Injection Without Distortion: Geometrically Constrained Knowledge Enhancement For Vision-Language Models

Zhongze Wu[1], Xiu Su[1*], Feng Yang[2], Shan You[3], Yueyi Luo[1], Jun Long[1*]

[1]Central South University, Big Data Institute, Changsha, China
[2]Southeast University, Nanjing, China
[3]SenseTime Research, Shanghai, China

Abstract

Vision-Language Models (VLMs) are widely used in tasks like Open-Vocabulary Object Detection and zero-shot Classification, owing to their powerful generalization. However, recent research reveals that VLMs exhibit significant performance instability when tasked with recognizing concepts at varying granularities (*e.g.*, "animal" *vs.* "dog"). Prevailing methods inject external knowledge from Large Language Models, but this unconstrained approach distorts the VLM's inherent hierarchical orthogonal geometry, leading to performance collapse on general concepts. To address this, we introduce ***GeCoin***, an innovative ***Ge***ometrically ***Co***nstra***in***ed framework that safely enhances existing VLMs with external knowledge for improved hierarchical understanding, without additional training. By projecting knowledge into the null-space of a query concept's feature space, *GeCoin* mathematically guarantees the preservation of general knowledge while integrating specialized information. Extensive experiments across large-scale benchmarks, diverse VLMs, and knowledge from various LLMs (*e.g.*, GPT-3.5, Claude-3, Gemini-Pro) show that *GeCoin* boosts performance by an average of 3.9% over the strongest baseline—crucially eradicating performance collapse on general concepts.

Introduction

Vision-Language Models (VLMs) (Radford et al. 2021; Zhai et al. 2023; Tschannen et al. 2025) are widely used in foundational tasks like Open-Vocabulary Object Detection (OvOD) (Zhou et al. 2022) due to their powerful generalization capabilities. OvOD leverages VLMs to detect a virtually unlimited range of objects from textual prompts (Zareian et al. 2021; Lin et al. 2022; Liu et al. 2024b). However, VLMs exhibit significant performance instability when confronted with concepts at varying levels of semantic granularity (*e.g.*, "animal" vs. "dog") (Liu et al. 2024a; Novack et al. 2023). To address this, the predominant research paradigm has focused on utilizing *external knowledge injection*, which aims to enrich the model's hierarchical semantic understanding by incorporating knowledge from external sources generated by Large Language Models (LLMs) (Brown et al. 2020; Team et al. 2024; Zhang et al. 2025).

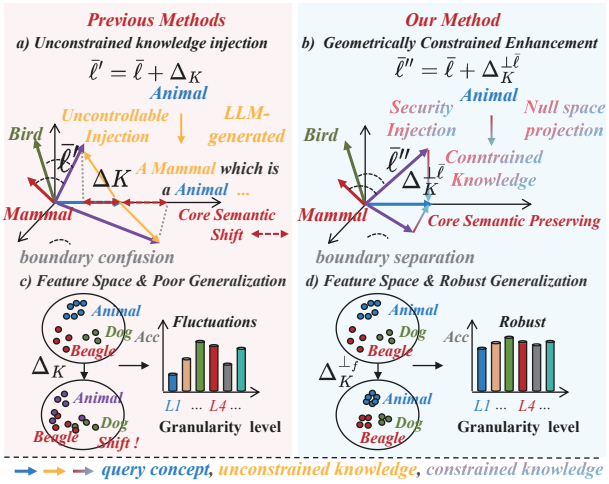
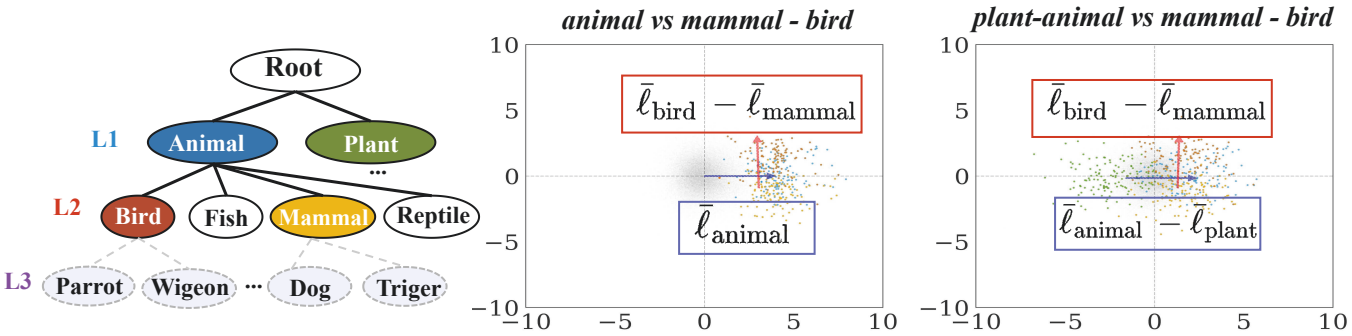


Figure 1: Comparison between previous methods and our approach. **(a) Previous Methods**: Unconstrained injection causes *semantic shift* and *boundary confusion*. **(b) Our Method**: GeCoin applies *null space projection* for *semantic preservation* and *boundary separation*. **(c)** Unconstrained injection causes feature space *shift* and performance *fluctuations*. **(d)** Our approach maintains stable clusters and delivers *robust* performance.

