

Few-shot Semantic Recovery Attacks on Image Embeddings

Yiyi Chen¹ Qiongkai Xu^{2*} Johannes Bjerva¹

¹Department of Computer Science, Aalborg University, Copenhagen, Denmark

²School of Computing, Macquarie University, Sydney, Australia

{yiyic, jbjerva}@cs.aau.dk qiongkai.xu@mq.edu.au

Abstract

Image embeddings are an essential commodity in modern AI: stored at scale in vector databases, transferred between edge devices and cloud APIs, and traded as paid services. Image embeddings have been implicitly treated as anonymous—an assumption that has not been rigorously tested even as embedding services scale to billions of stored vectors. *We show this assumption breaks.* We propose **Cross-modal Recovery Attack with Few-shot Transfer (CRAFT)**, a framework that recovers semantic meaning, including free-form captions and sensitive attributes. The attack focuses on the medical domain and facial features, targeting black-box image embeddings without reconstructing a single image in pixels and without any knowledge of the underlying model architecture, training data, or weights. **CRAFT** combines a public dual-encoder, a linear alignment learned from intercepted image-embedding pairs, and a fine-tuned embedding-to-text decoder; we further strengthen recovery with two augmentative techniques: cross-modal reranking against the attacker’s text encoder and chain-of-thought prompting on a diffusion-rendered low-fidelity reconstruction of the leaked embedding. Attacks succeed with as few as a single intercepted pair and transfer across five open-source and proprietary embedders. We disclose a privacy surface that current defenses cannot close. Every setting that preserves downstream utility also leaves the attack effective. Black-box image embeddings, as widely deployed in AI systems today, should not be treated as privacy-preserving artifacts.¹

1 Introduction

Embeddings are fundamental representations in contemporary AI. Image embeddings are optimized to preserve rich semantic and attributive infor-

mation from images (Chen et al., 2020), making them essential to multimodal systems and increasingly valuable as stored, transferred, and commercialized data assets (Radford et al., 2018; Cohere, 2025; Comanici et al., 2025). They allow large image archives to be searched, clustered, and compared efficiently, as illustrated by Google’s IDC image-search system for whole-slide pathology retrieval (Philbrick, 2026), and support high-performance medical image classification using lightweight classifiers, reducing the need for resource-intensive model training (Khoiwal and McMillan, 2024). However, the assumption underlying their deployment is the opposite, that the embedding is an opaque, privacy-preserving proxy, especially when the encoder weights are proprietary and only accessible through APIs.

This assumption has been broken for text embeddings; recent work has shown that an attacker can train an attack model that reverses the black-box embedding back to the original text in a few-shot setting, oftentimes verbatim (Chen et al., 2025b). Whether the same holds for image embeddings has not been extensively explored. First, a single image carries orders of magnitude more information than a sentence, so it is not obvious that a mean-pooled, fixed-dimensional vector retains enough of it to recover meaningful information, especially from unknown embedders. Second, the state-of-the-art image encoders that ship in production, such as COHERE (Cohere, 2025) and GEMINI (Lee et al., 2025), are opaque in training and weights, leaving no white-box surfaces open to attack (Bommasani et al., 2025). Beyond access, what it would mean to *reconstruct semantic meaning* from image embeddings has itself been under-defined and under-explored, except for recent work in white-box settings (Xiu and Zhang, 2025).

In this work, we demonstrate that image embeddings are, in fact, highly vulnerable to semantic recovery. Our framework, **Cross-modal Recovery**

*Correspondence author.

¹We open source our code for reproducibility and further research: <https://github.com/siebeniris/CRAFT>.

