Recovery Attacks on Distance-Preserving Encrypted Vector Databases

Chen Gong[1], Xiangyu Wang[1], Dan Zhu[2], Bowen Zhao[1], Jie Li[3], Xinghua Li[1], Jianfeng Ma[1]
[1]*Xidian University* [2]*Northwestern Polytechnical University* [3]*Huawei Technologies*
[1]*cgong_11@stu.xidian.edu.cn, wangxiangyu01@xidian.edu.cn, bwinzhao@gmail.com, xhli1@mail.xidian.edu.cn*
[1]*jfma@mail.xidian.edu.cn* [2]*zhudan@nwpu.edu.cn* [3]*lijie303@huawei.com*

***Abstract*—Private retrieval-augmented generation (RAG) systems based on distance-comparison-preserving encryption (DCPE) are widely adopted in industry to protect sensitive embeddings while maintaining low-latency search. However, DCPE inherently leaks the relational geometry of stored embeddings by preserving relative distance comparisons across ciphertexts, making its security in practical deployments questionable. This paper demonstrates that the geometric structure DCPE preserves to enable nearest-neighbor search directly exposes the underlying data manifold to full database recovery from snapshots. We introduce two practical recovery attacks: (1) Attack A, which requires only a leaked snapshot, the public encoder, and an auxiliary candidate corpus, achieving 100% record re-identification; and (2) Attack B, which requires no candidate corpus and uses a single known plaintext-ciphertext pair to resolve the secret scale in DCPE, reconstructing embeddings with 0.997 mean cosine similarity and restoring a functional search index. Furthermore, our evaluation indicates that existing defensive countermeasures offer minimal protection. Crucially, full-database reconstruction is computationally efficient, incurring an execution overhead that is even lower than that of the original encryption pass. Therefore, leaked DCPE snapshots must be treated as plaintext disclosures, underscoring the urgent need to explore and adopt truly secure cryptographic alternatives for private RAG.**

***Index Terms*—Encrypted vector databases, distance-preserving encryption, database recovery attacks, retrieval-augmented generation.**

1. Introduction

Retrieval-Augmented Generation (RAG) is widely adopted in enterprise pipelines and agentic AI systems [1]. In a normal RAG deployment, the system embeds private documents, records, and multimodal assets using a public foundation encoder into high-dimensional floating-point embeddings, which are stored in a vector database. When a user submits a query, the system embeds it using the same public encoder, computes pairwise similarities (e.g., cosine similarity) against the database, and retrieves the top-k nearest neighbors as grounding context for a large language model. At its core, vector retrieval relies on a single geometric primitive: *evaluating relative distances across high-dimensional embeddings*.

Embeddings do not function as one-way cryptographic hashes. Prior embedding-inversion attacks have shown that embeddings preserve semantic structure, enabling adversaries to reconstruct the underlying text [2], [3] and images [4] with high fidelity. Because persistent vector databases store sensitive data, encrypting stored embeddings at rest is a critical security requirement. Reconciling embedding confidentiality with low-latency retrieval, however, introduces a fundamental systems challenge. Standard cryptographic solutions rely on primitives such as homomorphic encryption [5], multi-party computation [6], private information retrieval [7], or oblivious RAM [8]. While these approaches provide high security guarantees by hiding search/access patterns, they incur considerable computational and communication overhead, often reducing query throughput by several orders of magnitude compared to native Approximate Nearest-Neighbor Search (ANNS).

To address this issue, practical deployments [9], [10], [11], [12] have converged on Distance-Comparison-Preserving Encryption (DCPE) [13], predominantly using the *Scale-and-Perturb (SAP)* construction. SAP scales plaintext embeddings by a shared secret scalar and adds bounded isotropic noise:

$$c_i = sp_i + \lambda_i, \quad s > 0, \quad \|\lambda_i\| \leq r, \tag{1}$$

where p_i is the plaintext embedding, s is the secret scale, and λ_i is the bounded perturbation. By design, SAP preserves relative distance comparisons. Because ciphertexts remain standard floating-point vectors, vector engines (e.g., HNSW [14], DiskANN [15], Milvus [16]) can index/search them approaching native speed with minimal recall loss.

However, the security of DCPE rests on an incomplete threat model. Prior analyses evaluate ciphertext indistinguishability only in isolation [13], [17]. In practice, the primary threat to data at rest is a *static database snapshot*, such as an exfiltrated backup, an exposed storage volume, or an insider leak. In a persistent database, all vectors are encrypted under the same secret key and scale. Consequently, the relative distance relations preserved for searchability directly expose the geometric structure of the stored data to snapshot recovery.

We formalize this exposure by identifying an inherent algebraic vulnerability in distance-preserving vector search. Cosine similarity is widely used in production vector databases to evaluate nearest neighbors. Under normalized

angular comparison against candidate embeddings e_j, the secret scale s in Equation (1) cancels out:

$$\cos(c_i, e_j) = \frac{\langle sp_i + \lambda_i, e_j \rangle}{\|sp_i + \lambda_i\| \|e_j\|} = \frac{\langle p_i + \lambda_i/s, e_j \rangle}{\|p_i + \lambda_i/s\| \|e_j\|}. \quad (2)$$

Normalization eliminates the global secret s, leaving an exposed metric channel corrupted only by the normalized perturbation λ_i/s.

Crucially, this noise cannot be arbitrarily increased to protect the underlying embeddings. To maintain acceptable retrieval utility, the perturbation radius must remain smaller than the semantic margins between data clusters. Any noise budget that allows legitimate queries to find nearest neighbors simultaneously allows an adversary to align ciphertexts with candidate records. We argue that just one reasonable weak assumption is enough for *distance-preserving encrypted vector search to be essentially equivalent to plaintext leakage*.

In this paper, we propose two practical attacks against DCPE-based encrypted vector databases.

- **Attack A (Ciphertext-Only Identity De-Anonymization):** Under a minimal threat model where the adversary possesses only the leaked snapshot, the public encoder, and an auxiliary candidate corpus, Equation (2) enables direct record re-identification without secret keys, queries, or known plaintext-ciphertext pairs. Attack A achieves 97.1% accuracy on 50,000 MS MARCO records against 100,000 candidates. Furthermore, formulating de-anonymization as a global assignment problem and solving it via the Hungarian algorithm raises the recovery rate to 100.0%, demonstrating that leaked snapshots can be fully de-anonymized at scale without cryptographic secrets.
- **Attack B (Anchor-Assisted Recovery):** When the candidate corpus is unavailable or non-enumerable, the adversary can still reconstruct the embeddings if they obtain as few as $m \geq 1$ known plaintext-ciphertext pairs (anchors). Attack B requires no candidate corpus. Instead, it estimates the secret scale s in closed form using an optimal least-squares estimator with a bounded error. Remarkably, a single anchor ($m = 1$) suffices to reconstruct raw embeddings with 0.997 cosine similarity, yielding a stolen search index that reproduces 100% of the original system's top-10 retrieval results.

Attack Impact. As shown in Figure 1(a), our direct linear recovery (Attack B) is as fast as a single encryption pass, reconstructing one million vectors in only 1.02 ms on a consumer GPU. As shown in Figure 1(b), legitimate retrieval utility degrades much faster than re-identification risk, destroying service utility before privacy can be achieved. Furthermore, recovering embeddings allows existing inversion decoders to reconstruct sensitive source text and images (see § 4). Finally, the success of our attacks against multiple proposed DCPE defenses demonstrates the fundamental limitation of distance-preserving designs (see § 5).

The remainder of this paper is organized as follows. Section 2 formalizes the DCPE, the Database Snapshot Confidentiality (DSC) objective, and the threat models. Section 3 presents our recovery attacks. Section 4 evaluates attack performance. Section 5 evaluates existing defensive countermeasures. Section 7 and 6 discuss related work and our results. Finally, Section 8 concludes the paper.

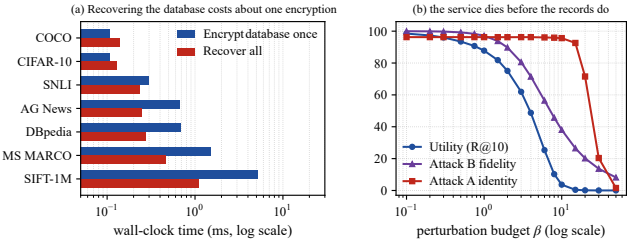

Figure 1. DCPE fails to protect vector databases. (a) The computational cost of Attack B recovery is lower than that of database encryption. (b) With the increase of encryption noise, the service utility of the vector search system encrypted by DCPE (Recall @10) is first destroyed.

TABLE 1. NOTATION SUMMARY.

Symbol	Description
Data and Models	
$D, \widehat{D}$	Private database and auxiliary candidate corpus
P, C	Plaintext embeddings and ciphertext database
p_i, c_i, e_j	Plaintext embedding, ciphertext vector, and candidate embedding
N, M, d	Database size, candidate corpus size, and embedding dimension
$\mathcal{M}$	Public foundation embedding model
DCPE Scheme and Attack Parameters	
$s, \hat{s}_{\mathrm{LS}}$	Secret global scale and its least-squares estimate
λ_i, r	Perturbation noise and its bounding radius
β	Scheme perturbation budget parameter
m	Number of anchor pairs
Evaluation Metrics	
Φ_{id}	Identity recovery rate
Φ_{num}	Vector reconstruction fidelity
Φ_{util}	Index retrieval utility
U, U_0	Utility under encrypted or clean queries

2. Problem Formulation and Threat Model

This section formalizes the target cryptographic construction, formalizes our deployment-level security definition, and defines the threat model. A summary of notations is provided in Table 1.

2.1. Distance-Comparison-Preserving Encryption

Let $\mathcal{P} \subset \mathbb{R}^d$ denote a high-dimensional vector space populated by dense embeddings. A dataset of private enterprise records is denoted as $D = (t_1, t_2, \ldots, t_N)$, where each entry t_i represents a textual document, code snippet, or multimodal asset. A public, deterministic foundation encoder $\mathcal{M} : \mathcal{T} \to \mathcal{P}$ maps each record into a normalized

plaintext embedding $p_i = \mathcal{M}(t_i)$, yielding the plaintext database matrix $P = [p_1, \ldots, p_N]^\top \in \mathbb{R}^{N \times d}$.

DCPE [13] is a symmetric property-preserving scheme designed to support ANNS directly over outsourced ciphertexts. Formally, a DCPE scheme consists of three polynomial-time algorithms:

- $\mathsf{KeyGen}(1^\kappa) \to K$: Outputs a secret key K. In the canonical Scale-and-Perturb (SAP) family [13], the key includes a secret global scaling factor $s \in \mathbb{R}^+$ and a perturbation parameter $\beta > 0$.
- $\mathsf{Enc}(K, p) \to c$: Encrypts a plaintext vector $p \in \mathcal{P}$ by scaling it and injecting bounded isotropic noise:

$$c = s \cdot p + \lambda, \quad \text{where } \|\lambda\|_2 \leq r. \tag{3}$$

  In baseline SAP, the perturbation vector λ is sampled uniformly from a d-dimensional ball of radius $r = s\beta/4$, ensuring that the relative displacement ratio is bounded by β.
- $\mathsf{Search}(c_q, \{c_i\}_{i=1}^N) \to \mathcal{I}_k$: Given an encrypted query c_q and a ciphertext database $C = [c_1, \ldots, c_N]^\top$, the server evaluates pairwise distance or similarity comparisons (e.g., cosine similarity) to output the indices $\mathcal{I}_k$ of the k-nearest neighbors.