The dominant approach within the external knowledge injection paradigm is what we term *unconstrained knowledge injection*, as illustrated in Figure 1(a). This framework first generates richer hierarchical or fine-grained semantic information for query concepts $\bar{\ell}$ using external knowledge sources, which have evolved from manually-curated hierarchies like WordNet (Fellbaum 1998) to the now-prevalent LLMs (Yao et al. 2024; Fu et al. 2025). Essentially, based on decomposable embedding principles (Trager et al. 2023), the original feature $\bar{\ell}$ is updated by adding Δ_K, a knowledge injection, to produce the new feature $\bar{\ell}'$. While appears intuitive, unconstrained injection creates a reverse phenomenon: it impairs the generalization capability on coarse-grained concepts (*e.g.*, L1 granularity), leading to unpredictable recognition results (Figure 1(c)) (Liu et al. 2024a).

To investigate this, our systematic analysis reveals that modern VLMs encode hierarchical concepts (Figure 2(a))

*These authors are the corresponding authors.



(a) Hierarchically related concepts **(b) Hierarchy is encoded as orthogonality in CLIP embedding space**

Figure 2: Hierarchical concepts in VLMs exhibit geometric orthogonality. (a) illustrates the conceptual hierarchy. (b) demonstrates that this hierarchy is encoded as orthogonality in CLIP's embedding space (ViT-B/32). The 2D projections show cross-specificity orthogonality relationships: $\text{span}\{\bar{\ell}_{\texttt{animal}}, \bar{\ell}_{\texttt{bird}} - \bar{\ell}_{\texttt{mammal}}\}, \text{span}\{\bar{\ell}_{\texttt{animal}} - \bar{\ell}_{\texttt{plant}}, \bar{\ell}_{\texttt{bird}} - \bar{\ell}_{\texttt{mammal}}\}$. Gray points represent the full vocabulary distribution. Blue vectors represent parent concept axes (e.g., *animal*), while red vectors represent child-concept difference vectors (e.g., *bird* − *mammal*). Colored points indicate specific concept categories.

with a consistent geometric structure. Specifically, the feature vector for a query concept is approximately orthogonal to the difference vector between its sub-concepts, as shown in Figure 2(b). However, the unconstrained knowledge injection Δ_K used by existing methods pulls the query concept's (*e.g.*, "animal") semantic representation $\bar{\ell}$ toward the distribution of the injected knowledge Δ_K (*e.g.*, "mammal"). This semantic shift inevitably distorts the hierarchical orthogonality that VLMs have carefully learned, thereby blurring cluster boundary and degrading their generalization capability (Figure 1(c)). The problem is further exacerbated when knowledge is sourced from LLMs, whose generated content often contains hallucinations (Brown et al. 2020).

To address these flaws, we reframe the problem from simple knowledge injection to **Ge**ometrically **Co**nstrained Knowledge **In**jection, introducing our method, **GeCoin**. Instead of performing a direct update, GeCoin purifies the injected knowledge Δ_K by projecting it onto the orthogonal null-space (Wang et al. 2021; Fang et al. 2024) of the original concept feature $\bar{\ell}$, yielding a pristine knowledge component $\Delta_K^{\perp\bar{\ell}}$. As shown in Figure 1(b), the feature is then updated to $\bar{\ell}' = \bar{\ell} + \Delta_K^{\perp\bar{\ell}}$. By leveraging the mathematical properties of null-space projection, our method guarantees that specialized information is added without altering the component of Δ_K in the direction of $\bar{\ell}$, thus preserving its core semantic. This ensures distributional invariance, fundamentally preventing the semantic *Shift* and producing a *Robust Concept* stable across all granularities (Figure 1(d)).

To validate the effectiveness of our method, we conducted extensive experiments on challenging hierarchical detection benchmarks (iNatLoc (Cole et al. 2022), FSOD (Fan et al. 2020)) and standard OvOD benchmarks (COCO (Lin et al. 2014), LVIS (Gupta, Dollar, and Girshick 2019)). The results compellingly demonstrate that GeCoin not only eradicates the performance collapse observed in prior works at the general concept level (L1) but also achieves over a 3.9% average improvement in fine-grained accuracy over the strongest baseline. Furthermore, we validate GeCoin's robustness against hallucinatory knowledge from various LLMs and its plug-and-play compatibility with different VLMs. GeCoin offers a safe, reliable paradigm for knowledge injection, underscoring the importance of preserving geometric integrity when using external, unreliable knowledge to expand the semantic capacity of multimodal models. Our contributions are summarized as follows:

1. We reveal that VLMs encode hierarchical concepts through *hierarchical orthogonality*—where general concepts are orthogonal to difference vectors between sub-concepts, providing a theoretical foundation for understanding VLM's geometric structure.
2. We propose **GeCoin**, a **Ge**ometrically **Co**nstrained Knowledge **In**jection framework that preserves VLM's intrinsic geometric structure during external knowledge integration through null-space projection, fundamentally preventing semantic drift.
3. We design adaptive null-space projection mechanisms that handle open-vocabulary scenarios while effectively mitigating LLM hallucination effects, ensuring robust knowledge integration across different paradigms.
4. We conduct extensive experiments across large-scale benchmarks with diverse VLMs (*e.g.*, CLIP, SigLip 2) and knowledge from various LLMs (*e.g.*, GPT-3.5, Claude-3, Gemini-Pro), demonstrating that GeCoin achieves 3.9% average improvement over baselines.