Attack with Few-shot Transfer (CRAFT), approaches the attack in two layers. The base attack fully leverages a pre-trained *dual encoder*, in which images and text are trained jointly, paired with a pre-trained language model and a closed-form linear alignment learned from a small set of intercepted image-embedding pairs. Two augmentative techniques further amplify the recovery: cross-modal reranking against the dual-encoder’s text encoder, and Chain-of-Thought Reasoning (CoT) over a low-fidelity reconstruction rendered by an off-the-shelf diffusion model. Without ever recovering the original pixels, an adversary can, therefore: (i) reconstruct the semantic meaning of the image as natural language; (ii) use a pre-trained diffusion model to render a low-fidelity intermediate reconstruction, not as a final output but as an additional visual signal that diversifies and grounds the recovered captions; and (iii) reveal sensitive attributes such as medical-examination metadata and identifiable facial features from the leaked vector. Using cross-model alignment, the same attack model transfers to any black-box embedder without target-embedder-specific retraining. Overall, our contributions are:

- We show that **image embeddings should be treated as privacy-sensitive as the source images and subject to the same data minimization and security practices**.
- We propose **CRAFT**, a novel attack framework that amortizes its training cost across all future victims via a closed-form per-victim alignment, exposing the universal vulnerability of image embeddings.
- We introduce two attack-time amplification techniques and extend the framework to direct attribute attacks on the victim embedding, demonstrating empirically that it matches the accuracy of a full-access attacker across medical and facial attributes.
- We test standard embedding-level defenses against **CRAFT** and find that none closes the leak without diminishing downstream utility, highlighting the severity of semantic recovery attacks on image embeddings.

2 Related Work

Embedding Inversion Attacks Ensuring secure embeddings is essential for contemporary AI systems, and embedding inversion attacks have gained

attention over the years, evolving from token-based reconstruction (Song and Raghunathan, 2020) to full-sentence generation (Li et al., 2023); from requiring a large volume of victim embeddings and their ground truth text (Morris et al., 2023) to few-shot (Chen et al., 2025b) and zero-shot settings (Zhang et al., 2025); extending to multilingual textual embedding inversion attacks (Chen et al., 2024, 2025a), and further focusing on cross-domain (Kim et al., 2026) and cross-lingual (Yu et al., 2025) optimization. In all of the above, embedding inversion attacks have exclusively focused on reconstructing text from textual embeddings.

Image Embedding Security While prior work has analyzed information leakage in encrypted images (Messadi et al., 2025), semantic leakage from image embeddings has received less attention. One line of work reconstructs faces from supposedly privacy-preserving face embeddings to attack facial recognition systems (Otroshi Shahreza et al., 2025, 2026; Wang et al., 2026). Chen et al. (2026b) similarly trains Stable Diffusion models to reconstruct semantically accurate images from CLIP embeddings. Closer to caption-level recovery, ZeroCap (Tewel et al., 2022) combines a frozen CLIP encoder with PPLM-style decoding on a frozen GPT-2 for zero-shot captioning, and CapRecover (Xiu and Zhang, 2025) recovers captions by training on white-box victim image embeddings. In contrast, **CRAFT** recovers semantic content without exact image reconstruction or training a massive decoder on victim embeddings.

3 Threat Model

We consider cross-modality inversion attacks in which an adversary aims to recover semantic information, such as image captions and labels, of a private image x from its embedding e_V in any victim model’s embedding space V . The embedding may have been intercepted from realistic attack surfaces: an embedding store accessed by an unauthorized party (Song and Raghunathan, 2020), an exfiltrated on-device cache (Sorensen and Khlaaf, 2024), a durable logged API response (Morris et al., 2023), or embeddings transmitted from an edge device to a cloud service that stores private content and serves downstream queries (Vepakomma et al., 2018).

Adversary’s Goal Let $\text{enc}_V : \mathcal{X} \rightarrow \mathbb{R}^d$ be a victim image encoder that maps images $x \in \mathcal{X}$ to

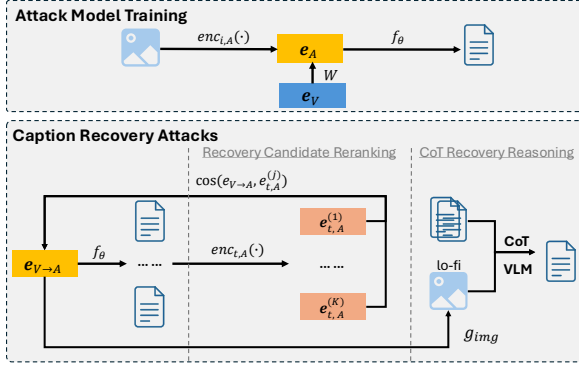


Figure 1: Few-shot Semantic Recovery Attack on Image Embeddings: 1) Attack Model Training; and 2) Caption recovery attacks, augmented by recovery candidate reranking and CoT recovery reasoning;

d -dimensional embeddings. Given $e_V = \text{enc}_V(x)$ for an unknown private image x , the adversary aims to i) reconstruct text $\hat{t}$ such that $\hat{t}$ preserves the semantic content of x , approximating the ground-truth captions t for x ; and ii) recover the label $\hat{y}$ for x , that matches the true label y .

