ORTHOGONAL CIPHER TOKEN ENCRYPTION: EQUIVARIANT UTILITY AND INVARIANT LEAKAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

We study orthogonal cipher token encryption, in which a client rotates latent token vectors with a secret orthogonal transform while an equivariant server computes directly on the rotated representations. Because the server is constructed from rotation-equivariant operations, the same server can process arbitrary client keys without key-specific retraining, and the client can decrypt the returned features before ordinary task decoding with little utility loss in the evaluated settings. We show, however, that direct no-key decoder failure is not sufficient evidence of representation privacy. Under closed-reference auxiliary knowledge, orthogonal invariants such as norms, pairwise similarities, and Gram matrices remain observable and can enable strong token-, instance-, and class-level leakage. Client pre-normalization suppresses norm fingerprints but leaves Gram structure intact. In a controlled Food-101 experiment, salted/input-protected rotation substantially reduces Gram retrieval while retaining authorized utility, whereas noise-only perturbation reduces leakage largely by destroying utility. These results position orthogonal ciphering as a key-agnostic equivariant utility interface rather than a privacy mechanism by itself; additional mechanisms are required to suppress invariant leakage.

1 INTRODUCTION

Learned systems increasingly exchange intermediate representations rather than raw inputs. A client may embed an image, text sequence, or patch grid locally, send the resulting tokens to a server, and receive task predictions from a model that never sees the original data. Orthogonal cipher token encryption asks whether this interface can be made useful under a secret rotation. If each vector token $x \in \mathbb{R}^d$ is encrypted as xR for a secret $R \in O(d)$, then an $O(d)$-equivariant server can satisfy $f(xR) = f(x)R$. The client decrypts with $R^\top$, while a no-key decoder should fail on the rotated representation.

This paper studies that privacy boundary across small dense image classifiers, equivariant CNNs, text language modeling, COCO captioning, MobileCLIP-S1 vision classifiers, projected Gram-statistic probes, and staged Food-101 noise sweeps. Across these settings, the same pattern appears: authorized encrypted computation can preserve utility, but invariants of the orthogonal action remain visible to the server. Norms and Gram matrices do not require the key, and they can be strong enough for exact retrieval or high-accuracy class prediction. We also evaluate noise and salted input-protected variants as empirical leakage mitigations, not privacy guarantees.

The exact retrieval attacks studied here assume auxiliary knowledge: the adversary has a closed reference set of unencrypted token embeddings, instance features, or sequence/gallery signatures. Thus, exact token and instance recovery should be read as closed-reference leakage rather than unconstrained open-world generation of raw inputs. The use of a shared rotation across tokens is also structurally important: attention-style computations depend on cross-token inner products, and independent per-token rotations would generally destroy those equivariant dot products.

The main contributions of this paper are:

- We audit orthogonal cipher token encryption as an equivariant client–server interface and separate authorized utility from privacy leakage.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- We demonstrate that direct no-key decoder failure is not sufficient evidence of privacy under auxiliary reference sets.
- We formalize and evaluate norm-retrieval and Gram-statistic attackers that exploit invariants preserved by orthogonal rotation.
- We show that pre-normalization is a targeted defense against norm fingerprints but not a general defense against Gram or inner-product leakage.
- We consolidate measured leakage benchmarks across vision and text representation transformations, including closed-reference NN, norm, Gram, and EDNN-style attacks.

2 Related Work

The experiments use standard image, text, and captioning benchmarks: MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky, 2009), WikiText-103 (Merity et al., 2016), and COCO with the common captioning split lineage (Lin et al., 2014; Karpathy & Fei-Fei, 2015). Captioning is evaluated with BLEU, METEOR, ROUGE-L, CIDEr, and SPICE (Papineni et al., 2002; Banerjee & Lavie, 2005; Lin, 2004; Vedantam et al., 2015; Anderson et al., 2016).

Orthogonal and covariant obfuscation methods transform embeddings or model components so inference can proceed over hidden coordinates. Privacy from Symmetry introduces orthogonal obfuscation and ConjFormer, an $O(d)$-equivariant Transformer for LLM inference (Yukhimchuk et al., 2026); AloePri studies covariant obfuscation at industrial scale (Lin et al., 2026); SentinelLMs adapt language models on passkey-encrypted tokenization and embeddings (Mishra et al., 2024). Our work is complementary: it audits invariant leakage channels that remain even when coordinate-space decoding fails.