Recent works, such as CAPRISE [17], introduce an asymmetric noise allocation under the hypothesis that obscuring inter-database distances thwarts geometric inference on stored vectors. In CAPRISE, the database encryption samples noise from a larger radius $r_{\text{db}} = 3s\beta/8$, whereas query encryption injects a smaller perturbation $r_{\text{query}} = s\beta/8$. Our analysis evaluates the general design pattern of Equation (3) and directly examines whether this asymmetric allocation offers tangible protection.

2.2. Database Snapshot Confidentiality

The cryptographic evaluations of DCPE focus on indistinguishability games over isolated pairs of ciphertexts [13]. However, this primitive-level formulation fails to capture the security of persistent vector stores at rest, in which an adversary obtains a static copy of the entire dataset encrypted under a single key. To formalize this operational threat, we define *Database Snapshot Confidentiality (DSC)*.

2.2.1. The DSC Game. We formalize DSC as an experiment $\mathsf{Exp}^{\text{DSC}}_{\Pi,\mathcal{A}}(\lambda)$ played between a challenger and an adversary $\mathcal{A}$:

1) **Setup:** The challenger samples a secret key $K \leftarrow \mathsf{KeyGen}(1^\kappa)$ and selects a private corpus $D = (t_1, \ldots, t_N)$. It computes the embeddings $P = \mathcal{M}(D)$ and encrypts each record $c_i \leftarrow \mathsf{Enc}(K, p_i)$ to form the ciphertext database $C \in \mathbb{R}^{N \times d}$.
2) **Snapshot Exposure:** The challenger gives the static ciphertext snapshot C and auxiliary information to $\mathcal{A}$.
3) **Reconstruction:** The adversary outputs an estimate of the plaintext records $\widehat{D} = (\widehat{t}_1, \ldots, \widehat{t}_N)$ and/or an estimate of the embedding values $\widehat{P} = [\widehat{p}_1, \ldots, \widehat{p}_N]^\top$.

2.2.2. Database Recovery Metrics. To measure the extent of information leakage from a static database snapshot, we evaluate recovery along three dimensions:

- **Identity Recovery Rate (Φ_{id}):** Measures the proportion of ciphertext rows whose corresponding plaintext records are correctly de-anonymized:

$$\Phi_{\text{id}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\widehat{t}_i = t_i), \tag{4}$$

  where $\mathbb{I}(\cdot)$ denotes the indicator function. This metric assesses whether an adversary can link encrypted rows back to real-world documents or entities. By default, Φ_{id} reports row-wise top-1 matching.
- **Vector Reconstruction Fidelity (Φ_{num}):** Measures the average cosine similarity between the reconstructed embeddings and the ground-truth embeddings:

$$\Phi_{\text{num}} = \frac{1}{N} \sum_{i=1}^{N} \cos(\widehat{p}_i, p_i) = \frac{1}{N} \sum_{i=1}^{N} \frac{\langle \widehat{p}_i, p_i \rangle}{\|\widehat{p}_i\|_2 \|p_i\|_2}. \tag{5}$$

  A value close to 1.0 indicates near-lossless directional recovery of the original embedding matrix P.
- **Index Retrieval Utility (Φ_{util}):** Measures whether the reconstructed embeddings remain operational as a functional search engine:

$$\Phi_{\text{util}} = \text{Recall@}k(\widehat{P}; Q), \tag{6}$$

  defined as the top-k nearest-neighbor overlap between queries evaluated on the recovered vector matrix $\widehat{P}$ and queries evaluated on the original plaintext database P.

The adversary advantage in the DSC game under metric $\Phi \in \{\Phi_{\text{id}}, \Phi_{\text{num}}, \Phi_{\text{util}}\}$ is defined as:

$$\mathbf{Adv}^{\text{DSC}}_{\Pi,\mathcal{A},\Phi} = \mathbb{E}\left[\Phi\left(\mathcal{A}(C, \mathcal{M}, \mathsf{aux}), D\right)\right]. \tag{7}$$

2.3. Threat Model

We consider a passive, persistent snapshot adversary $\mathcal{A}$ who obtains a static dump of the encrypted vector database $C = [c_1, \ldots, c_N]^\top \in \mathbb{R}^{N \times d}$. We assume no specific compromise vector, and the snapshot reflects realistic data-at-rest leakages. In line with enterprise RAG architectures, the embedding model $\mathcal{M}$ is public deployment metadata rather than a secret cryptographic parameter. Vector database clients routinely require the exact model identifier and embedding dimension d to produce query vectors, and foundation model weights (e.g., OpenAI text-embedding, BGE, GTR, CLIP) are publicly accessible or callable via open APIs. The adversary therefore possesses access to $\mathcal{M}$, allowing local generation of candidate embeddings.

Beyond the ciphertext snapshot and the public encoder, the adversary's capability is parameterized by the available auxiliary external information. We design database recovery under two disjoint auxiliary settings:

1) **Ciphertext-Only.** The adversary possesses an auxiliary candidate corpus $\widehat{D} = \{\widehat{t}_1, \ldots, \widehat{t}_M\}$ representing publicly available or crawlable domain text. In this regime,

the adversary holds no cryptographic keys, issues no queries to the production service, and possesses no known plaintext-ciphertext pairs ($m = 0$). The adversary aims to maximize Φ_{id} by directly de-anonymizing the private records.

2) **Anchor-Informed.** The adversary is assumed to acquire a minimal budget of $m \geq 1$ known plaintext-ciphertext pairs (as anchors) $\{(p_a, c_a)\}_{a=1}^{m}$. In commercial RAG systems, such pairs can be obtained when an adversary interacts with multi-tenant databases as a legitimate client. Specifically, any $m \geq 1$ pairs are enough. The adversary is indifferent to whether these pairs are obtained from known-plaintext (KPA) or chosen-plaintext (CPA) observations. The adversary aims to maximize Φ_{num} and Φ_{util}.

Trust Assumptions and Non-Goals To ensure a rigorous and realistic security evaluation, we explicitly establish the following operational boundaries: *Key Secrecy:* The cryptographic key management system is fully trusted. The adversary cannot extract the secret scale s or cryptographic random seeds via side channels, memory leakage, or key-schedule compromises. *Data-in-Use Protections:* We do not target runtime memory execution. Our attack targets only persistent data at rest. *No Production Query Access:* The adversary cannot observe runtime query traffic, communication volumes, or server-side access patterns.

2.4. Auxiliary Corpus Coverage

In real-world deployments, an auxiliary candidate corpus $\widehat{D}$ rarely constitutes a byte-for-byte replica of the protected private database D. To rigorously model partial adversary knowledge, we define the corpus coverage parameter:

$$\kappa = \frac{|\widehat{D} \cap D|}{|D|} \in [0, 1], \tag{8}$$

where $|\widehat{D} \cap D|$ denotes the number of distinct records co-occurring in both the attacker's auxiliary candidate collection and the target database. A coverage of $\kappa = 1.0$ indicates that every private record is enumerable within the candidate envelope, whereas $\kappa < 1.0$ captures realistic scenarios where the candidate set misses a fraction of private records or is diluted by large volumes of irrelevant, out-of-domain distractor documents. We explicitly isolate per-covered-record recovery (A_{cov}) from whole-database recovery (A_{all}) throughout our evaluation.

3. The Recovery Attacks

Equation (2) establishes that cosine ranking evaluates ciphertexts independently of the secret global scale s. The normalized ciphertext acts as a metric channel corrupted solely by the scaled perturbation λ_i/s. In this section, we develop two recovery attacks targeting this channel:

- **Attack A (§3.1):** Recovers identities by matching embeddings in the corpus against the leaked snapshot.
- **Attack B (§3.2):** Reconstructs embeddings and clones the search index using minimal anchor supervision.

In the SAP construction, the positive scalar s is the only shared secret. Under cosine similarity (and normalized inner products), s cancels out. For Attack A, when $m = 0$, an adversary requires no key material to identify records. For Attack B, when $m \geq 1$, estimating s in closed form avoids the only transformation applied to the plaintext embeddings, leaving a residual that consists entirely of zero-mean, isotropic perturbation noise. Because the perturbation noise is isotropic and statistically independent of the underlying embedding distribution, and the non-linear model cannot exploit non-existent correlations with the data manifold. Direct linear recovery is geometrically sufficient, and we provide a theoretical error bound for Attack B.

3.1. Attack A

Attack A is a ciphertext-only identity de-anonymization. It addresses the auxiliary-informed adversary model ($\mathcal{A}_{\mathrm{aux}}$): the adversary possesses the leaked snapshot C, the public encoder $\mathcal{M}$, and an auxiliary candidate corpus $\widehat{D} = \{\widehat{t}_1, \ldots, \widehat{t}_M\}$, but holds no secret keys and zero anchor pairs ($m = 0$). The attack proceeds in two steps: First, the adversary uses scale-invariant cosine similarity to locally match each leaked ciphertext to its closest candidate embedding (§ 3.1.1). Second, to resolve candidate collisions caused by perturbation noise, the adversary models database de-anonymization as a maximum-weight matching problem and solves it globally (§ 3.1.2).

3.1.1. Scale-Invariant Ranking and Geometric Analysis. The adversary embeds the candidate corpus locally using the public model, generating candidate embeddings $E = \{e_1, \ldots, e_M\}$, where $e_j = \mathcal{M}(\widehat{t}_j)$. For each leaked ciphertext c_i, the initial candidate match is determined greedily by:

$$\widehat{t}_i = \arg\max_{\widehat{t}_j \in \widehat{D}} \cos\left(c_i, \mathcal{M}(\widehat{t}_j)\right). \tag{9}$$

The validity of directly comparing ciphertexts with candidate embeddings relies on the scale invariance of cosine similarity. Substituting $c_i = sp_i + \lambda_i$:

$$\cos(c_i, e_j) = \frac{s\langle p_i, e_j\rangle + \langle \lambda_i, e_j\rangle}{\|sp_i + \lambda_i\|_2 \|e_j\|_2} = \frac{\langle p_i, e_j\rangle + \langle \lambda_i/s, e_j\rangle}{\|p_i + \lambda_i/s\|_2 \|e_j\|_2}. \tag{10}$$

For a fixed ciphertext c_i, the denominator $\|p_i + \lambda_i/s\|_2$ is identical across all candidate indices j. Furthermore, because the secret scale satisfies $s > 0$, the relative ordering induced by $\langle p_i, e_j\rangle$ remains unchanged unless the projection of λ_i/s exceeds the angular margin between candidates. The perturbation λ_i/s acts merely as additive angular noise over the metric geometry that DCPE is designed to preserve.