Methodology

The Geometric Representations of VLM

The remarkable capabilities of VLMs stem from their ability to map multimodal information into a shared embedding space $\mathcal{S} \subset \mathbb{R}^d$. Foundational research reveals that this space adheres to a strict ***linear compositionality*** (Fellbaum 1998; Trager et al. 2023; Saglam et al. 2025), where the representation of a complex concept can be precisely decomposed into a vector sum of its constituent semantic factors. A formal definition is stated as follows:

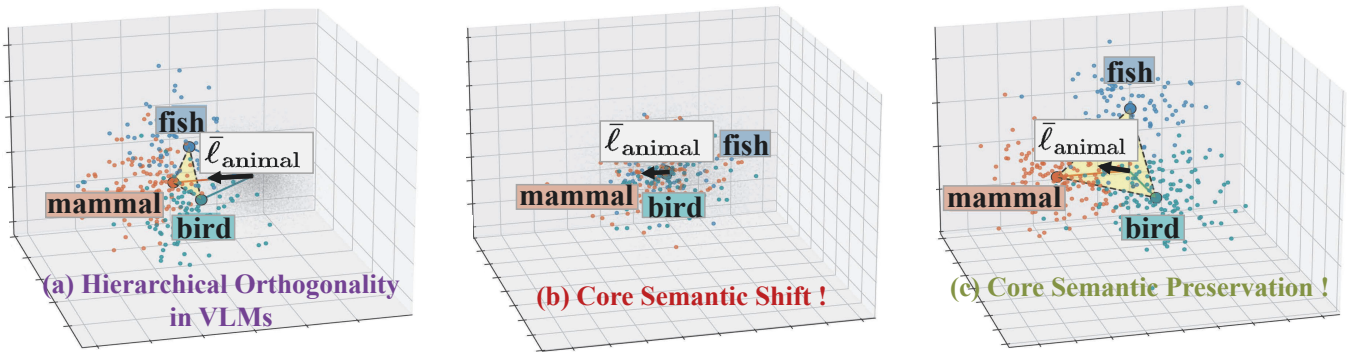


Figure 3: Feature space evolution from original representations using CLIP ViT-B/32 to unconstrained knowledge injection and our geometrically constrained approach. **(a)**: Original hierarchical feature space where concepts maintain clear geometric orthogonality, preserving natural cluster boundaries. **(b)**: Unconstrained external knowledge injection from LLMs causes unpredictable feature drift and cluster boundary contamination, disrupting the original semantic structure. **(c)**: Our GeCoin method maintains distinct cluster boundaries between hierarchical concepts, preserving both *Generality* (semantic anchoring) and *Specificity* (sub-concept differentiation) while safely integrating external knowledge through null-space projection.

Definition 1 (Decomposable embeddings)**.** An embedding $\bar{\ell}_z$ of a concept z, parameterized by multiple factors $\mathcal{Z} = \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_k$, is decomposable if there exists a global offset vector $\bar{\ell}_0$ and "ideal word" vectors $\bar{\ell}_{z_i}$ corresponding to each factor value, such that:

$$\bar{\ell}_z = \bar{\ell}_0 + \sum_{i=1}^{k} \bar{\ell}_{z_i} \quad \text{s.t.} \quad \forall i, \sum_{z_i \in \mathcal{Z}_i} \bar{\ell}_{z_i} = \mathbf{0}. \tag{0.1}$$

The centering condition ensures the uniqueness of the decomposition. This linear structure exhibits an even more elegant property when handling hierarchical concepts (e.g., "animal" vs. "dog"). The following theorem from (Park et al. 2024) formalizes this geometric relationship within LLMs.

Theorem 2 (Hierarchical Orthogonality in LLMs)**.** *Suppose there exist vector representations for hierarchical concepts. Let q represent a query concept (e.g., "animal") and subordinate concepts. The embedding space exhibits two fundamental orthogonality properties:*

(a) General-Specific Orthogonality: $\bar{\ell}_q \perp (\bar{\ell}_{s_1} - \bar{\ell}_{s_2})$ for any s_1, s_2 such that $s_1 \prec q$ and $s_2 \prec q$.

(b) Cross-Hierarchy Orthogonality: $(\bar{\ell}_{q_1} - \bar{\ell}_{q_0}) \perp (\bar{\ell}_{s_1} - \bar{\ell}_{s_0})$ for any four concepts where (q_0, s_0) and (q_1, s_1) form parallel parent-child pairs, i.e., $s_0 \prec q_0$ and $s_1 \prec q_1$.

The Theorem 2 reveals that the vector for a general concept, such as "animal", is orthogonal to the difference vector between its child concepts, such as "bird" and "mammal". As shown in Figure 2(b), we find that VLMs exhibit a similar hierarchical orthogonality to LLMs (see Supplementary Material for detailed experiments and proof).

Unconstrained Knowledge Injection

Despite possessing this inherent geometric structure, VLMs still struggle with performance instability across semantic granularities. To address this, a predominant paradigm is **external knowledge injection**. This approach leverages richer semantic information generated by LLMs to enhance VLM representations. Based on the **linear decomposition** principle of VLM feature spaces, essentially, the new semantic information ($\bar{\ell}'_q$) is injected by adding an information increment (Δ_K) to the original concept feature ($\bar{\ell}_q$), aiming to obtain more granular and generalized representations :

$$\bar{\ell}'_q = \bar{\ell}_q + \Delta_K. \tag{0.2}$$

However, while theoretically promising, this unconstrained method often distorts the underlying geometric structure of query concepts (Proof see Supplementary Material). As our experiments show, this leads to a feature space *Shift* (Figure 3) and degrades generalization, especially on broader concepts (Table 1).

Problem Formulation The core issue with the unconstrained knowledge injection paradigm is what we term **Geometric Pollution** (Fig. 3(b)). This problem arises from the component of the knowledge increment Δ_K that is not orthogonal to the original feature $\bar{\ell}_{q,L_i}$ at a given granularity level L_i. We quantify this pollution component as:

$$\Delta_K^{\text{pollute}} = \text{Proj}_{\bar{\ell}_{q,L_i}}(\Delta_K) = \frac{\langle \Delta_K, \bar{\ell}_{q,L_i} \rangle}{\|\bar{\ell}_{q,L_i}\|^2} \bar{\ell}_{q,L_i}. \tag{0.3}$$

This pollution term shifts the concept's semantic anchor, violating the hierarchical orthogonality principle. This violation compromises cross-granularity consistency, leading to performance degradation in detection and classification, particularly at broader levels (smaller L_i) that act as shared semantic anchors. Noisy sourced Δ_K exacerbates this issue.