We compare against external leakage baselines and related split-inference defenses. EDNN attacks obfuscated embedding matrices through differential signatures (Lin et al., 2024); SIMBA, InfoDecom, and Split-and-Denoise study split-inference leakage, noise, and denoising mechanisms (Singh et al., 2024; Deng et al., 2026; Mai et al., 2024). Vector Neurons and VN-Transformer motivate rotation-equivariant neural operations (Deng et al., 2021; Assaad et al., 2022), while homomorphic private inference pursues stronger cryptographic security goals (Zimerman et al., 2024). Our noise variants are empirical invariant-leakage probes, not formal DP or cryptographic guarantees.

3 Methods

3.1 Problem Setting

3.1.1 Candidate Tokens and Latent Sequence

Let

$$\mathcal{T}_{\text{cand}} = \{t_1, \dots, t_N\}, \qquad x_i = E_\theta(t_i) \in \mathbb{R}^d, \qquad X = [x_{j_1}; \dots; x_{j_L}] \in \mathbb{R}^{L \times d}. \tag{1}$$

Here t_i may be a text token, image patch, region, or instance, but the cipher acts only on the latent vector coordinate. A client uses a secret orthogonal key $R \in O(d)$ and sends

$$y_\ell = x_{j_\ell} R, \qquad Y = XR \equiv X_{\text{enc}}. \tag{2}$$

Batch, patch, spatial, or channel axes are implementation bookkeeping; the rule xR is applied independently to each d-dimensional latent vector.

3.1.2 Adversary and Authorized Server

An adversary may observe $y = x_j R$ or $Y = XR$, but not R. Retrieval attacks use an auxiliary reference set

$$\mathcal{A}_{\text{aux}} = \{a_1, \dots, a_M\}, \qquad a_i \in \mathbb{R}^d, \tag{3}$$

such as public token embeddings, gallery features, or Gram-derived reference signatures. Orthogonal rotation hides coordinates but preserves key invariants:

$$\|x_j R\|_2 = \|x_j\|_2, \qquad (x_i R)(x_j R)^\top = x_i x_j^\top. \tag{4}$$

The authorized server receives $Y = XR$ and should emit rotated outputs:

$$F_\phi(XR) \approx F_\phi(X)R. \tag{5}$$

Thus the intended split is: the server computes in encrypted coordinates, the client decrypts, and a no-key adversary is limited except for invariant leakage.

Threat model and claim boundary. The server and attacker are assumed to know the model family, the protection mechanism, and any public auxiliary references used by the benchmark, but not the client's secret key R. The leakage results are therefore not attacks on raw inputs in an unconstrained open world; they are closed-reference tests of what can be inferred from protected representations under auxiliary knowledge. Conversely, authorized utility is not a privacy claim: it only states that an equivariant server can compute on rotated coordinates and return an output the client can decrypt.

3.2 Authorized Encrypted Inference

The server applies F_ϕ directly to X_{enc}:

$$Z_{\text{enc}} = F_\phi(XR), \qquad F_\phi(XR) \approx F_\phi(X)R. \tag{6}$$

The client decrypts and applies the task head:

$$Z_{\text{dec}} = Z_{\text{enc}}R^\top, \qquad Z_{\text{dec}} \approx F_\phi(X), \qquad \hat{y}_{\text{enc}} = H_\psi(Z_{\text{dec}}). \tag{7}$$

The unencrypted prediction is

$$\hat{y}_{\text{plain}} = H_\psi(F_\phi(X)), \qquad \hat{y}_{\text{enc}} = \hat{y}_{\text{plain}} \text{ under exact equivariance.} \tag{8}$$