Figure 2 shows the geometric condition under which this ranking is preserved. For a target embedding p whose true nearest candidate is e_2, correct re-identification succeeds whenever the effective angular noise cone θ_{noise} remains

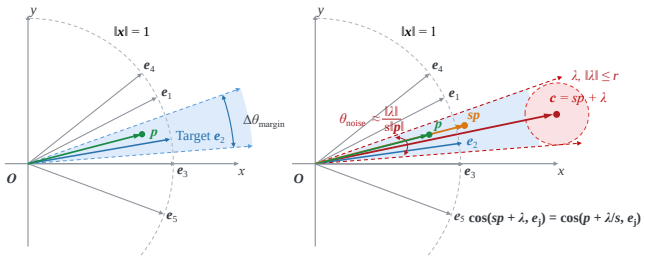

Figure 2. Mechanism of scale-invariance and ranking preservation.

within the candidate acceptance region bounded by the angular margin $\Delta\theta_{\text{margin}}$.

Crucially, this geometric containment is an unavoidable consequence of system functionality. DCPE is deployed specifically to support nearest-neighbor retrieval over encrypted data. If a deployer increases the perturbation budget β such that $\theta_{\text{noise}} \geq \Delta\theta_{\text{margin}}$, ciphertexts cross the decision boundaries of their true semantic clusters. Under this noise level, legitimate query vectors fail to retrieve their correct nearest neighbors, causing service recall to collapse. Therefore, to maintain operational utility, the system must set the noise radius such that $\theta_{\text{noise}} < \Delta\theta_{\text{margin}}$, which inherently keeps ciphertexts aligned with their corresponding candidate embeddings (as formalized in Appendix B).

This geometric constraint explains our empirical findings in Section 4.4: as perturbation β increases, legitimate retrieval utility (U) collapses long before the attacker's re-identification success rate (A) begins to degrade, leading to the observed situation $A \geq U_0 \geq U$.

3.1.2. Global Assignment via Hungarian Matching. Equation (9) defines an independent, row-wise decision rule. However, records within an enterprise database typically correspond to distinct entities. Due to perturbation noise, greedy row-wise matching can occasionally produce collisions, in which multiple ciphertexts select the same candidate record.

To resolve collisions and enforce a one-to-one mapping, we formulate database de-anonymization as a maximum-weight bipartite matching problem:

$$\max_{\pi \in \Pi} \sum_{i=1}^{N} S_{i,\pi(i)}, \tag{11}$$

where $S \in \mathbb{R}^{N \times M}$ is the pairwise cosine matrix $S_{i,j} = \cos(c_i, e_j)$, and Π denotes the set of injective assignments from $[N]$ to $[M]$. We solve Equation (11) using the Hungarian algorithm [18]. By resolving candidate contention globally, this assignment avoids collisions and raises the recovery rate to 100.0%.

3.2. Attack B

Attack B is an anchor-assisted database reconstruction. When a candidate corpus is unavailable or non-enumerable, Attack B reconstructs the continuous embeddings directly from the leaked snapshot using a small set of $m \geq 1$ known plaintext-ciphertext pairs (anchors) $\{(p_a, c_a)\}_{a=1}^{m}$.

Attack B proceeds in three steps: (1) Estimating the secret global scale s via ordinary least squares (§ 3.2.1). (2) Inverting the snapshot through scalar division to recover embeddings and search functionality (§ 3.2.2). Crucially, as we show below, high-dimensional concentration of measure ensures that a single anchor pair ($m = 1$) is sufficient. (3) An optional post-processing step, using local neighborhood smoothing to reduce residual perturbation noise (§ 3.2.3).

3.2.1. Linear Scale Estimation and Inversion. Reconstructing continuous embedding coordinates requires resolving the secret scale s. Given m known anchor pairs, the encryption relation $c_a = sp_a + \lambda_a$ forms an overdetermined linear system for s. We formulate the least-squares objective:

$$\min_{s} \sum_{a=1}^{m} \|c_a - sp_a\|_2^2. \tag{12}$$

Setting the derivative with respect to s to zero yields the closed-form estimator:

$$\hat{s}_{\text{LS}} = \frac{\sum_{a=1}^{m} \langle p_a, c_a \rangle}{\sum_{a=1}^{m} \|p_a\|_2^2}. \tag{13}$$

Because the perturbation noise λ_a is zero-mean and isotropic in expectation ($\mathbb{E}[\lambda_a] = 0$ and $\mathbb{E}[\langle p_a, \lambda_a \rangle] = 0$), the estimator $\hat{s}_{\text{LS}}$ is unbiased ($\mathbb{E}[\hat{s}_{\text{LS}}] = s$).

With $\hat{s}_{\text{LS}}$ obtained, the adversary reconstructs the entire database by dividing each leaked ciphertext vector by the estimated scale:

$$\widehat{P} = C/\hat{s}_{\text{LS}}, \qquad \widehat{p}_i - p_i = \left(\frac{s}{\hat{s}_{\text{LS}}} - 1\right) p_i + \frac{\lambda_i}{\hat{s}_{\text{LS}}}. \tag{14}$$

Equation (14) shows that the reconstruction error decomposes into two terms: a collinear scaling of the plaintext embedding and the scaled perturbation.

Because the collinear term is a positive scalar multiple of p_i, it identically cancels out under cosine similarity:

$$\widehat{p}_i = \frac{s}{\hat{s}_{\text{LS}}} \left(p_i + \frac{\lambda_i}{s}\right) \implies \cos(p_i, \widehat{p}_i) = \cos\left(p_i,\ p_i + \lambda_i/s\right). \tag{15}$$

Importantly, Equation (15) is independent of the scale estimate $\hat{s}_{\text{LS}}$. This linear reconstruction is therefore directionally exact: it eliminates the only structured transformation applied by the encryption scheme. The remaining residual consists solely of the uncorrelated perturbation λ_i/s.

3.2.2. Estimation Uncertainty. Next, we analyze the number of anchor pairs m needed to accurately recover s. Substituting $c_a = sp_a + \lambda_a$ into Equation (13) gives:

$$\hat{s}_{\text{LS}} - s = \frac{\sum_{a=1}^{m} \langle p_a, \lambda_a \rangle}{\sum_{a=1}^{m} \|p_a\|_2^2}. \tag{16}$$

Because the perturbations λ_a are drawn independently and uniformly from a d-dimensional ball of radius r, isotropy

implies $\mathbb{E}[\langle u, \lambda \rangle^2] = r^2/(d+2)$ for any unit vector u.[1] For each anchor, the variance of the inner product is:

$$\mathrm{Var}\big(\langle p_a, \lambda_a \rangle\big) = \|p_a\|_2^2 \,\mathrm{Var}\big(\langle u_a, \lambda_a \rangle\big) = \|p_a\|_2^2 \frac{r^2}{d+2}. \tag{17}$$

Summing across independent anchors yields the variance of the estimator:

$$\mathrm{Var}(\hat{s}_{\mathrm{LS}}) = \frac{r^2}{d+2} \cdot \frac{1}{\sum_{a=1}^{m} \|p_a\|_2^2}. \tag{18}$$

Assuming anchor embeddings have comparable norms ($\|p_a\|_2 \approx \|p\|_2$), the relative standard deviation of $\hat{s}_{\mathrm{LS}}$ satisfies:

$$\frac{\mathrm{sd}(\hat{s}_{\mathrm{LS}})}{s} \approx \frac{r}{s\|p\|_2\sqrt{m(d+2)}} = \frac{\rho}{\sqrt{m(d+2)}}, \tag{19}$$

where $\rho = r/(s\|p\|_2)$ is the nominal noise-to-signal ratio.

Equation (19) reveals that the uncertainty in estimating s decreases with vector dimension d. For typical embedding dimensions (e.g., $d = 768$), a single anchor pair ($m = 1$) reduces the relative scale uncertainty to:

$$\frac{\mathrm{sd}(\hat{s}_{\mathrm{LS}})}{s} \approx \frac{\rho}{\sqrt{770}} \approx 0.036 \cdot \rho. \tag{20}$$

The uncertainty in estimating s is bounded by 3.6% of the encryption noise itself. High-dimensional concentration of measure effectively aggregates d independent coordinate measurements, avoiding the need for multiple anchor pairs.

Why Additional Anchors Are Redundant. Equation (19) also shows why acquiring additional anchors ($m > 1$) does not improve search-based recovery. As established in Equation (15), angular metrics (including cosine similarity, top-k retrieval recall, and index rankings) depend solely on the noise floor λ_i/s and are invariant to the positive scalar $\hat{s}_{\mathrm{LS}}$. Additional anchors only refine the absolute Euclidean norm of recovered vectors, which normalized vector databases discard during similarity search. Directional recovery reaches its theoretical ceiling with just a single anchor.

3.2.3. Post-Recovery Denoising via Local Smoothing.

In high-noise situations (e.g., when a defender aggressively inflates β in an attempt to disrupt reconstruction), isotropic perturbation variance dominates embedding features. After recovering the unscaled vectors $\widehat{P}$, the adversary can optionally apply a lightweight k-nearest-neighbor (kNN) smoothing step to reduce residual perturbation noise:

$$\widehat{p}_i' = (1-\gamma)\widehat{p}_i + \gamma \left(\frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \widehat{p}_j \right), \tag{21}$$

1. Let σ^2 denote the noise variance along any unit direction: isotropy implies $\mathbb{E}[\lambda\lambda^\top] = \sigma^2 I_d$. For any unit vector u, $\mathbb{E}[\langle u, \lambda \rangle^2] = u^\top \mathbb{E}[\lambda\lambda^\top] u = \sigma^2$. Taking traces yields $d\sigma^2 = \mathbb{E}[\|\lambda\|^2]$. Because λ is uniform in the ball of radius r, the radial density is $f_R(R) = dR^{d-1}/r^d$ on $[0, r]$, giving $\mathbb{E}[\|\lambda\|^2] = \int_0^r R^2 \frac{dR^{d-1}}{r^d} dR = \frac{d}{d+2} r^2$. Therefore, $\sigma^2 = r^2/(d+2)$.

TABLE 2. EVALUATION DATASETS AND EXPERIMENTAL OBJECTIVES

Dataset	Modality	Scale	Target Encoder
MS MARCO	Text	50K	GTR-T5 (d=768)
AG News	Text	100K	GTR-T5 (d=768)
DBpedia-14	Text	100K	GTR-T5 (d=768)
COCO 2017	Image	5K	CLIP ViT-L (d=768)
CIFAR-10	Image	10K	Conv-AE (d=256)
SIFT-1M	-	1M	- (d=128)

where $\mathcal{N}_k(i)$ denotes the k-nearest neighbors of $\widehat{p}_i$ within $\widehat{P}$, and $\gamma \in [0, 1]$ is a smoothing coefficient. Because the encryption noise λ_i is independent and zero-mean across records, while true embeddings exhibit local semantic correlation, averaging over neighbors suppresses the noise variance by approximately $1/k$ without external supervision. The kNN smoothing acts as an effective denoising regularizer that filters out high-amplitude isotropic noise in noise-dominated environments (as validated in Appendix C).