Adversary’s Access We assume a realistic *black-box few-shot* scenario. The adversary has no access to weights, gradients, training data, or training procedures of $\text{enc}_V(\cdot)$, and *cannot query* at attack time. It holds only a small alignment set $D_{\text{pair}} = \{(x^b, e_V^b)\}_{b=1}^N$ of N paired tuples where both the image x_b and its victim embedding e_V^b are captured and N ranges from $\{1, 10, 10^2, 10^3, 10^4\}$, well within what a single compromised storage snapshot or a brief interception window would expose (The OWASP Foundation, 2025). The adversary has no captions or labels in the victim’s space, whereas it has unrestricted access to public image corpora with captions and concept and modality labels, $\mathcal{D}_{\text{pub}}$, drawn from a broad but likely mismatched distribution.

4 Cross-Modal Recovery Attack with Few-shot Transfer

We introduce a novel framework for a **cross-modal recovery attack with few-shot transfer on image embeddings**.

4.1 Attack Model Training on Public Data

Let $\text{enc}_A(\cdot)$ be the *accessible dual encoders*, a jointly trained image–text model with image encoder $\text{enc}_{i,A}(\cdot)$ and text encoder $\text{enc}_{t,A}(\cdot)$; both are frozen and the adversary uses only $\text{enc}_{i,A}$ to train the attack model. Separately, let $(\text{enc}_{\text{LM}}, \text{dec}_{\text{LM}})$

be an off-the-shelf, *text-only* pretrained encoder–decoder language model.

As shown in Fig. 1 (Top), the adversary trains a lightweight *auxiliary* image-embedding-to-text model $f_\theta = (\phi_\theta, \text{dec}_{\text{LM}})$ from scratch on a public image–caption corpus $\mathcal{D}_{\text{pub}}$, where $\phi_\theta : \mathbb{R}^{d_A} \rightarrow \mathbb{R}^{T \times h}$ is a lightweight multilayer perceptron (MLP), trained from scratch, that projects the L_2 -normalized image embedding $\text{enc}_{i,A}(x)$ into a sequence of T pseudo-tokens of hidden dimension h , and dec_{LM} is the decoder of the pretrained LM, initialized from its pretrained weights and fine-tuned. The pseudo-tokens $\mathbf{c}_x$ replace the LM encoder’s output in the decoder’s cross-attention, so enc_{LM} is discarded. Both trainable components - ϕ_θ and dec_{LM} , with parameters collected in θ - are optimized end-to-end on $\mathcal{D}_{\text{pub}}$, with the other parameters frozen, by maximum likelihood with teacher forcing:

$$\mathcal{L}_{\text{dec}}(\theta) = - \sum_{(x,t) \in \mathcal{D}_{\text{pub}}} \sum_{k=1}^{|t|} \log p_{\text{dec}_{\text{LM}}}(t_k \mid t_{<k}, \mathbf{c}_x) \quad (1)$$

where $\mathbf{c}_x := \phi_\theta(\text{enc}_{i,A}(x))$, optimized with AdamW.