3.3 Equivariant Server Layer Constraints

The server must avoid operations that mix encrypted coordinates in a non-equivariant way, such as arbitrary dense projections over the encrypted vector axis, coordinate-wise nonlinearities on that axis, or standard LayerNorm over that axis. The admissible operations used here are summarized by

$$\mathcal{L}(X) = AX, \qquad \mathcal{L}(XR) = \mathcal{L}(X)R, \tag{9}$$

$$\mathcal{S}(x) = \alpha(\|x\|_2)x, \qquad \mathcal{S}(xR) = \mathcal{S}(x)R, \tag{10}$$

$$\mathcal{N}(x) = x/(\|x\|_2 + \epsilon), \qquad \mathcal{N}(xR) = \mathcal{N}(x)R, \tag{11}$$

$$(q_iR)(k_jR)^\top = q_ik_j^\top. \tag{12}$$

Here A mixes slots but not encrypted coordinates, scalar gates depend only on norms, and attention weights depend on invariant inner products. Residual addition and composition preserve equivariance.

Proof sketch. The layer constraints above are sufficient for equivariance by induction over the computation graph. Slot mixing commutes with right multiplication by R, scalar gates and normalization depend only on $\|x\|_2$, and attention weights are functions of invariant dot products. If two blocks F and G satisfy $F(XR) = F(X)R$ and $G(XR) = G(X)R$, then

$$(F+G)(XR) = (F+G)(X)R, \qquad (G \circ F)(XR) = G(F(X)R) = G(F(X))R. \tag{13}$$

Thus CNN-style slot convolutions, Transformer-style attention blocks, residual paths, and norm-based gates remain equivariant as long as they avoid coordinate-specific nonlinearities on the encrypted vector axis.

Architectural consequence. The proof is useful because it is key-agnostic. Once a server block is built from the equivariant primitives above, the same weights work for every client key R; the server does not need to know R, retrain for a new key, or conjugate weights by R. A convolutional block satisfies the same condition when it treats each encrypted vector as an atomic feature:

$$\mathcal{C}(X)_{i:} = \sum_{\delta \in \mathcal{N}} B_\delta X_{i+\delta:} \quad \Longrightarrow \quad \mathcal{C}(XR) = \mathcal{C}(X)R, \tag{14}$$

where B_δ mixes slots or channels but not the encrypted coordinate axis. Likewise, if queries, keys, and values rotate together, attention scores $q_i k_j^\top$ are invariant and the weighted sum of values rotates by R:

$$a_{ij}(X) = \frac{\exp(q_i k_j^\top/\sqrt{d})}{\sum_m \exp(q_i k_m^\top/\sqrt{d})}, \quad o_i(XR) = \sum_j a_{ij}(X)(v_j R) = \left(\sum_j a_{ij}(X) v_j\right) R. \tag{15}$$

The same invariant dot products that make attention useful also expose Gram-statistic leakage, so coordinate-hiding does not imply privacy. Forbidden operations include coordinate-wise nonlinearities, coordinate-specific affine maps, and normalizations over the encrypted vector coordinate.

3.4 Attacker Objectives

3.4.1 Retrieval and Rotated-Token Task Prediction

Single-reference retrieval ranks auxiliary references by

$$\widehat{i}_{\text{single}} = \arg\min_{i \in \{1,\dots,M\}} d_{\text{single}}(y, a_i). \tag{16}$$

Pair and sequence attacks use invariant relational signatures:

$$\begin{aligned} s_{\text{pair}}(y_p, y_q) = y_p y_q^\top = x_p x_q^\top, \qquad G(Y) = YY^\top = XX^\top, \\ \widehat{k}_{\text{seq}} = \arg\min_k d_{\text{seq}}(G(Y), G(A^{(k)})). \end{aligned} \tag{17}$$

A direct no-key decoder instead predicts a task output from rotated features,

$$\widehat{o}_{\text{rot}} = g_\omega(Y), \tag{18}$$

without multiplying by $R^\top$. This covers unauthorized classification, token prediction, captioning, transcription, and probing.