4. Evaluations

4.1. Purpose and Methodology

We conduct evaluations to answer the questions below:

- **Identity re-identification** (§4.2): Can an adversary deanonymize records by performing Attack A?
- **Embedding reconstruction** (§4.3): How are the reconstruction effect and performance of Attack B?
- **Impact of DCPE parameters** (§4.4, §4.7): How do the DCPE parameters affect legitimate search and recovery? Is there a parameter setting that preserves search utility while preventing recovery?
- **Performance in non-ideal settings** (§4.5): How robust is our attack in non-ideal settings?
- **Computational overhead** (§4.6): Is our attack lightweight enough to prove that DCPE cannot provide computational protection?

Every dataset in this evaluation is summarized in Table 2. Unless otherwise specified, the private database contains the first $N = 50,000$ records, while the adversary's candidate corpus is initialized with $M = 100,000$ passages (yielding a baseline coverage of $\kappa = 0.5$ with 50% distractors). The default secret scale is set to $s = 3$, and the reference perturbation is $\beta = 0.3$. For visual retrieval, legitimate search utility is measured using independently encrypted queries with fresh noise per query, matching real-world client-side SDK implementations. All experiments are executed on a single workstation equipped with an NVIDIA RTX 5090 GPU (24GB VRAM), an AMD Ryzen 9 9950X3D CPU, 48GB of RAM, PyTorch 2.11, and CUDA 12.8. Run-to-run dispersion across five independent noise seeds is strictly below 0.10 percentage points.

4.2. Efficacy of Attack A

Table 3 reports identity recovery across all six datasets at $\beta = 0.3$ under the minimal adversary model. Row-wise top-1 matching achieves 96.98%-100.00% accuracy across five

TABLE 3. IDENTITY RECOVERY ($\beta = 0.3$)

Dataset	**Row-wise top-1**	**Global assignment** (Re-ID)
MS MARCO	96.98%	100.00%
AG News	98.84%	100.00%
DBpedia-14	100.00%	100.00%
SNLI	27.82%	99.06%
COCO 2017	100.00%	100.00%
SIFT-1M	100.00%	100.00%
Uniform-random	0.002%	—

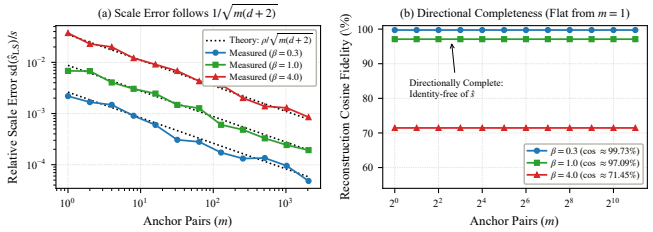

Figure 3. The impact of the number of anchors.

datasets. On SNLI, where extensive duplicate text causes greedy selection collisions (27.82%), global assignment resolves these conflicts, raising the recovery rate to 99.06% and bringing five of the six datasets to an exact 100.00% recovery rate.

Row-wise top-1 evaluation operates greedily and independently per record. On $N = 5,000$ records matched against 10,000 candidates, 144 records (2.88%) experience selection collisions, where multiple ciphertexts select the identical candidate. As shown in Table 3 (row 3), solving the global maximum-weight assignment problem via the Hungarian algorithm correctly resolves all 144 collisions, achieving 100.00% database-level recovery across the evaluated subsets. At scale ($N = 20,000$, candidates $M = 40,000$), conflict-resolving greedy assignment achieves 96.92%, which is statistically indistinguishable from row-wise matching (97.17%). Greedy row-wise top-1 matching, therefore, represents a conservative lower bound on the adversary's true recovery advantage.

To verify that re-identification stems from preserved geometric alignment rather than nearest-neighbor clustering artifacts, we run a negative control by uniformly permuting the ground-truth candidate identifiers. Under this control, recovery collapses to 0.002% (Table 3), demonstrating that the attack strictly exploits true semantic alignment.

4.3. Efficacy of Attack B

4.3.1. Impact of the number of anchors. Figure 3(a) plots the empirical relative standard deviation $\mathrm{sd}(\hat{s}_{\mathrm{LS}})/s$ against anchor budget m on logarithmic axes. The measured scale estimation error tracks the theoretical $1/\sqrt{m(d+2)}$ bound with near-perfect fidelity across all noise settings. At $d = 768$, a single anchor yields an estimation error of 0.22% at $\beta = 0.3$ and 0.68% at $\beta = 1.0$.

Figure 3(b) displays the downstream vector reconstruction fidelity. In complete agreement with Equation (15), the reconstruction cosine similarity is strictly constant across anchor counts: for $\beta = 0.3$, the cosine similarity is exactly 0.9973 at $m = 1$ and remains exactly 0.9973 at $m = 2,048$. Because cosine similarity cancels the positive scaling factor $\hat{s}_{\mathrm{LS}}$, additional anchors provide zero marginal gain in directional recovery.

4.3.2. Functional Index Replication Φ_{util}. We evaluate whether reconstructed embeddings can replicate the search functionality of the production service. Dividing ciphertexts by $\hat{s}_{\mathrm{LS}}$ preserves exact cosine rankings among records: the top-10 retrieval sets produced by the recovered database are 100.00% identical to those returned by the original ciphertext index across all tested queries.

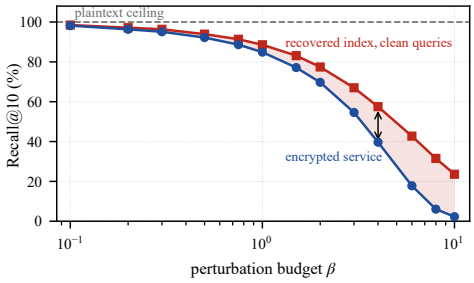

Figure 4. Φ_{util} against the attacked service. Blue is the legitimate encrypted service answering encrypted queries with fresh noise; red is the recovered index answering the same queries without query-side noise.

As shown in Figure 4, executing clean queries against the recovered database circumvents the query-side noise that the deployed service must tolerate. At $\beta = 4.0$, legitimate encrypted retrieval achieves only 39.7% Recall@10 due to query noise; in contrast, searching clean queries over the recovered index achieves 57.8% Recall@10, effectively restoring utility that the encryption layer destroyed.

4.3.3. Visual Reconstruction. While numerical cosine fidelity (Φ_{num}) quantifies geometric alignment, it is essential to demonstrate what vector recovery implies at the human-readable source-data level. To evaluate end-to-end data exposure, we pass the recovered embeddings $\widehat{p}$ into an off-the-shelf, external embedding-to-text inversion model (Huang et al. [4]). We treat the inversion model strictly as a downstream visualization reader to examine how closely the decoded text approximates the sensitive original record.

Figure 5 shows representative decoding outputs on an MS MARCO document under $\beta = 0.2$ and $s = 3$. We compare four distinct states of the representation:

1) **Clean Embedding Ceiling:** Passing the unperturbed plaintext embedding p to the inversion tool establishes the practical ceiling of the external decoder, reproducing the original sentence almost verbatim.
2) **Reconstructed Embedding:** Core-sensitive entities, including "Manhattan Project", "atomic bomb", "World

Case 1: Ground-Truth Document (Stored in Victim Database)
"The Manhattan Project and its atomic bomb helped bring an end to World War II. Its legacy of peaceful uses of atomic energy continues to have an impact on modern medicine."
Case 2: Decoded from Clean Embedding [Ceiling]
"The Manhattan Project and its atomic bomb helped bring an end to World War II. Its legacy of peaceful uses of atomic energy continues to have an impact on medicine."
Case 3: Decoded from Attack B Recovery (Ours) [Recovered]
"The Manhattan Project and its atomic bomb helped bring an end to World War II. Its legacy of peaceful atomic energy use has continued to have an impact on health."
Case 4: Decoded from Raw DCPE Ciphertext [Corrupted]
"With the help of the Manhattan Project and its legacy of peaceful energy, the atomic bomb brought an end to World War II. Manhattan nuclear power left a heavy mark."

Figure 5. Text reconstruction via an external embedding inversion model (Huang et al. [4]) on MS MARCO ($\beta = 0.2, s = 3$).

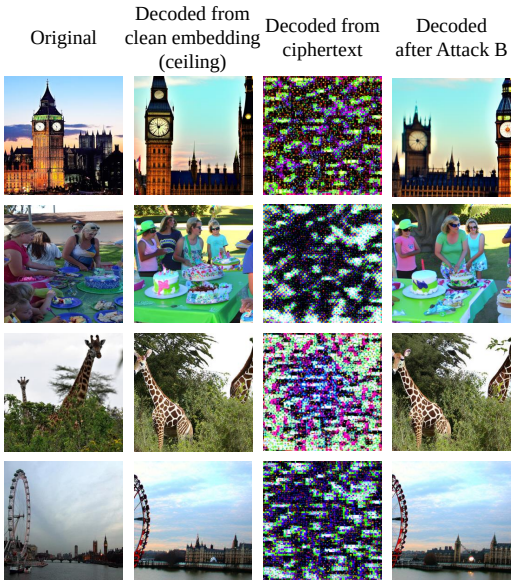

Figure 6. Image reconstruction on COCO ($\beta = 0.3$, $s = 3$).

War II", and "peaceful atomic energy", are fully recovered with coherent grammatical structure.

3) **Raw DCPE Ciphertext Baseline:** In stark contrast, attempting to decode the raw, unrescaled ciphertext $c = sp + \lambda$ directly through the inversion tool produces severe text distortion. The decoded text hallucinates distorted phrases (such as "legacy of peaceful energy" and "Manhattan nuclear power left"), failing to reconstruct the true sentence structure.

These qualitative results demonstrate that Attack B's closed-form scale resolution is essential. By removing the secret scale s and realigning the vector back to the natural embedding manifold, Attack B allows existing inversion decoders to extract human-readable plaintexts from stolen database snapshots.

This semantic exposure is not confined to discrete textual semantics. Figure 6 evaluates whether Attack B enables image reconstruction on COCO CLIP image embeddings ($d = 768$, $\beta = 0.3$, $s = 3$), passing the representations through a public unCLIP decoder. The first column shows the original image, and the second column shows the image directly decoded from clean embeddings, serving as the ceiling for Attack B. The third column is the unreadable noise obtained by directly decoding the DCPE ciphertext vector. The fourth column shows the image decoded from the vector recovered by our proposed Attack B, which is semantically identical to the image decoded directly from the original embedding.