Geometrically Constrained Knowledge Injection

To address geometric pollution, we introduce GeCoin, a training-free framework for safely integrating external knowledge while preserving the VLM's intrinsic structure through null-space projection.

The Principle of Null-Space Projection The concept of a null-space is core to our work. Formally, a matrix $\mathbf{B}$ is

Figure 4: Comparison of detection results between Ground-truth and GeCoin on the iNatLoc dataset. The figure shows that under query terms spanning different semantic granularities (*e.g.*, L1 - L6), GeCoin consistently produces correct predictions. This stability reflects the effectiveness of our approach in generating bias-reduced, granularity-invariant features.

in the null-space of a matrix $\mathbf{A}$ if and only if $\mathbf{BA} = \mathbf{0}$ (Wang et al. 2021). In our context, this means ensuring the knowledge increment Δ_K is projected into the null-space of the query concept's feature $\bar{\ell}_q$, which we denote as $\Delta_K^{\perp \bar{\ell}_q}$. This projection guarantees that the pollution term $\Delta_K^{\text{pollute}} = \text{Proj}_{\bar{\ell}_q}(\Delta_K)$ is completely removed.

Consequently, the update process preserves the geometric integrity of the original concept. This implies that the projected knowledge increment $\Delta_K^{\perp \bar{\ell}_q}$ will not disrupt the core semantic geometry of the preserved query concept. The updated feature $\bar{\ell}_q'' = \bar{\ell}_q + \Delta_K^{\perp \bar{\ell}_q}$ thus satisfies:

$$\langle \bar{\ell}_q'', \bar{\ell}_q \rangle = \langle \bar{\ell}_q + \Delta_K^{\perp \bar{\ell}_q}, \bar{\ell}_q \rangle = \langle \bar{\ell}_q, \bar{\ell}_q \rangle + \langle \Delta_K^{\perp \bar{\ell}_q}, \bar{\ell}_q \rangle = \|\bar{\ell}_q\|^2, \tag{0.4}$$

where the term $\langle \Delta_K^{\perp \bar{\ell}_q}, \bar{\ell}_q \rangle = 0$ by definition of the null-space projection. This mechanism is how GeCoin preserves a concept's core semantic anchor, preventing the feature drift that causes generalization failure in other methods.

The GeCoin Framework. Our framework is designed for both scenarios with known hierarchies and truly open-vocabulary concepts. GeCoin operationalizes this principle through a new two-step process:

Null-Space Projector Constructing. To project knowledge into the null-space, we first define the "preserved knowledge" that must remain invariant. Our framework adapts to the nature of the query concept q:

Hierarchical Concept (with sub-concepts). If the query concept q has a set of known sub-concepts $\{c_{s_1}, c_{s_2}, \ldots, c_{s_m}\}$ (e.g., from a knowledge base or LLM generated), its semantic identity is fundamentally supported by the collective space formed by itself and its subordinate concepts. Therefore, to protect this entire semantic structure, we form a preserved knowledge matrix $\mathbf{K}_0 = [\bar{\ell}_q, \bar{\ell}_{c_{s_1}}, \ldots, \bar{\ell}_{c_{s_m}}]$. We then apply Singular Value Decomposition (SVD):

$$\{\mathbf{U}, \mathbf{\Lambda}, \mathbf{V}^\top\} = \text{SVD}(\mathbf{K}_0). \tag{0.5}$$

We consider singular values below $\varepsilon = 10^{-2}$ as zero to form $\hat{\mathbf{V}}$. The null-space is spanned by the columns in $\mathbf{V}$ corresponding to zero singular values. We form a submatrix $\hat{\mathbf{V}}$ from these columns to construct the projector:

$$\mathbf{P}_q = \hat{\mathbf{V}}\hat{\mathbf{V}}^\top. \tag{0.6}$$

Leaf Concept. For truly novel, open-vocabulary concepts where no hierarchy is available a priori, the system defaults to treating q as a leaf node. The preserved knowledge collapses to its own feature vector, $\mathbf{K}_0 = \bar{\ell}_q$. The projector then has a direct closed-form solution:

$$\mathbf{P}_q = \mathbf{I} - \frac{\bar{\ell}_q \bar{\ell}_q^\top}{\|\bar{\ell}_q\|^2}. \tag{0.7}$$

This projector, $\mathbf{P}_q$, serves as the fundamental tool for all subsequent constrained operations.

Knowledge Injection with Null-Space Projection. To implement knowledge injection, we explore two modern paradigms for leveraging LLMs:

Hierarchical Knowledge Injection (HKI). This approach (Liu et al. 2024a) uses an LLM to first discover parent and child concepts for a query q. Then, these relationships are structured into a set of templated sentences, e.g., *"a {child}, which is a {query}, which is a {parent}."* This produces a set of K sentences $\{\mathbf{s}_1, \ldots, \mathbf{s}_K\}$ that explicitly encode the hierarchy.

Fine-Grained Knowledge Injection (FKI). This approach (Jin et al. 2024) prompts an LLM to generate a rich, fine-grained paragraph detailing the unique visual and semantic characteristics of the query concept q.