3.4.2 Norm, Gram, and EDNN Attackers

The norm attacker uses the fingerprint $\|x_j R\|_2 = \|x_j\|_2$:

$$\hat{i} = \arg\min_i \left| \|y\|_2 - \|a_i\|_2 \right|. \tag{19}$$

Pre-normalization targets this channel by replacing x_ℓ with

$$\widetilde{x}_\ell = \frac{x_\ell}{\|x_\ell\|_2 + \epsilon}, \qquad \|\widetilde{x}_\ell R\|_2 = \|\widetilde{x}_\ell\|_2 \approx 1. \tag{20}$$

It does not remove angular or Gram structure:

$$G(Y) = YY^\top = XRR^\top X^\top = XX^\top, \qquad \hat{c} = h_\omega(G(Y)). \tag{21}$$

EDNN is included as an external embedding-inversion baseline. For $\text{shift}(v) = [v_2, \dots, v_d, v_1]$,

$$\Delta(v) = v - \text{shift}(v), \qquad D_i = \|\Delta(z) - \Delta(a_i)\|_2, \qquad \hat{i} = \arg\min_i D_i. \tag{22}$$

Lower attacker accuracy is better; none of these attacks or defenses establishes cryptographic security.

3.5 Salted, Input-Protected, and Noise Variants

The evaluated protection maps are

$$z_{\text{rot}} = xR_d, \tag{23}$$

$$z_{\text{noise}} = xR_d + \sigma\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_d), \tag{24}$$

$$\bar{x} = [x\ s] \in \mathbb{R}^{d+r}, \qquad s \sim \mathcal{N}(0, I_r), \qquad z_{\text{salt}} = \bar{x} R_{d+r}, \tag{25}$$

$$\hat{x} = \rho_d(z_{\text{salt}} R_{d+r}^\top). \tag{26}$$

Noise-only perturbation changes the observed value in the original dimension; salted/input-protected variants augment the representation before rotation and recover the signal coordinates through ρ_d. The Food-101 sweeps instantiate these mechanisms with 8D post-rotation noise and an 8D signal plus 4D salt before 12D rotation.

4 DATASETS

MNIST. Grayscale digit images with labels 0–9, used for dense EqCipher MLP, Eq-CNN, and norm-retrieval experiments (LeCun et al., 1998). **Fashion-MNIST.** Grayscale fashion images with 10 clothing labels, used for dense EqCipher and lifted/equivariant CNN experiments (Xiao et al., 2017). **CIFAR-10.** Color natural images with 10 object labels, used for dense EqCipher representation and norm-retrieval experiments (Krizhevsky, 2009).
WikiText-103. Natural-language text for next-token prediction, used for EqTextCipherLM utility, no-key decoding, perplexity, and invariant attacks (Merity et al., 2016). **COCO Karpathy.** Natural images with captions, using the Karpathy split, used for MobileCap/Nano-LLaMA and EqCipher captioning (Lin et al., 2014; Karpathy & Fei-Fei, 2015).
Food-101. Food images with 101 labels, used for MobileCLIP-S1 representation attacks and noise/salting sweeps (Bossard et al., 2014). **Stanford Cars.** Fine-grained car images (Krause et al., 2013). **Flowers-102.** Flower images with 102 labels (Nilsback & Zisserman, 2008). **CUB-200.** Fine-grained bird images (Wah et al., 2011). Frozen MobileCLIP-S1 (Vasu et al., 2023) maps these four datasets into 1024-dimensional visual representations attacked in Table 2. **GPT-2/RoBERTa-base.** These are text representation sources rather than datasets; their token embeddings are auxiliary reference representations, and Table 2 averages attack success across the two embedding spaces (Radford et al., 2019; Liu et al., 2019).

5 EXPERIMENTS AND RESULTS

The experiments have two layers. First, we test whether authorized EqCipher models preserve task utility while no-key decoders fail. Second, we test closed-reference leakage by instantiating $\mathcal{A}_{\text{aux}}$ as token embedding tables, image galleries, or labeled Gram features.

5.1 AUTHORIZED UTILITY, NO-KEY DECODING, AND ABLATIONS

Aim. This experiment separates two questions. The authorized-utility question is: after the server processes rotated latent tokens, can the client decrypt the server output and use it for the original task? The unauthorized-access question is: if a model or probe sees only rotated features and no key, does task decoding or invariant leakage remain possible? We do not claim that the server's rotated output is directly human-interpretable or directly usable by an ordinary task head without client decryption.