These results demonstrate that once a static snapshot is recovered via Attack B, off-the-shelf foundation decoders can directly expose sensitive visual and textual records without requiring adversary training.

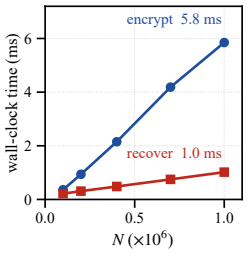

Figure 7. Cost of Attack B against database size on SIFT-1M. Medians of ten timed runs on one RTX 5090. At $N{=}10^6$ recovery takes 1.02 ms. Encryption takes $\sim 5.7\times$ longer.

4.3.4. Scalability to Million-Vector Stores. The matter at production scale is whether Attack B can practically reconstruct million-scale vectors. We evaluate computational scalability on SIFT-1M (from $N = 10^5$ to $N = 10^6$). As shown in Figure 7, the cost of recovery is linear in N, and at $N{=}10^6$ it takes 1.02 ms against 5.85 ms to encrypt the same store, so the gap widens as the deployment grows. Reconstructing a production-scale database requires neither complex iterative solvers nor distributed infrastructure.

4.4. Impact of Perturbation on Utility and Attack Performance

In the SAP construction, the perturbation budget β is the primary parameter governing privacy: larger values inject more noise, obscuring the underlying vector geometry. A natural defensive strategy is to increase β. However, adding noise inherently degrades the search recall. We therefore evaluate the system across a wide range of perturbation levels to examine whether a feasible parameter setting exists that balances legitimate retrieval utility with protection against snapshot recovery.

Figure 8 plots the simultaneous trajectories of legitimate retrieval utility (U, Recall@10), Attack B metric reconstruction (Φ_{num}, cosine similarity), and Attack A identity re-identification (A, $m = 0$) across all six datasets as perturbation β scales from 0.1 to 50. Across all datasets, the empirical curves demonstrate a consistent order of degradation: legitimate search utility drops first and most rapidly, continuous coordinate recovery degrades second, and record identity re-identification remains effective until extreme noise levels.

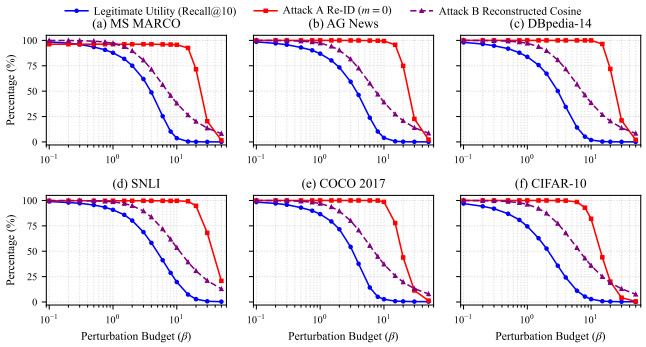

Figure 8. Impact of perturbation budget on legitimate search utility and adversarial recovery success across six datasets.

Across moderate perturbation ranges ($\beta \in [1.0, 10.0]$), legitimate search utility suffers steep degradation. On MS MARCO, service Recall@10 drops from 87.78% at $\beta = 1.0$ to 3.66% at $\beta = 10.0$, with a parallel collapse on COCO from 99.84% to 12.51%. In contrast, Attack B metric recovery degrades far more slowly, maintaining 71.45% cosine fidelity on MS MARCO at $\beta = 4.0$ and 36.17% even at $\beta = 10.0$, preserving structural correlation for downstream inversion. Attack A identity recovery exhibits even greater persistence: across all text datasets, re-identification remains above 95.6% up to $\beta = 10.0$. On DBpedia-14, identity recovery remains at 99.98% at $\beta = 10.0$, where legitimate recall is merely 1.88%. Identity matching degrades only under extreme perturbation ($\beta \geq 15.0$), where nominal noise exceeds signal norms by an order of magnitude.

DCPE cannot guarantee both confidentiality and utility simultaneously. These trajectories demonstrate that the DCPE defensive parameter surface is empty. A secure operational point requires high utility ($U \approx 100\%$) alongside low leakage ($A, \Phi_{\text{num}} \approx 0\%$). However, Figure 8 shows that no such window exists: 1) Any parameter choice that preserves acceptable search utility ($\beta \leq 1.0$) guarantees near-perfect identity recovery ($A \geq 96.2\%$) and near-lossless metric reconstruction ($\Phi_{\text{num}} \geq 0.97$); 2) Suppressing Attack B by increasing noise to $\beta = 4.0$ destroys search utility while leaving Attack A completely unimpeded; 3) Suppressing Attack A requires scaling noise past $\beta = 20.0$, where legitimate retrieval recall drops below 0.1%, and the database ceases to function as a searchable index. Parameter tuning cannot decouple searchability from snapshot recoverability.

4.5. Robustness in Non-Ideal Settings

In practical deployments, an adversary may face non-ideal conditions, such as noisy candidate corpora, incomplete database coverage, or hidden embedding models. This section evaluates the resilience of our attacks under these realistic constraints.

4.5.1. Robustness to Candidate Distractors. In real-world leaks, an adversary rarely possesses a clean, 1-to-1 candidate set. To evaluate whether unrelated documents degrade matching accuracy, we dilute the 50,000-record MS MARCO candidate pool with out-of-domain text from AG News and DBpedia-14, expanding the candidate set up to 195,847 passages. As shown in Table 4, Attack A maintains an exact 100.00% re-identification rate across all distractor sizes. This resilience occurs because semantically unrelated distractor embeddings reside in distant regions of the vector space, creating angular margins that prevent them from competing with the true candidate embeddings during cosine matching. On de-duplicated datasets from other text domains, Attack A achieves 99.59%–100.00% re-identification, and Attack B recovers embeddings with > 0.997 cosine similarity.

TABLE 4. ATTACK PERFORMANCE ACROSS CANDIDATE DISTRACTORS, TEXT DOMAINS, AND DATA MODALITIES ($\beta = 0.3$).

Dataset / Condition	Records N	Candidates M	Attack A (Re-ID)	Attack B (Cosine)
Candidate Corpus Dilution (MS MARCO)				*m=1*
Zero Distractors	50,000	50,000	100.00%	0.9973
+100K	50,000	149,608	100.00%	0.9973
+150K	50,000	195,847	100.00%	0.9973
Cross-Domain Text Benchmarks				
AG News	25,000	49,613	100.00%	0.9975
DBpedia-14	25,000	49,995	100.00%	0.9974
SNLI	13,754	13,754	99.59%	0.9987

Note: Attack A matches ciphertexts against the M candidates. Attack B is independent of the candidate pool.

An adversary might only be able to harvest candidates for a fraction of the private database. To evaluate this partial-knowledge setting, we vary the candidate coverage ratio $\kappa = |\widehat{D} \cap D|/|D|$ from 0.05 to 1.0. As depicted in Figure 9, across five of the six datasets (MS MARCO, AG News, DBpedia-14, COCO, and CIFAR-10), the whole-database recovery rate scales strictly proportionally with candidate coverage: $A_{\text{all}} \approx \kappa \cdot A_{\text{cov}}$. For instance, on MS MARCO, when candidate text covers only 25% of the database ($\kappa = 0.25$), the adversary successfully de-anonymizes 24.28% of the entire database, while the accuracy on the covered subset (A_{cov}) remains high at 97.10%. Partial auxiliary knowledge therefore yields a predictable, linear proportion of recovered records rather than causing the attack to fail.

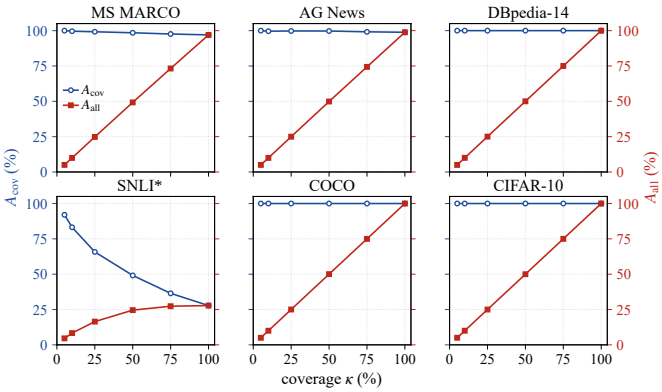

Figure 9. Impact of partial candidate coverage κ.

TABLE 5. COMPUTATIONAL COST COMPARISON: ENCRYPTION VS. RECOVERY ($s = 3$, $\beta = 1.0$)

Dataset	Records N	Encryption (t_{enc})	Attack B (t_{B})	Ratio ($t_{\text{B}}/t_{\text{enc}}$)	Attack A (t_{A})
MS MARCO	50,000	1.51 ms	0.47 ms	0.31×	0.21 s
AG News	25,000	0.67 ms	0.25 ms	0.37×	0.06 s
DBpedia-14	25,000	0.68 ms	0.28 ms	0.41×	0.06 s
SNLI	13,754	0.30 ms	0.24 ms	0.79×	0.01 s
COCO 2017	4,990	0.11 ms	0.14 ms	1.30×	1.40 ms
CIFAR-10	10,000	0.11 ms	0.13 ms	1.20×	2.28 ms
SIFT-1M	1,000,000	5.16 ms	1.02 ms	0.22×	27.69 s

The sole outlier to this linear trend occurs on SNLI (Figure 9, bottom-left), where A_{cov} declines from 92.40% down to 27.82% as κ reaches 1.0, causing A_{all} to stay around 27.82%. This behavior is directly attributable to the sentence redundancy inherent in the SNLI corpus, in which identical premise sentences are repeatedly paired with multiple distinct hypotheses across thousands of samples. For low-coverage, candidate collisions remain rare, allowing high per-covered precision. However, as the candidate corpus expands towards full coverage, identical and near-duplicate candidate embeddings fill up the search space, causing greedy row-wise matching to degenerate into widespread collisions. As established in § 4.2 and Table 3, this limitation is purely an artifact of independent greedy selection. Resolving these conflicts globally via the Hungarian algorithm restores the database-level recovery rate to 99.06%.

4.5.2. Impact of a Secret Model. A potential defensive countermeasure against Attack A is *model secrecy*, keeping the victim's embedding model private so the adversary cannot generate candidate embeddings using the exact encoder $\mathcal{M}$. We evaluate whether model secrecy can thwart snapshot recovery by simulating an adversary who only has access to a surrogate encoder (BGE-base) to attack a target database embedded with GTR-T5 ($d = 768$).

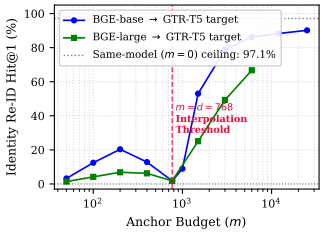

Figure 10. Cross-model transfer attack using a surrogate BGE-base encoder against a target GTR-T5 index ($d = 768$). While zero-anchor transfer fails ($m = 0$), acquiring anchors allows linear alignment, recovering up to 90.15% of records at $m \approx 32d$.