The embeddings of the N external knowledge texts e^q (from either paradigm), encoded by $\mathcal{E}_\text{T}$, are averaged to form a consolidated external knowledge representation:

$$\mathbf{K}_{ext} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{E}_\text{T}(e_i^q). \tag{0.8}$$

Then, the projection isolates this knowledge representation to the null-space of the original feature, ensuring the update does not alter the core semantic direction, thus making the injection robust to factual inaccuracies or noise from the external knowledge source:

$$\mathbf{K}_{ext}^\perp = \mathbf{P}_q \mathbf{K}_{ext}. \tag{0.9}$$

Finally, the purified knowledge is injected into the $\bar{\ell}_q$:

$$\bar{\ell}_q'' = \bar{\ell}_q + \mathbf{K}_{ext}^\perp. \tag{0.10}$$

Dataset	Gran	Level	ResNet-50 Backbone						Swin-B Backbone					
			I - LVIS			II - LVIS + IN-L			III - LVIS + IN-21k			IV - LVIS & COCO + IN-21k		
			Detic	SHiNe	GeCoin	Detic	SHiNe	GeCoin	Detic	SHiNe	GeCoin	Detic	SHiNe	GeCoin
iNatLoc	F ↓ C	L6	32.0	52.8	**54.7** (+1.9)	35.2	58.3	**60.7** (+2.4)	58.6	84.5	**86.0** (+1.5)	60.2	82.7	**85.0** (+2.3)
		L5	28.2	41.1	**48.9** (+7.8)	30.3	46.6	**53.7** (+7.1)	54.9	76.3	**80.9** (+4.6)	57.5	76.1	**81.0** (+4.9)
		L4	40.1	50.4	**53.9** (+3.5)	43.4	57.5	**60.8** (+3.3)	73.1	84.0	**86.8** (+2.8)	74.9	83.4	**86.4** (+3.0)
		L3	38.8	57.2	**59.3** (+2.1)	41.6	61.7	**63.3** (+1.6)	63.8	83.6	**85.3** (+1.7)	67.2	81.7	**83.7** (+2.0)
		L2	34.4	43.9	**49.5** (+5.6)	39.3	50.5	**54.7** (+4.2)	65.3	77.2	**78.2** (+1.0)	67.2	74.5	**75.6** (+1.1)
		L1	31.6	33.5	**41.9** (+8.4)	32.5	36.9	**43.8** (+6.9)	65.4	63.8 (-1.6)	**72.1** (+8.3)	64.4	62.1 (-2.3)	**67.5** (+5.4)
		Average	34.2	46.5	**51.4** (+4.9)	37.1	51.9	**56.2** (+4.3)	63.5	78.2	**81.6** (+3.4)	65.2	76.8	**79.9** (+3.1)
FSOD	F ↓ C	L3	49.7	52.2	**54.0** (+1.8)	51.9	53.7	**55.1** (+1.4)	66.0	66.3	**66.6** (+0.3)	65.6	66.4	**66.5** (+0.1)
		L2	28.2	30.9	**34.9** (+4.0)	27.8	29.8	**32.8** (+3.0)	38.4	40.3	**40.7** (+0.4)	39.4	41.5	**42.1** (+0.6)
		L1	16.0	22.0	**27.0** (+5.0)	16.5	21.0	26.0 (+5.0)	24.7	30.2	**35.4** (+5.2)	25.0	29.6	**34.1** (+4.5)
		Average	31.3	35.0	**38.6** (+3.6)	32.1	34.8	**38.0** (+3.2)	43.0	45.6	**47.6** (+2.0)	43.3	45.8	**47.6** (+1.8)

Table 1: Zero-shot detection performance on the iNatLoc and FSOD datasets. We compare the performance of Detic, SHiNe and GeCoin. Results are reported for each granularity level, ranging from the finest (F) to the coarsest (C), with mAP50 (%) for each level and four types of supervisory signal combinations (*e.g.*, I - LVIS etc). Bold values indicate the best performance.

Level	# Classes	CLIP	H-CLIP	CHiLS	SHiNe	UnSec
L1	10	56.2	67.9	**73.8**	50.4	61.3
L2	29	56.8	69.3	67.2	60.9	67.9
L3	128	43.3	62.4	62.2	54.7	59.8
L4	466	55.2	69.6	70.1	70.3	**71.1**
L5	591	62.4	65.9	64.5	69.1	**71.0**
L6	98	73.1	75.4	73.5	78.9	**80.5**
FPS		152	3	28	150	146
Average		57.8	68.4	68.5	64.1	**68.6**

Table 2: Zero-shot classification performance on ImageNet-1k (Deng et al. 2009) using the BREEDS structure (Santurkar, Tsipras, and Madry 2020) across different granularity levels. All methods utilize CLIP ViT-B/16, with Top-1 accuracy (%) reported for six label granularity levels.

These enhanced features $\bar{\ell}''_q$ are computed offline and cached for downstream tasks. The null-space projection, which guarantees orthogonality, serves a controlled role by ensuring the knowledge is injected only in directions that do not pollute the preserved semantics of the original concept, making it a safe and effective paradigm for semantic expansion.

Experiments

In this section, we conduct extensive experiments to address the following research questions (RQ):

- **RQ1 (Effectiveness):** Can GeCoin effectively enhance fine-grained recognition while preventing performance degradation on general concepts, thereby surpassing current SOTA methods?
- **RQ2 (Generality):** Can GeCoin's advantages generalize across different knowledge types (hierarchical vs. fine-grained), diverse base detectors, and various downstream tasks (detection vs. classification)?

