnamese,” in Findings of ACL: EMNLP, 2020, pp. 1037–1042.

[8] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” IEEE Trans. Big Data, vol. 7, no. 3, pp. 535–547, 2021.

[9] Y. A. Malkov and D. A. Yashunin, “Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 4, pp. 824–836, 2020.

[10] K. T. Lam and T. T. Binh, “Chu VN Song Song 4.0,” Vietnam copyright registration no. 1850/2020/QTG, 2020.

[11] L. Ngo, “CVNSS4.0 IR-based reference IDs and precomputed semantic vectors for Vietnamese digital infrastructure,” unpublished conceptual manuscript, 2026.