As shown in Figure 10, with zero known pairs ($m = 0$), cross-model re-identification yields 0.00%, showing that representation spaces across different model families are not directly aligned out of the box. However, when the adversary acquires anchor pairs $\{(p_a, c_a)\}_{a=1}^{m}$, they can train a linear projection from the surrogate space to the target space via least squares. Performance reaches 20.35% at $m = 200$, dips to 1.80% near the interpolation threshold ($m = d = 768$) due to noise overfitting on a square linear system, and rebounds to 90.15% once $m \approx 32d$. This result demonstrates that model secrecy is not an absolute defense. It merely raises the adversary's anchor requirement.

TABLE 6. INVARIANCE TO SECRET SCALE s ON MS MARCO ($d = 768$, $\beta = 1.0$)

Secret Scale s	Deployment Noise	Recon. Cosine	Re-ID
1.0	24.2%	0.9709	100.00%
3.0	72.7%	0.9709	100.00%
10.0	242.2%	0.9709	100.00%
100.0	2,422.4%	0.9709	100.00%
1000.0	24,224.2%	0.9709	100.00%

4.6. Computational Overhead of Recovery

A secure encryption scheme should be efficient for legitimate users while being computationally expensive for an adversary to invert. To evaluate this overhead, Table 5 compares the wall-clock execution time of database encryption against our two recovery attacks on a single workstation with an NVIDIA RTX 5090 GPU. Recovering the database via Attack B is actually faster than encrypting it. On MS MARCO ($N = 50{,}000, d = 768$), database encryption takes 1.51 ms, whereas Attack B estimates the scale and reconstructs all 50,000 embeddings in just 0.47 ms, using only 31% of the encryption time ($t_{\text{B}}/t_{\text{enc}} = 0.31$). On SIFT-1M ($N = 1{,}000{,}000$), encryption takes 5.16 ms, while Attack B reconstructs the entire database in 1.02 ms ($t_{\text{B}}/t_{\text{enc}} = 0.22$). Encryption requires sampling random vectors uniformly from high-dimensional spheres, which incurs overhead for pseudo-random number generation. In contrast, Attack B requires only vectorized dot products and scalar division, which execute at peak GPU memory throughput. Table 5 also shows the runtime of Attack A. Because candidate matching relies on parallel matrix multiplication, matching the 50,000 MS MARCO ciphertexts against 96,239 candidate passages takes 0.21 seconds. On SIFT-1M, matching one million ciphertexts against one million candidates completes in 27.69 seconds on a single GPU.

DCPE provides no computational protection. An attacker with a leaked snapshot can de-anonymize records or reconstruct embeddings in milliseconds to seconds.

4.7. Scale Sensitivity

We assume that increasing s scales up the injected noise, thereby overwhelming the underlying semantic signal. This intuition fails algebraically because scaling impacts the signal and perturbation by the identical factor $c_i/s = p_i + \lambda_i/s$. The secret scale s identically cancels out during scale normalization, leaving the attacker-facing normalized perturbation ratio $\rho = \|\lambda\|_2/(s\|p\|_2) = \beta/(4\|p\|_2)$ unchanged. Table 6 provides empirical confirmation: increasing the secret scale s from 1 to 1,000 inflates the nominal deployment noise by three orders of magnitude (from 24.2% to 24,224.2%). Across this sweep, every metric remains strictly constant. Attack A re-identification stays at 100%, numerical reconstruction fidelity remains at 0.9709 cosine similarity.

5. Evaluation of Existing Defenses

A concurrent study by Fuchsbauer et al. [19] proposed several defensive techniques to mitigate leakage in distance-preserving vector search: (1) Rotate-and-Perturb (RAP), (2) secret projection, and (3) ciphertext shuffling. These defenses follow two distinct approaches. RAP and secret projection modify the vector coordinates while preserving relative distance comparisons for search. In contrast, ciphertext shuffling leaves the vector values unchanged and permutes the database rows to prevent known-plaintext alignment. Since this violates our basic assumption, we only evaluate RAP and secret projection. Besides, CAPRISE [17] was also proposed as an enhanced distance-preserving scheme. CAPRISE aims to thwart embedding analysis by asymmetrically allocating perturbation noise. In this section, we describe the specific design and evaluate each defense with our attack metrics: Φ_{id}, Φ_{num}, and Φ_{util}.

5.1. Rotate-and-Perturb

Rotate-and-Perturb (RAP) replaces the shared secret scale s in standard DCPE with a secret orthogonal matrix $M \leftarrow \mathcal{O}(d)$, where $\mathcal{O}(d)$ denotes the orthogonal group. For a normalized plaintext embedding $p_i \in \mathbb{R}^d$, the ciphertext is generated as $c_i = Mp_i + \eta_i$, where η_i is perturbation noise sampled uniformly from a d-dimensional ball of radius $\beta/4$. The core motivation of RAP is to eliminate the global scalar parameter s. Because orthogonal matrices are isometric ($\|Mx\|_2 = \|x\|_2$ and $\langle Mx, My\rangle = \langle x, y\rangle$), RAP preserves Euclidean distances and cosine similarities without modifying the relative geometry between vectors, thereby retaining native search performance.

Without anchors, Attack A fails completely. The unknown rotation disrupts angular alignment with the candidate corpus. Stopping zero-anchor matching is RAP's sole defensive benefit. Given m anchor pairs, the adversary recovers M via the linear system $C_{\text{anchor}} = P_{\text{anchor}}M^\top + E$. Because text embeddings have ill-conditioned covariance matrices (condition number $\approx 2.6 \times 10^3$), ordinary least squares overfits noise (cosine 0.78). Regularized ridge regression resolves this, boosting estimation cosine fidelity to 0.98. Once $\widehat{M}$ is obtained, inverting the rotation ($\widehat{p}_i = c_i\widehat{M}$) directly reuses the baseline DCPE recovery pipeline. At $m = d$, Attack B achieves $\Phi_{\text{id}} = 99.95\%$, $\Phi_{\text{num}} = 0.963$, and index utility $\Phi_{\text{util}} = 92.32\%$. Thus, RAP only raises the entry barrier from $m = 1$ to $m \approx d$, without lowering the attack ceiling. As shown in Figure 11, across six datasets, Φ_{id} reaches 90% at 10^2 anchors ($m \ll d$), while Φ_{num} requires $m \approx d$ anchors.

5.2. Secret Projection

Secret projection introduces irreversible dimensionality reduction to prevent full recovery of coordinates. Let $R \in \mathbb{R}^{m_{\text{proj}} \times d}$ be a secret projection matrix with orthonormal rows, where the projection output dimension $m_{\text{proj}} < d$. The

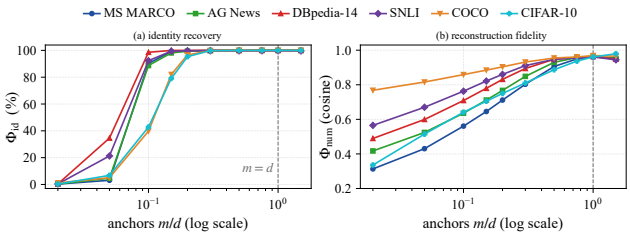

Figure 11. Recovery performance against RAP.

combined projected RAP scheme encrypts an embedding p_i as $c_i = R(Mp_i) + \eta_i$, where $c_i \in \mathbb{R}^{m_{\text{proj}}}$ and η_i is perturbation noise sampled uniformly from a d-dimensional ball of radius $\beta/4$.

Because $\text{rank}(R) = m_{\text{proj}} < d$, the linear map has a non-trivial null space. The components of p_i orthogonal to the row space of R are permanently lost. Even if an adversary determines both R and M, they can at most reconstruct the projection of p_i (as shown in Figure 12), theoretically bounding coordinate reconstruction fidelity by $\Phi_{\text{num}} \approx \sqrt{m_{\text{proj}}/d}$. The legitimate recall of the system will also decrease. Secret Projection cannot be adopted as an effective defense.

5.3. Asymmetric Noise Allocation (CAPRISE)

CAPRISE [17] modifies the standard Scale-and-Perturb construction by dividing the noise budget β asymmetrically between stored embeddings and runtime queries. For a secret scale s, stored embeddings and queries are encrypted as:

$$c_i = sp_i + \lambda_i^{\text{db}}, \quad \|\lambda_i^{\text{db}}\|_2 \le r_{\text{db}}, \tag{22}$$

$$c_q = sq + \lambda_q^{\text{query}}, \quad \|\lambda_q^{\text{query}}\|_2 \le r_{\text{query}}, \tag{23}$$

where the database noise radius is intentionally larger than the query noise radius (typically $r_{\text{db}} = 3s\beta/8$ and $r_{\text{query}} = s\beta/8$, yielding a 3:1 ratio). The rationale behind CAPRISE is that allocating most of the error budget to the stored database will obscure inter-database metric relationships,

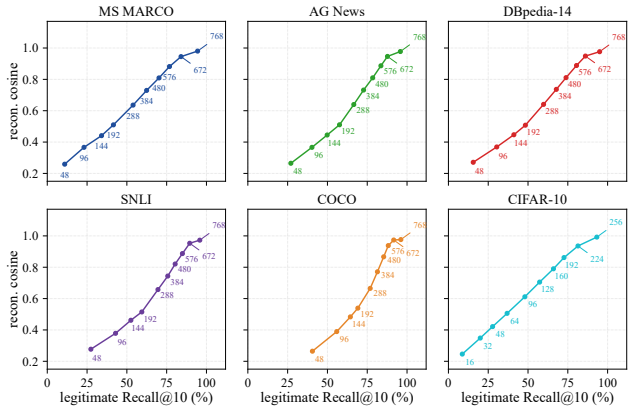

Figure 12. Φ_{num} vs. legitimate Recall. There is no projection dimension that enables a high recall while keeping Φ_{num} below 50%.

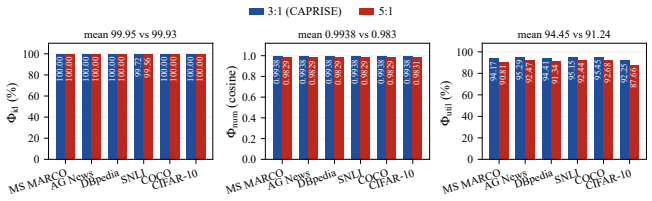

Figure 13. Φ_{id}, Φ_{num}, and Φ_{util} with different noise ratio.

thereby hindering geometric inference from static snapshots while preserving search accuracy through low query noise.