Dataset	Level	ResNet-50 Backbone			Swin-T Backbone		
		CoDet	SHiNe	GeCoin	LLM-Det	SHiNe	GeCoin
iNatLoc	L6	31.2	54.8	**57.8**	61.0	66.0	**67.5**
	L5	22.5	35.5	**39.0**	49.0	54.0	**55.5**
	L4	21.4	35.8	**39.8**	60.0	62.0	**63.8**
	L3	25.8	43.3	**47.0**	64.0	66.0	**67.9**
	L2	19.7	24.6	**28.6**	53.0	54.5	**56.5**
	L1	17.7	12.6	**19.7**	39.0	37.5	**43.5**
	Average	23.1	34.4	**38.6**	54.3	56.7	**59.1**
FSOD	L3	60.5	61.6	**64.6**	54.8	58.3	**59.2**
	L2	33.5	36.6	**40.1**	26.3	29.8	**33.9**
	L1	19.9	25.4	**29.9**	15.2	18.7	**25.6**
	Average	38.0	41.2	**44.9**	32.1	35.6	**39.6**

Table 3: Comparison of different methods on iNatLoc and FSOD datasets using CoDet (Ma et al. 2024), SHiNe (Liu et al. 2024a), GeCoin and LLM-Det (Fu et al. 2025) with ResNet-50 and Swin-T backbones. Our method is directly applied to the pre-trained CoDet and LLM-Det OvOD detectors. mAP50 (%) is reported.

- **RQ3 (Ablation):** What are the individual contributions of GeCoin's core components, particularly the null-space projection and the hierarchical protection strategy?
- **RQ4 (Robustness):** How robust is GeCoin to external knowledge of varying quality, such as outputs from different LLMs, human-annotated ground-truth hierarchies, and irrelevant vocabulary?

Datasets and Protocols. We evaluate our approach on benchmarks for downstream tasks where VLMs are widely used. This includes hierarchical detection benchmarks iNatLoc (Cole et al. 2022) and FSOD (Fan et al. 2020), and zero-

Level	Detic	Unconstrained Injection		Ours (GeCoin)	
		w/HKI	w/FKI	w/HKI	w/FKI
L6 †	60.2	82.7	**85.5**	85.0	**85.8**
L5	57.5	76.1	**76.5**	81.0	**81.2**
L4	74.9	83.4	**84.6**	86.4	**86.5**
L3	67.2	**81.7**	80.7	**83.7**	83.5
L2	67.2	**74.5**	72.0	**75.6**	75.0
L1 †	**64.4**	62.1 (-2.3)	60.6 (-3.8)	**67.5** (+5.4)	65.0 (+4.4)
Avg.	65.2	76.8	76.6	**79.9** (+3.1)	79.5 (+2.9)

Table 4: Ablation study on iNatLoc comparing unconstrained knowledge injection with our final GeCoin method. All experiments use the Swin-B backbone under the "IV - LVIS & COCO + IN-21k" setting. mAP50 (%) is reported.

shot classification on ImageNet-1k (Deng et al. 2009). We follow the protocol in (Liu et al. 2024a), using mAP50 for iNatLoc/FSOD and Top-1 accuracy for ImageNet-1k.

Baselines. We apply GeCoin to various baselines including Detic (Zhou et al. 2022), CoDet (Ma et al. 2024), SHiNe (Liu et al. 2024a), and the SOTA LLM-Det (Fu et al. 2025).

Implementation Details. Our experiments are primarily conducted using Swin-B (Liu et al. 2021) and ResNet-50 (He et al. 2016) backbones paired with a CLIP ViT-B/32 (Radford et al. 2021) text encoder. HKI default to 10 sub & parent nodes, FKI default to 1 sentence. Following (Liu et al. 2024a), all label embeddings and projection matrices $\mathbf{P}_q$ are computed once offline and cached for reuse, requiring no additional training.

Concept Generalizability Preservation (RQ1)

We evaluate GeCoin against SOTA methods on hierarchical datasets. Table 1 presents results across four supervisory signal combinations. Our observations:

- **Obs 1: GeCoin eliminates performance collapse on broader concepts.** On iNatLoc, GeCoin eradicates the performance degradation on general concepts (e.g., L1) caused by unconstrained knowledge injection (-2.3%), and improves L1 performance to 67.5% (+5.4%).
- **Obs 2: GeCoin achieves superior performance across all granularity levels.** On iNatLoc and FSOD, GeCoin achieves average improvements of 3.1% and 1.8% over SHiNe across all levels. This consistent superiority stems from GeCoin's safely integrating external knowledge, ensuring stable representations at every granularity.

Cross-domain Generality and Detector Compatibility Analysis (RQ2)

We test GeCoin across zero-shot classification tasks (Table 2), base detector architectures (Table 3) and knowledge injection paradigms (Table 4). Key observations:

- **Obs 4: GeCoin's benefits extend to zero-shot classification with high efficiency.** On ImageNet-1k classification (Table 2), GeCoin achieves a competitive 68.6% average accuracy while maintaining 146 FPS, far exceeding expensive alternatives like H-CLIP (3 FPS). It shows particular strength on general concepts (L1: 61.3% vs. SHiNe's 50.4%) while remaining competitive on fine-grained ones.
- **Obs 5: GeCoin provides substantial gains across diverse detector architectures.** Applied to CoDet and LLM-Det (Table 3), GeCoin delivers average mAP50 gains of +1.7% and +3.7% on iNatLoc, and +3.5% and +4.0% on FSOD, respectively. Notably, it consistently resolves the L1 performance collapse seen in SHiNe (-5.1% with CoDet), converting it to a significant gain (+7.1% with CoDet, +8.9% with LLM-Det).
- **Obs 6: GeCoin demonstrates robust superiority across knowledge paradigms.** Table 4 shows GeCoin outperforms unconstrained injection with both HKI (+3.1%) and FKI (+2.9%), showing adaptability to diverse knowledge sources. Crucially, it eradicates the L1 performance drop of unconstrained methods (e.g., HKI: +4.4% vs. -3.8%), proving its robustness by preserving geometric integrity.