Setup.

- MNIST, Fashion-MNIST, and CIFAR-10 test dense and convolutional EqCipher utility together with norm-leakage ablations.
- WikiText-103 tests EqTextCipherLM perplexity after authorized decryption, no-key language-model decoding, and invariant retrieval attacks.
- COCO Karpathy tests captioning after authorized decryption and compares it with a no-key captioner applied without $R^\top$.
- MobileCLIP-S1 and Food-101 probes test coordinate leakage, Gram-derived class leakage, and noise/salting behavior.

Table 1 summarizes representative utility and leakage measurements across the evaluated modalities. Additional utility, probe, captioning, and sweep results are reported in Supplementary Appendix A.

Observation. The results show the intended authorized-utility behavior: when the server computation is equivariant, the client can decrypt the server output and use the normal task head with little loss in the reported settings. No-key task decoders often collapse, but that does not imply privacy because invariant leakage remains. Pre-normalization removes some norm attacks but not Gram attacks. In the controlled Food-101 comparison, salted/input-protected rotation preserves high utility while reducing Gram retrieval, whereas noise-only perturbation reduces leakage with severe utility loss. Server equivariance checks pass for EqText, MNIST Eq-CNN, Fashion-MNIST lifted Eq-CNN, and the COCO multi-head caption server; the largest caption-server check reports max error 8.28×10^{-7} at threshold 2×10^{-6}.

Table 1: Representative utility and leakage measurements. Task metrics use their native direction; unauthorized/leakage metrics are lower better.

Experiment	Authorized task metric	Unauthorized/leakage metric	Interpretation
COCO captioning	Plain baseline: 1.0908 CIDEr ↑; authorized decrypted EqCipher: 1.0770 CIDEr ↑	No-key captioner: 0.0003 CIDEr ↓	The client can recover a usable captioning representation after decryption; a no-key captioner cannot use the rotated representation.
WikiText-103	Authorized PPL ↓: 122.31 without pre-norm; 126.31 with pre-norm	Norm Top-1 ↓: $100.00 \rightarrow 0.00$; Gram Top-1 ↓: 100.00	Pre-normalization removes the norm attack but not the Gram attack.
Fashion-MNIST CNN	Authorized accuracy ↑: 92.20 without pre-norm; 91.37 with pre-norm	Norm Top-1 ↓: $100.00 \rightarrow 0.0014$	Norm retrieval can be suppressed with small utility loss.
Controlled Food-101, $\sigma = 10$	Accuracy ↑: 89.20 rotated; 88.99 pre-norm; 88.34 salted; 3.92 noise-only	Gram Top-10 ↓: 100.000 rotated/pre-norm; 0.230 salted; 0.040 noise-only	Salted/input-protected rotation preserves utility while sharply reducing Gram retrieval; noise-only reduces leakage by destroying utility.

5.2 Closed-Reference Text and Vision Retrieval

Aim. This experiment asks which information remains recoverable from protected representations when the attacker has an explicit auxiliary reference set but not the secret key.

Setup.

- Text-token retrieval hides a vocabulary entry t_j. GPT-2 and RoBERTa-base provide frozen embedding-table references $a_i = e_i \in \mathbb{R}^{768}$ (Radford et al., 2019; Liu et al., 2019), and attacks rank candidates by Eq. (16).
- Vision-instance retrieval hides a particular image in a closed gallery. CUB-200, Flowers-102, Food-101, and Stanford Cars are encoded with frozen MobileCLIP-S1, giving $a_i \in \mathbb{R}^{1024}$; attacks use NN, grouped norm, Gram/similarity-profile, or EDNN scores.
- Comparator rows are protected-representation transformations under the same retrieval protocol. Plain leaves x unchanged; Rotated uses xR_d; Pre-norm Rotated normalizes before rotation; Noise-only perturbs after rotation; Salted appends random coordinates before rotation.
- ConjFormer-style, Sentinel-style, and SnD-style rows are representation-level leakage comparators, not full end-to-end reproductions. SnD-style follows the Split-and-Denoise motivation of perturbing intermediate representations before server-side processing; η indexes the perturbation scale.