We evaluate whether this asymmetric noise split provides meaningful protection against static snapshot recovery. When a snapshot adversary evaluates a stored ciphertext c_i against an unencrypted candidate embedding $e_j = \mathcal{M}(\widehat{t}_j)$, the cosine similarity satisfies:

$$\cos(c_i, e_j) = \frac{\langle p_i + \lambda_i^{\mathrm{db}}/s,\ e_j \rangle}{\|p_i + \lambda_i^{\mathrm{db}}/s\|_2 \|e_j\|_2}. \tag{24}$$

Equation (24) reveals that the query-side noise radius r_{query} does not enter the adversary's matching computation. The adversary evaluates candidates directly in the unencrypted space. We evaluate Φ_{id}, Φ_{num}, and Φ_{util} under the published 3:1 ratio as well as 5:1 asymmetric allocations. Unfortunately, Figure 13 shows that as the ratio increases, this approach can only provide minimal defense.

5.4. Remarks on Ciphertext Shuffling

Ciphertext shuffling applies a secret permutation π to the rows of the database matrix C. Unlike RAP and secret projection, shuffling does not modify vector coordinates. Instead, it alters the threat model: it targets adversaries that rely on a known row-by-row correspondence between plaintexts and ciphertexts.

While shuffling disrupts static row-aligned pairing, we observe that this protection relies on the assumption of a purely passive, read-only observer. In operational RAG deployments where clients have data-ingestion privileges (such as multi-tenant vector databases), an adversary can reconstruct paired anchors by inserting d known documents and computing the differential snapshot between the states. Consequently, row shuffling does not reduce the geometric leakage of the ciphertexts themselves.

6. Discussion and Limitations

6.1. The Impossibility of Distance-Preserving

In this paper, we arrive at a fundamental conclusion: **any cryptographic scheme that preserves relative distance comparisons to support native approximate nearest-neighbor search inherently exposes the database snapshot to geometric reconstruction.**

This vulnerability is an inevitable consequence of the geometry preserved for search. Transformations that preserve relative distances, such as scalar scaling, coordinate permutations, and orthogonal rotations, leave pairwise inner products and angular topology invariant. Because vector retrieval engines rely on this exact geometry to rank nearest neighbors, an adversary can exploit the same metric relations to de-anonymize records (Attack A) or reconstruct embeddings (Attack B). Consequently, legitimate retrieval utility (U) and adversarial recovery (A) are fundamentally consistent, and increasing isotropic perturbation noise degrades service utility long before it prevents recovery.

6.2. Practical Implications for Private RAG

Based on the Attack evaluation results, we offer several suggestions for designing private vector databases.

1) Leaked DCPE snapshots must be treated as direct plaintext disclosures: When an auxiliary candidate corpus is publicly crawlable, Attack A can de-anonymize encrypted records at scale in seconds on a single GPU. More critically, even when the private database is entirely non-enumerable, Attack B breaches embeddings confidentiality under the minimal assumption of acquiring as few as a single known plaintext-ciphertext anchor. With this single anchor, an adversary estimates the secret scale and reconstructs embeddings with high cosine similarity. Furthermore, the adversary can fully restore retrieval functionality, enabling inversion models to decode sensitive source text and image assets from the recovered embeddings. Because either crawlable public domain text or minimal legal users interaction is sufficient to compromise the database, distance-preserving encryption offers negligible confidentiality.

2) Defensive countermeasures based on DCPE fail to provide practical protection: First, adjusting system parameters cannot secure the database. The secret scale s cancels out under cosine normalization, a single anchor ($m = 1$) is enough to estimate s accurately, and increasing the perturbation noise β destroys search recall long before it suppresses re-identification. Second, keeping the embedding model secret is also insufficient. While hiding the encoder prevents zero-anchor matching, an adversary using a public surrogate model can train a linear projection with anchor pairs and recover over 90% of records. More broadly, further modifications such as secret rotations, projections, and asymmetric noise allocations also fail to rescue the design. As long as an encryption scheme preserves relative distance comparisons directly over ciphertexts, the system is not confidential. If the system maintains high retrieval recall, it preserves enough distance relations needed for our attacks, leaving the database vulnerable to near-perfect recovery. Conversely, if an operator injects massive noise or removes dimensions to prevent recovery, the retrieval recall drops so severely that the search service becomes unusable.

3) Achieving confidentiality requires abandoning direct distance preservation: Because our attacks demonstrate that preserving relative distance comparisons exposes the database to snapshot recovery, a secure system must eliminate the metric channel that enables candidate alignment. One promising direction is to conceal the distance evaluations. Systems can compute vector similarities inside trusted

hardware enclaves, evaluate encrypted inner products using homomorphic encryption, or traverse a search database through oblivious RAM. While these cryptographic primitives introduce considerable latency, recent optimizations in batched private retrieval and hardware acceleration are actively narrowing this overhead.

6.3. Limitations and Future Work

We identify four primary limitations in our study that represent promising directions for future research. First, Attack A relies on semantic overlap between the auxiliary candidate corpus and the target database. Although our experiments demonstrate that database-level recovery scales linearly with candidate coverage, our threat model does not account for the monetary cost of corpus collection or API embedding generation. Second, our primary attacks assume access to the public foundation encoder. Although cross-model alignment demonstrates that surrogate models can bridge this gap when anchors are present, evaluating heavily customized models with unknown architectures remains an open problem. Finally, Attack B requires correctly paired plaintext-ciphertext anchors. While multi-tenant vector databases enable anchor fetching through differential snapshots, the complexity of fetching varies across deployment environments.

7. Related Work

7.1. Property-Preserving and Encrypted Vector Search

Property-preserving encryption allows untrusted servers to evaluate search predicates directly over ciphertexts without decryption. While early order-preserving schemes [20], [21] argued security on isolated ciphertexts, subsequent leakage-abuse attacks [22], [23] and volumetric analyses [24], [25], [26] proved that preserving relations inevitably leaks dataset-level geometry. Order-revealing primitives were similarly shown to expose unexpected statistical structure [27], [28].

Modern systems avoid these leakages by employing heavier cryptographic primitives, such as homomorphic evaluation and oblivious memory (in the area of vector search, Compass [8], Pacmann [7], Panther [6], Tiptoe [5], and Wally [29]). However, these constructions incur severe performance penalties, running several orders of magnitude slower than plaintext search [30]. To maintain real-time throughput, commercial RAG deployments widely converged on Distance-Comparison-Preserving Encryption and Scale-and-Perturb constructions [13], [17], [9]. This paper demonstrates that DCPE's efficient encrypted vector search comes at the cost of confidentiality. An adversary who can observe the encrypted database snapshot can practically reconstruct the original data.

7.2. Embedding Inversion Across Modalities

Dense embeddings preserve rich semantic topology. Embedding-inversion techniques [2], [3], [31], [32] demonstrate that text and images can be reconstructed from clean embeddings with high fidelity, transferring across encoders without querying victim models [4], [33]. However, this literature assumes an adversary who already holds plaintext vectors. We show that distance-preserving ciphertexts can be inverted back to the natural embedding, after which off-the-shelf inversion models can decode human-readable records without adversarial retraining.

7.3. Membership Inference Attacks

Membership inference attacks (MIA) evaluate privacy by determining whether a specific record contributed to a target dataset or model [34], [35], with recent variants targeting RAG retrieval interfaces [36], [37]. The concurrent work by Fuchsbauer et al. [19] formulates membership inference against DCPE snapshots under public encoders. They demonstrate that second-moment statistics allow determining the presence of a record without anchors, prove theoretical bounds for Rotate-and-Perturb on uniform spherical distributions, and suggest secret projection as a mitigation.

Our work addresses a fundamentally different threat model and objective. While membership inference resolves a binary decision problem (determining whether an individual record is present), our attacks address the more catastrophic problem of database reconstruction: Attack A recovers up to 100.0% of record identities without keys, while Attack B reconstructs continuous vector values at 0.997 cosine fidelity from a single anchor, restoring a functional surrogate search index. Furthermore, while they suggest several defensive countermeasures for DCPE, our attacks on real-world datasets show that these enhancements fail to provide practical protection.

8. Conclusion

Distance-preserving encryption is widely adopted to enable efficient nearest-neighbor search over encrypted vector databases while preserving relative distance relations across ciphertexts. In this work, we evaluate whether this preserved geometry exposes static database snapshots to recovery. We show that an adversary without secret keys can directly align leaked ciphertexts with public candidate embeddings, achieving large-scale identity de-anonymization. Furthermore, obtaining as few as a single known plaintext-ciphertext anchor allows the adversary to estimate the secret scale and reconstruct embeddings with 0.997 cosine similarity, fully restoring the search index. Across extensive experiments, existing defensive countermeasures fail to prevent recovery without damaging retrieval recall. In conclusion, leaked distance-preserving snapshots should be treated as plaintext disclosures.

References

[1] J. Kim, W. Guo, and D. Song, "SoK: Attack and defense landscape of agentic AI systems," in *Proceedings of the 35th USENIX Security Symposium (USENIX Security)*, 2026.

[2] C. Song and A. Raghunathan, "Information leakage in embedding models," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2020, pp. 377–390.

[3] J. X. Morris, V. Kuleshov, V. Shmatikov, and A. M. Rush, "Text embeddings reveal (almost) as much as text," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 12 448–12 460.

[4] Y.-H. Huang, Y. Tsai, H. Hsiao, H.-Y. Lin, and S.-D. Lin, "Transferable embedding inversion attack: Uncovering privacy risks in text embeddings without model queries," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, pp. 4193–4205.

[5] A. Henzinger, E. Dauterman, H. Corrigan-Gibbs, and N. Zeldovich, "Private web search with tiptoe," in *Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP)*, 2023, pp. 396–416.

[6] J. Li, Z. Huang, M. Zhang, C. Hong, J. Liu, T. Wei, and W. Chen, "Panther: Private approximate nearest neighbor search in the single server setting," in *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2025, pp. 365–379.

[7] M. Zhou, E. Shi, and G. Fanti, "Pacmann: Efficient private approximate nearest neighbor search," in *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[8] J. Zhu, L. Patel, M. Zaharia, and R. A. Popa, "Compass: Encrypted semantic search with high accuracy," in *19th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. USENIX Association, 2025, pp. 915–938.

[9] Y. Liu, Y. Zhang, J. Xie, H. Li, J. X. Yu, and J. Cui, "Privacy-preserving approximate nearest neighbor search on high-dimensional data," in *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, 2025, pp. 3017–3029.

[10] IronCore Labs, "Cloaked AI product documentation," Online documentation, https://ironcorelabs.com/products/cloaked-ai/, accessed September 2026.

[11] Alibaba Cloud, "Deploy confidential RAG applications on heterogeneous confidential computing instances," Online documentation, https://help.aliyun.com/zh/egs/use-cases/deploy-confidential-rag-applications-on-heterogeneous-confidential-computing-instances.

[12] Volcano Engine, "Jeddak AICC: Vector encryption service," Online documentation, https://docs.volcengine.com/docs/JeddakAICC/VectorEncryptionService.

[13] G. Fuchsbauer, R. Ghosal, N. Hauke, and A. O'Neill, "Approximate distance-comparison-preserving symmetric encryption," in *Security and Cryptography for Networks (SCN)*, 2022, pp. 117–144.