Level	Detic	w/o Proj.	w/ Case 2	GeCoin (Case 1)
L6	60.2	82.7	84.1	**85.0**
L5	57.5	76.1	79.5	**81.0**
L4	74.9	83.4	85.0	**86.4**
L3	67.2	81.7	82.8	**83.7**
L2	67.2	74.5	75.1	**75.6**
L1	64.4	62.1 (-2.3)	65.4 (+1.0)	**67.5** (+3.1)
Avg.	65.2	76.8	78.7	**79.9**

Table 5: Ablation study of GeCoin's core components using HKI on iNatLoc (Swin-B).

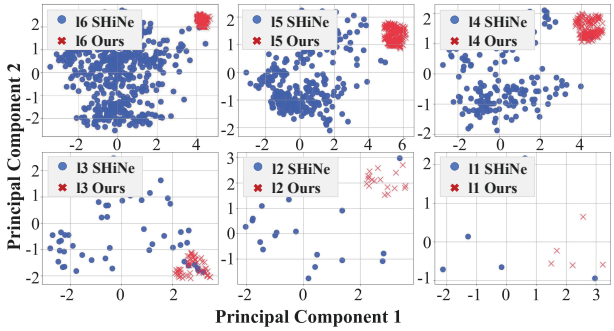


Figure 5: PCA visualization of the semantic space across different text label granularities on the iNatLoc dataset using CLIP ViT-B/32 encoders. Blue dots represent SHiNe, while red crosses represent GeCoin. GeCoin forms a more compact semantic space across all granularities, demonstrating that null-space projection preserves geometric integrity while integrating enhanced knowledge, preventing the semantic drift that causes performance instability.

Component-wise Ablation Analysis (RQ3)

We systematically evaluate GeCoin's components through ablation studies comparing unconstrained injection, Case 2

LLM	Clean		+20% Noise	
	SHiNe	GeCoin	SHiNe	GeCoin
GPT-3.5	46.5	**51.4**	43.0	**49.5**
Claude-3.7-Sonnet	47.8	**52.0**	43.9	**50.4**
Gemini-2.5-Pro	47.3	**51.7**	44.0	**50.0**
Std. Dev.	0.54	**0.25**	0.45	**0.37**

Table 6: Robustness to LLM Variation and Knowledge Noise. We report average mAP50 (%) on iNatLoc with a ResNet-50 backbone. "Clean" refers to the original LLM-generated hierarchy, while "+20% Noise" refers to adding 20% irrelevant concepts. GeCoin shows higher stability and lower Performance Drop.

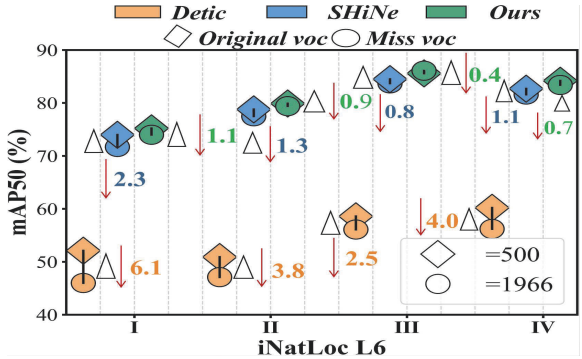


Figure 6: Evaluation of performance with noisy, mis-specified label vocabularies on iNatLoc L6. We compare the detection results of GeCoin (green), SHiNe (blue) and Detic (orange) across supervision signals. Performance is shown for original ($\diamond$) and expanded mis-specified ($\bigcirc$) vocabularies, with mAP50 drops (Δ) highlighted by red arrows.

(query concept-only protection), and Case 1 (hierarchical protection) on iNatLoc. Table 5 presents the analysis:

- **Obs 7: Null-space projection is essential for preventing geometric distortion.** Unconstrained injection (w/o Proj.) improves fine-grained performance over the Detic baseline, but this gain comes at the severe cost of degrading general concept recognition (L1: -2.3%) due to semantic anchor shift. This quantitative collapse is substantiated in Figure 3, illustrating how unconstrained injection causes severe feature drift and cluster boundary contamination. In contrast, our GeCoin method maintains distinct cluster boundaries by safely integrating knowledge.
- **Obs 8: Hierarchical protection strategy significantly reduces broad concept performance degradation and enhances fine-grained recognition.** The systematic progression from no projection (-2.3% L1 drop) to single concept protection (+1.0% improvement) to full hierarchical protection (+3.1%) demonstrates the cumulative benefit of preserving hierarchical orthogonality at multiple levels. This occurs because null-space injection better isolates irrelevant knowledge, making feature representations more compact with clearer cluster boundaries. This aligns with VLMs' intrinsic use of hierarchical orthogonality to distinguish hierarchical nodes.