We attack the transformed representation directly and do not model a denoising stage or claim a formal local-DP guarantee.

Observation. Arbitrary rotation destroys coordinate-space NN and EDNN-style retrieval, but it does not hide invariant structure. Norm and Gram attacks remain strong under pure rotation; pre-normalization removes norm leakage but not vision Gram leakage. Perturbation, salting, and SnD-style perturbation can reduce attack success, but Table 2 is not a privacy–utility frontier: some transformations may reduce leakage by damaging task-relevant information.

Table 2: Average Top-10 attack success on vision and text representation benchmarks. Vision values average over CUB-200, Flowers-102, Food-101, and Stanford Cars using MobileCLIP-S1 representations. Text values average over GPT-2 and RoBERTa-base token embeddings. Lower is better for privacy. Gram is omitted for text because the token-level setup does not define a meaningful multi-token Gram statistic. The table reports leakage under the stated retrieval protocol and does not imply downstream task utility for every transformation.

Condition	Vision NN ↓	Vision Norm ↓	Vision Gram ↓	Vision EDNN ↓	Text NN ↓	Text Norm ↓	Text EDNN ↓
Plain	100.000	100.000	100.000	100.000	100.000	100.000	100.000
Rotated	**0.000**	100.000	100.000	**0.000**	0.002	100.000	0.002
Pre-norm Rotated	**0.000**	0.125	100.000	**0.000**	**0.000**	0.020	**0.000**
Noise-only ($\sigma = 10$)	62.783	0.141	0.130	36.089	0.072	0.024	0.068
Salted ($\sigma = 10$)	**0.000**	**0.119**	0.474	**0.000**	0.009	0.020	0.015
Sentinel Glide	99.125	30.697	0.910	100.000	83.072	0.026	100.000
Sentinel Shuffle	4.129	100.000	N/A	4.743	0.031	0.028	0.025
Sentinel Glide + Shuffle	1.961	5.830	N/A	4.743	0.018	0.019	0.025
ConjFormer-style	0.065	56.257	100.000	0.073	0.015	0.111	0.014
SnD-style ($\eta = 1.0$)	0.393	0.121	**0.129**	0.383	0.015	0.019	0.023
SnD-style ($\eta = 10.0$)	95.142	0.176	0.140	91.007	0.072	**0.018**	0.070

Table 3: Controlled Food-101 privacy–utility comparison using frozen MobileCLIP-S1 features. Utility is authorized Top-1 accuracy; attack columns are Top-10 retrieval success.

Condition	Utility ↑	NN ↓	Norm ↓	Gram ↓	EDNN ↓
Plain	89.41	100.000	100.000	100.000	100.000
Rotated	89.20	**0.000**	100.000	100.000	**0.000**
Pre-norm Rotated	88.99	**0.000**	**0.040**	100.000	**0.000**
Noise-only ($\sigma = 10$)	3.92	63.739	**0.040**	**0.040**	32.087
Salted ($\sigma = 10$)	88.34	**0.000**	0.044	0.230	**0.000**

5.3 Controlled Privacy–Utility Trade-off on Food-101

The existing staged Food-101 sweeps compare constructions that differ in representation dimension, noise placement, normalization, optimization schedule, and seed history. To isolate the protection mechanism itself, we run a controlled experiment in which all factors other than the protection transformation are held fixed.

Protocol. Food-101 images are encoded with frozen MobileCLIP-S1 using the same train/test split as the existing Food-101 experiments. All conditions use a fixed representation dimensionality, the same sampled orthogonal key R, optimizer and training setup, classifier/probe protocol, and random-seed schedule. The comparison includes Plain, Rotated, Pre-normalized + Rotated, Rotated + Gaussian noise at $\sigma = 10$, and Salted/Input-protected + Rotated at $\sigma = 10$.

Metrics. Each condition reports authorized Food-101 Top-1 accuracy ↑, Gram attack Top-10 retrieval success ↓, Norm attack Top-10 retrieval success ↓, NN Top-10 retrieval success ↓, and EDNN Top-10 retrieval success ↓. Table 3 gives a controlled privacy–utility comparison rather than combining utility and leakage values from incomparable staged setups.