[14] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 824–836, 2020.

[15] S. J. Subramanya, F. Devvrit, H. V. Simhadri, R. Krishnawamy, and R. Kadekodi, "DiskANN: Fast accurate billion-point nearest neighbor search on a single node," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[16] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu, K. Yu, Y. Yuan, Y. Zou, J. Deng, R. Hu, T. Liu, J. Xu, C. Li, S. Lin, Y. Li, and X. Wang, "Milvus: A purpose-built vector data management system," in *Proceedings of the 2021 International Conference on Management of Data (SIGMOD)*, 2021.

[17] H. Ye, J. Guo, Z. Liu, and K.-Y. Lam, "Efficient privacy-preserving retrieval augmented generation with distance-preserving encryption," in *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, 2025, pp. 668–676.

[18] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[19] G. Fuchsbauer, C. Luck, A. O'Neill, and I. Rapko, "Membership inference in privacy-preserving vector search," Cryptology ePrint Archive, Paper 2026/2570, 2026. [Online]. Available: https://eprint.iacr.org/2026/2570

[20] A. Boldyreva, N. Chenette, Y. Lee, and A. O'Neill, "Order-preserving symmetric encryption," in *Advances in Cryptology – EUROCRYPT*, 2009, pp. 224–241.

[21] O. Pandey and Y. Rouselakis, "Property preserving symmetric encryption," in *Advances in Cryptology – EUROCRYPT*, 2012.

[22] D. Cash, P. Grubbs, J. Perry, and T. Ristenpart, "Leakage-abuse attacks against searchable encryption," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2015, pp. 668–679.

[23] M. Naveed, S. Kamara, and C. V. Wright, "Inference attacks on property-preserving encrypted databases," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2015, pp. 644–655.

[24] G. Kellaris, G. Kollios, K. Nissim, and A. O'Neill, "Generic attacks on secure outsourced databases," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 1329–1340.

[25] P. Grubbs, M.-S. Lacharité, B. Minaud, and K. Paterson, "Pump up the volume: Practical database reconstruction from volume leakage on range queries," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2018, pp. 315–331.

[26] M.-S. Lacharité, B. Minaud, and K. G. Paterson, "Improved reconstruction attacks on encrypted data using range query leakage," in *2018 IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 297–314.

[27] F. B. Durak, T. M. DuBuisson, and D. Cash, "What else is revealed by order-revealing encryption?" in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 1155–1166.

[28] P. Grubbs, K. Sekniqi, V. Bindschaedler, M. Naveed, and T. Ristenpart, "Leakage-abuse attacks against order-revealing encryption," Cryptology ePrint Archive, Paper 2016/895, 2016, https://eprint.iacr.org/2016/895.

[29] H. Asi, F. Boemer, N. Genise, M. H. Mughees, T. Ogilvie, R. Rishi, K. Talwar, K. Tarbe, A. Wadia, R. Zhu, and M. Zuliani, "Wally: Batched private nearest neighbor search at scale," arXiv preprint arXiv:2406.06761, 2024, https://arxiv.org/abs/2406.06761.

[30] A.-M. Kermarrec, R. Pires, M. Randl, and M. de Vos, "A unified benchmark for privacy-preserving vector search," arXiv preprint arXiv:2608.01192, 2026, https://arxiv.org/abs/2608.01192.

[31] H. Li, M. Xu, and Y. Song, "Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 14 022–14 040.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[33] R. Jha, C. Zhang, V. Shmatikov, and J. X. Morris, "Harnessing the universal geometry of embeddings," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025, pp. 45 963–45 987.

[34] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy (S&P)*, 2017.

[35] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership inference attacks from first principles," in *IEEE Symposium on Security and Privacy (S&P)*, 2022.

[36] S. Zeng, J. Zhang, P. He, Y. Xing, Y. Liu, H. Xu, J. Ren, S. Wang, D. Yin, Y. Chang, and J. Tang, "The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG)," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 4505–4524.

[37] M. Anderson, G. Amit, and A. Goldsteen, "Is my data in your retrieval database? membership inference attacks against retrieval augmented generation," in *International Conference on Information Systems Security and Privacy (ICISSP)*, 2025, pp. 474–485.

Appendix A. Ethics Considerations

In this paper, we evaluate the algorithmic foundations of distance-preserving encryption and its existing security hardening in an offline experimental environment. All evaluations were conducted using public benchmark datasets and open foundation encoders without accessing private user records. We did not attack, query, or disrupt any live commercial systems, and our study caused no operational interruptions or financial losses to any organization.

Our primary motivation in disclosing these findings is to demonstrate the fundamental security limits of distance-preserving vector search, enabling system operators to recognize these exposures in a timely manner and transition to genuinely secure cryptographic primitives. To support this transition, we also discuss several constructive defensive directions, including adopting homomorphic encryption and oblivious RAM to protect sensitive embeddings. Finally, in accordance with established responsible disclosure principles, we have formally notified the relevant companies and organizations via official communication channels to inform them of our findings and provide technical context on these exposures.

Appendix B. Theoretical guarantee of Attack A

B.1. Row-wise top-1 correctness under angular separation

Let $p_i \in \mathbb{R}^d$ be a unit-norm plaintext embedding, let $c_i = sp_i + \lambda_i$ with $\|\lambda_i\|_2 \le r = s\beta/4$, and let $\hat{D}$ be an enumerable candidate corpus with embeddings $E = \{\mathcal{M}(\hat{t}_j)\}$. Let $e_1 = \mathcal{M}(t_i)$ be the true candidate, and define the angular margin

$$\Delta\theta_{\min}(p_i) \;=\; \min_{e \in E \setminus \{e_1\}} \left(\arccos\langle p_i, e\rangle - \arccos\langle p_i, e_1\rangle\right),$$

with $\Delta\theta_{\min}(p_i) > 0$ whenever t_i has a unique nearest candidate. If

$$\arcsin\left(\frac{\beta}{4}\right) \;<\; \frac{1}{2}\,\Delta\theta_{\min}(p_i),$$

then $\cos(c_i, e_1) > \cos(c_i, e)$ for every $e \in E \setminus \{e_1\}$. Consequently, row-wise top-1 matching recovers t_i.

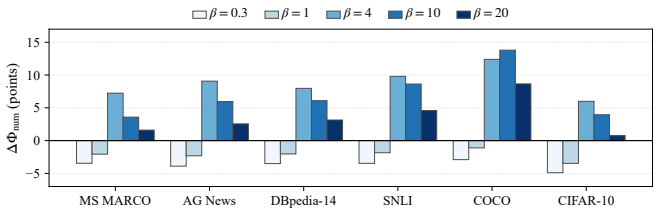

Figure 14. $\Delta\Phi_{\text{num}}$ in points after post-recovery k-NN smoothing.

B.2. Proof of Proposition 1

Since cosine is invariant to positive scaling, write $p_i' = p_i + \lambda_i/s$, so that $\cos(c_i, e) = \cos(p_i', e)$ for every candidate e. The perturbation satisfies

$$\|p_i' - p_i\|_2 = \|\lambda_i\|_2/s \le r/s = \beta/4.$$

Let $\phi = \angle(p_i, p_i')$. Because $\|p_i\|_2 = 1$,

$$\sin\phi \le \frac{\|p_i' - p_i\|_2}{\|p_i\|_2} \le \frac{\beta}{4},$$

hence $\phi \le \arcsin(\beta/4) < \Delta\theta_{\min}(p_i)/2$. For any $e \in E \setminus \{e_1\}$, the triangle inequality for angles gives $\angle(p_i', e_1) \le \angle(p_i, e_1) + \phi \le \angle(p_i, e) - \Delta\theta_{\min}(p_i) + \phi < \angle(p_i, e) - \phi \le \angle(p_i', e)$, where the last inequality uses $\angle(p_i', e) \ge \angle(p_i, e) - \phi$. Therefore p_i' is strictly closer in angle to e_1 than to any other candidate, and since cos is strictly decreasing on $[0, \pi]$, $\cos(c_i, e_1) > \cos(c_i, e)$. Row-wise top-1 matching therefore selects e_1, recovering t_i.

Appendix C. Impact of k-NN Denoising

In §3.2.3, we formulated an unsupervised post-recovery k-nearest-neighbor (k-NN) smoothing step to mitigate residual isotropic noise. To evaluate its efficacy under varying defensive configurations, we conduct an ablation study across five distinct perturbation budgets $\beta \in \{0.3, 1, 4, 10, 20\}$, ranging from standard operational noise to extreme noise regimes in which legitimate utility collapses. Figure 14 reports the delta in numerical reconstruction fidelity $\Delta\Phi_{\text{num}} = \left(\Phi_{\text{num}}(\widehat{P}') - \Phi_{\text{num}}(\widehat{P})\right) \times 100$ (points). Across all six different types of datasets, Figure 14 demonstrates a consistent, three-phase behavior governing local denoising:

1) Low-Noise. When perturbation is mild, raw Attack B reconstruction already achieves near-lossless directional fidelity ($\Phi_{\text{num}} > 0.97$). In this situation, the perturbation variance is negligible relative to the natural semantic margins separating distinct clusters on the data manifold. Averaging across the k nearest neighbors inevitably blends adjacent, semantically distinct embeddings, introducing an oversmoothing distortion. Consequently, smoothing causes a modest decline of 1.5 to 4.9 points across all datasets. The direct linear estimate is already optimal, making denoising unnecessary.

2) High-Noise. When the defender increases the noise budget to prevent recovery, the noise variance overwhelms fine-grained embedding coordinates (e.g., dropping raw fidelity to $\approx 71.45\%$ on MS MARCO at $\beta = 4$). In this noise-dominated situation, the benefits of local smoothing peak decisively. Because the isotropic perturbations λ_i are zero-mean and independent across stored records ($\mathbb{E}[\lambda_i \lambda_j^\top] = 0$ for $i \neq j$), averaging across k neighbors suppresses the random noise variance by roughly $1/k$, while the underlying manifold coordinates reinforce one another. At $\beta = 4$, fidelity increases by $+9.8$ points on SNLI, $+9.1$ points on AG News, $+7.9$ points on DBpedia-14, and $+7.2$ points on MS MARCO. At $\beta = 10$, gains remain exceptionally robust ($+3.6$ to $+8.6$ points). On CIFAR-10, smoothing achieves up to a $+6.0$ points gain. On COCO (CLIP embeddings), fidelity improves by $+12.4$ points at $\beta = 4$ and peaks at $+13.7$ points at $\beta = 10$.

3) Extreme Noise. Under extreme noise, the perturbation magnitude dwarfs the raw embedding radius by an order of magnitude. At this level, severe displacement due to noise perturbs a vector across cluster boundaries, causing its retrieved Euclidean nearest neighbors in the ciphertext space to include incorrect records. This noisy neighbor sampling decreases the averaging effect, causing the net fidelity delta to taper off ($\Delta\Phi_{\mathrm{num}}$ remains positive across datasets).