Robustness to Knowledge Quality and Source Variations (RQ4)

To evaluate GeCoin's robustness against knowledge imperfections and source variations, we evaluate robustness across LLM variations with 20% noise injection (Table 6), and vocabulary misspecification using expanded label sets (Figure 6). Qualitative analysis includes detection visualizations (Figure 4) and semantic space visualization (Figure 5):

- **Obs 10: GeCoin demonstrates superior stability and noise resistance across different LLMs.** Table 6 shows that GeCoin consistently outperforms SHiNe across all three LLMs with significantly improved stability (standard deviation: 0.25 vs. 0.54 for clean data, 0.37 vs. 0.45 for noisy data). When 20% noise is injected, GeCoin shows remarkable resilience with only -1.7 average performance drop compared to SHiNe's -3.6 collapse, while preserving the natural performance hierarchy among LLMs (Claude: -1.6 $>$ Gemini: -1.7 $>$ GPT-3.5: -1.9). This 2.1× better noise resistance, 2.2× improved clean stability, and 1.2× improved noise stability demonstrate that GeCoin's null-space projection effectively filters semantic contamination regardless of knowledge source quality.
- **Obs 11: GeCoin exhibits superior resilience to vocabulary misspecification and maintains compact semantic representations.** To evaluate robustness to mis-specified vocabularies, we expanded the leaf vocabularies (L6) by adding 1966 classes from OpenImages (Kuznetsova et al. 2020) and LVIS datasets. Figure 6 shows that while Detic's mAP50 drops by -6.1, GeCoin maintains the smallest decline of -1.5 under I: LVIS+IN-L. The qualitative analysis (Figure 4) demonstrates consistent detection across granularity levels, while PCA visualization (Figure 5) reveals that GeCoin produces significantly more compact semantic clustering compared to SHiNe's scattered distribution, indicating effective noise filtering and semantic consistency preservation.

Conclusion

In this work, we identified and addressed a fundamental problem of geometric distortion in VLM knowledge injection. We proposed GeCoin, a training-free framework that safely enhances VLMs by projecting external knowledge into a constrained null-space. Our extensive experiments systematically demonstrated GeCoin's superior effectiveness over SOTA methods (RQ1), its broad generality across different settings (RQ2), the critical role of its geometric constraint components (RQ3), and its strong robustness against noisy and varied knowledge sources (RQ4). GeCoin offers a new paradigm for reliable knowledge integration in VLMs, highlighting that preserving geometric integrity is paramount for safe and effective semantic expansion, with promising implications for broader applications.

Acknowledgments

This research is funded by the National Key Research and Development Program of China (Grant 2021YFC3340800), the National Natural Science Foundation of China (No. 62406347), the Hunan Provincial Natural Science Foundation of China (No. 2024JJ5448), and the Science and Technology Research and Development Program of China State Railway Group Co., Ltd (No. N2024W006).

References

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

Cole, E.; Wilber, K.; Van Horn, G.; Yang, X.; Fornoni, M.; Perona, P.; Belongie, S.; Howard, A.; and Aodha, O. M. 2022. On label granularity and object localization. In *European Conference on Computer Vision*, 604–620. Springer.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Fan, Q.; Zhuo, W.; Tang, C.-K.; and Tai, Y.-W. 2020. Few-shot object detection with attention-RPN and multi-relation detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4013–4022.

Fang, J.; Jiang, H.; Wang, K.; Ma, Y.; Jie, S.; Wang, X.; He, X.; and Chua, T.-S. 2024. Alphaedit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355*.

Fellbaum, C. 1998. *WordNet: An electronic lexical database*. MIT press.

Fu, S.; Yang, Q.; Mo, Q.; Yan, J.; Wei, X.; Meng, J.; Xie, X.; and Zheng, W.-S. 2025. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14987–14997.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5356–5364.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jin, S.; Jiang, X.; Huang, J.; Lu, L.; and Lu, S. 2024. Llms meet vlms: Boost open vocabulary object detection with fine-grained descriptors. *arXiv preprint arXiv:2402.04630*.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7): 1956–1981.

Lin, C.; Sun, P.; Jiang, Y.; Luo, P.; Qu, L.; Haffari, G.; Yuan, Z.; and Cai, J. 2022. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, M.; Hayes, T. L.; Ricci, E.; Csurka, G.; and Volpi, R. 2024a. SHiNe: Semantic Hierarchy Nexus for Open-vocabulary Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16634–16644.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Ma, C.; Jiang, Y.; Wen, X.; Yuan, Z.; and Qi, X. 2024. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in neural information processing systems*, 36.

Novack, Z.; McAuley, J.; Lipton, Z. C.; and Garg, S. 2023. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, 26342–26362. PMLR.

Park, K.; Choe, Y. J.; Jiang, Y.; and Veitch, V. 2024. The geometry of categorical and hierarchical concepts in large language models. *arXiv preprint arXiv:2406.01506*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Saglam, B.; Kassianik, P.; Nelson, B.; Weerawardhena, S.; Singer, Y.; and Karbasi, A. 2025. Large Language Models Encode Semantics in Low-Dimensional Linear Subspaces. *arXiv preprint arXiv:2507.09709*.

Santurkar, S.; Tsipras, D.; and Madry, A. 2020. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Trager, M.; Perera, P.; Zancato, L.; Achille, A.; Bhatia, P.; and Soatto, S. 2023. Linear spaces of meanings: compositional structures in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15395–15404.

Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M. F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Wang, S.; Li, X.; Sun, J.; and Xu, Z. 2021. Training Networks in Null Space of Feature Covariance for Continual Learning. In *CVPR*, 184–193. Computer Vision Foundation / IEEE.

Yao, L.; Pi, R.; Han, J.; Liang, X.; Xu, H.; Zhang, W.; Li, Z.; and Xu, D. 2024. DetCLIPv3: Towards Versatile Generative Open-vocabulary Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27391–27401.

Zareian, A.; Rosa, K. D.; Hu, D. H.; and Chang, S.-F. 2021. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14393–14402.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, C.; Cote, M.-A.; Albada, M.; Sankaran, A.; Stokes, J. W.; Wang, T.; Abdi, A.; Blum, W.; and Abdul-Mageed, M. 2025. DefenderBench: A Toolkit for Evaluating Language Agents in Cybersecurity Environments. *arXiv preprint arXiv:2506.00739*.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, 350–368. Springer.