Observation. Plain, Rotated, and Pre-norm Rotated retain high authorized accuracy, but Gram Top-10 leakage remains 100.000. Pre-normalization removes norm leakage without addressing Gram leakage. Noise-only perturbation suppresses Gram and norm retrieval but collapses utility to 3.92. Salted/input-protected rotation keeps utility at 88.34 while reducing Gram Top-10 retrieval to 0.230, giving the strongest controlled utility–leakage trade-off among the measured Food-101 conditions.

5.4 Vision-Class Leakage and Appendix Expansion

Aim. This experiment asks whether class information remains visible even when instance retrieval or direct coordinate decoding is suppressed.

Setup. The vision-class protocol measures label leakage rather than instance recovery. Labeled MobileCLIP-S1 features $x_i \in \mathbb{R}^{1024}$ are protected by $z_i = \Phi(x_i)$, and leakage is held-out probe

accuracy on protected coordinates or Gram-derived invariant features. Class probes, no-key captioning results, language-model probes, and Food-101 sweeps are reported in Supplementary Appendix A. Per-dataset and per-model retrieval results are reported in Supplementary Appendix B.

Observation. The direct-coordinate probes can be near random after rotation, but Gram-derived probes can remain highly informative. This is the same qualitative pattern as the retrieval benchmark: no-key coordinate failure is useful evidence, but it is not sufficient evidence of representation privacy under auxiliary invariant attacks.

6 Discussion

The consolidated results separate utility preservation from privacy. Orthogonal equivariance gives a clean authorized computation path: if the server respects the group action, decryption restores the feature orientation needed by the task head, as in Eq. (7) and Eq. (8). Once built from equivariant operations, the same server can process any client key R without key-specific retraining or weight conjugation.

The main lessons are:

- **Direct decoder failure is not privacy.** Orthogonal transformations preserve norms, pairwise similarities, and Gram matrices, which are sufficient for exact or near-exact recovery in several dense image and text settings.
- **Useful equivariance creates a leakage surface.** Attention and other equivariant blocks intentionally depend on invariant inner products, because those quantities let a server compute relationships without knowing R. The same quantities support norm, similarity-profile, and Gram attacks.
- **Pre-normalization is partial.** It suppresses norm fingerprints, but it does not remove angular or Gram structure. In vision probes, the direct attacker can be near random while a Gram classifier remains near the authorized classifier.
- **Salted/input-protected variants are promising but empirical.** Appending random coordinates before rotation changes the representation observed by the attacker and, in the staged Food-101 sweep, preserves high authorized utility under perturbation where the 8D noise-only construction loses utility. This is not a proof of privacy, and adaptive or repeated-query attacks remain open.

The controlled Food-101 study in Section 5.3 separates genuine privacy–utility effects from the confounded 8D/12D staged sweeps by holding dimensionality, optimization, data split, key, and seed protocol fixed.

The proof material also clarifies what should not be changed casually. Independent rotations per token would hide some cross-token geometry, but would also break the shared dot products needed by attention-style equivariant computation. Coordinate-wise nonlinearities or LayerNorm over the encrypted vector axis may improve ordinary neural modeling but generally violate arbitrary-R equivariance. Thus future mechanisms must trade off three quantities explicitly: authorized equivariant utility, invariant leakage, and compatibility with practical neural blocks.

The benchmark tables report leakage only, not complete privacy–utility frontiers. Noise-only or high-η privatization may suppress retrieval while degrading utility, and the ConjFormer-style, Sentinel-style, and SnD-style rows are representation-level attack-protocol comparators rather than full system reproductions. Orthogonal ciphering is therefore an equivariant interface, not a complete privacy layer; stronger constructions must hide Gram structure, reduce invariant sufficiency, or combine equivariance with noise, padding, aggregation, access control, or cryptographic protocols.

7 Limitations

The retrieval attacks are closed-reference attacks: text uses public token embedding tables, and vision instance retrieval assumes an auxiliary gallery of unencrypted reference features. They show

invariant leakage under auxiliary knowledge, not open-world reconstruction. The attack tables report leakage only; full privacy–utility comparison requires paired task-utility measurements for every perturbation condition. The staged Food-101 curves remain descriptive because they differ in dimension, noise placement, normalization, scheduler, and seed history; the controlled Food-101 comparison addresses this issue for the reported $\sigma = 10$ setting, while a dense perturbation sweep remains future work. Salted/input-protected variants reduce observed leakage in some protocols, but the experiments do not establish differential privacy, cryptographic indistinguishability, membership-inference resistance, adaptive reconstruction resistance, or side-channel resistance.

8 CONCLUSION

Orthogonal cipher token encryption enables key-agnostic equivariant computation: an equivariant server can process rotated latent representations without knowing the client key, and the client can recover task-usable features after decryption. However, orthogonal rotation alone does not provide representation privacy. Norms, pairwise similarities, and Gram structure are preserved and can support strong closed-reference leakage attacks even when direct no-key decoding fails.

Pre-normalization removes some norm-based leakage but does not remove Gram leakage. In the controlled Food-101 setting, salted/input-protected rotation substantially reduced Gram retrieval while retaining authorized utility, suggesting a promising empirical direction but not a privacy guarantee. The central conclusion is therefore that the same invariants that enable useful equivariant computation also create leakage channels; practical private inference requires additional mechanisms that explicitly weaken or hide those invariants.

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pp. 382–398, 2016.

Serge Assaad, Carlton Downey, Rami Al-Rfou, Nigamaa Nayakanti, and Ben Sapp. Vn-transformer: Rotation-equivariant attention for vector neurons, 2022. URL https://arxiv.org/abs/2206.04176.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461, 2014.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J. Guibas. Vector neurons: A general framework for so(3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12200–12209, 2021.

Ruijun Deng, Zhihui Lu, and Qiang Duan. Infodecom: Decomposing information for defending against privacy leakage in split inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(25):20737–20745, 2026. doi: 10.1609/aaai.v40i25.39212. URL https://ojs.aaai.org/index.php/AAAI/article/view/39212.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop on Text Summarization Branches Out*, pp. 74–81, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Yu Lin, Qizhi Zhang, Quanwei Cai, Jue Hong, Wu Ye, Huiqi Liu, and Bing Duan. An inversion attack against obfuscated embedding matrix in language model inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2100–2104, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.126. URL https://aclanthology.org/2024.emnlp-main.126/.

Yu Lin, Qizhi Zhang, Wenqiang Ruan, Daode Zhang, Jue Hong, Ye Wu, Hanning Xia, Yunlong Mao, and Sheng Zhong. Towards privacy-preserving llm inference via covariant obfuscation, 2026. URL https://arxiv.org/abs/2603.01499.

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

Peihua Mai, Ran Yan, Zhe Huang, Youjia Yang, and Yan Pang. Split-and-denoise: Protect large language model inference with local differential privacy. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34281–34302. PMLR, 2024. URL https://proceedings.mlr.press/v235/mai24a.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Abhijit Mishra, Mingda Li, and Soham Deo. Sentinellms: Encrypted input adaptation and fine-tuning of language models for private and secure inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21403–21411, 2024. doi: 10.1609/aaai.v38i19.30136. URL https://ojs.aaai.org/index.php/AAAI/article/view/30136.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://cdn.openai.com/better-language-models/language-models.pdf.

Abhishek Singh, Vivek Sharma, Rohan Sukumaran, John J. Mose, Jeffrey K. Chiu, Justin Yu, and Ramesh Raskar. Simba: Split inference–mechanisms, benchmarks and attacks. In *European Conference on Computer Vision*, pp. 214–232, 2024.

Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training, 2023. URL https://arxiv.org/abs/2311.17049.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Alexander Yukhimchuk, Andrey Shulga, Mladen Kolar, and Martin Takáč. Privacy from symmetry: Orthogonally equivariant transformers for llm inference, 2026. URL https://arxiv.org/abs/2606.16461.

Itamar Zimerman, Moran Baruch, Nir Drucker, Gilad Ezov, Omri Soceanu, and Lior Wolf. Converting transformers to polynomial form for secure inference over homomorphic encryption. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62803–62814. PMLR, 2024. URL https://proceedings.mlr.press/v235/zimerman24a.html.