4.2 Embedding Alignment

We adopt the linear embedding alignment method from Chen et al. (2025b) to map victim embedding e_V to attack space A with a learned linear mapping matrix $\mathbf{W}$, i.e., $e_{V \rightarrow A} = e_V \mathbf{W}$. The optimal alignment matrix $\mathbf{W}$ is obtained by solving the least-squares minimization $\min_{\mathbf{W}} \|\mathbf{E}_A - \mathbf{E}_V \cdot \mathbf{W}\|^2$, where $\mathbf{E}_V = [e_V^{1\top}, \dots, e_V^{N\top}]^\top \in \mathbb{R}^{N \times d_V}$ and $\mathbf{E}_A = [e_A^{1\top}, \dots, e_A^{N\top}]^\top \in \mathbb{R}^{N \times d_A}$, and N is the number of alignment samples, whereas d_V and d_A are the embedding dimensions in victim and attack spaces respectively. In result, the solution to the least squares loss is:

$$\mathbf{W} = (\mathbf{E}_V^\top \mathbf{E}_V)^{-1} \mathbf{E}_V^\top \mathbf{E}_A. \quad (2)$$

The aligned embedding matrix from V to A is thus:

$$\mathbf{E}_{V \rightarrow A} = \mathbf{E}_V \mathbf{W}, \quad (3)$$

where $\mathbf{E}_{V \rightarrow A} \in \mathbb{R}^{N \times d_A}$. We use D_{pair} to align victim vector e_V to the attack space A , i.e., $e_{V \rightarrow A}$.

4.3 Caption Recovery Attacks

As shown in Fig. 1 (Bottom), we use the trained decoder f_θ to reconstruct the semantic meaning

of the aligned embedding $e_{V \rightarrow A}$ via beam search, producing a candidate pool of K captions:

$$\{\hat{t}^{(1)}, \dots, \hat{t}^{(K)}\} = \text{BeamSearch}_K(e_{V \rightarrow A}; f_\theta). \quad (4)$$

Recovery Candidate Reranking. To further improve attack performance, the adversary leverages the attacker text encoder $\text{enc}_{t,A}(\cdot)$ to embed each candidate,

$$e_{t,A}^{(j)} = \text{enc}_{t,A}(\hat{t}^{(j)}), \quad j = 1, \dots, K, \quad (5)$$

and ranks the candidates by cosine similarity to the aligned image embedding in the shared attacker-side joint space:

$$\hat{t} = \arg \max_{j \in \{1, \dots, K\}} \langle e_{t,A}^{(j)}, e_{V \rightarrow A} \rangle. \quad (6)$$

This grounds the decoded text against the attack joint embedding space without extra training cost.

CoT Recovery Reasoning. The candidate pool $\{\hat{t}^{(j)}\}_{j=1}^K$ and the aligned embedding $e_{V \rightarrow A}$ encode complementary information about the source image x : the captions identify entities, counts, and actions reliably but are linguistically sparse, while the embedding additionally retains coarse visual structure (colors, spatial layout, scene type) that the decoder f_θ tends to neglect based on the trained data. To exploit both, the adversary first renders a low-fidelity reconstruction

$$\hat{x}^{\text{lo-fi}} = g_{\text{img}}(e_{V \rightarrow A}), \quad (7)$$

where g_{img} is an off-the-shelf diffusion model, and we use the same alignment method to align the victim embedding into the embedding space of g_{img} .

The pair $(\hat{x}^{\text{lo-fi}}, \{\hat{t}^{(j)}\}_{j=1}^K)$ is then handed to an off-the-shelf vision-language model $\text{VLM}(\cdot)$ with a chain-of-thought prompt (Wei et al., 2022) that explicitly instructs the model to (i) identify the consensus content across the K captions, (ii) extract visual cues from $\hat{x}^{\text{lo-fi}}$, (iii) reconcile the two sources, preferring captions for entities and counts, and the image for color, layout, and atmosphere, and (iv) generate a final caption:

$$\hat{t} = \text{VLM}(\hat{x}^{\text{lo-fi}}, \{\hat{t}^{(j)}\}_{j=1}^K; \pi_{\text{CoT}}), \quad (8)$$

with π_{CoT} the structured prompt template. CoT reasoning enables VLM to act as a learned prior over plausible scenes, modulating decoder hallucinations that contradict the visual evidence $\hat{x}^{\text{lo-fi}}$ and adding visually salient details that the candidate K captions omit.

4.4 Attribute Attacks

Beyond free-form caption recovery, the adversary can directly query $e_{V \rightarrow A}$ for sensitive attributes (e.g., facial features, clinical findings), which does not require training an attack model. Let $\mathcal{A} = \{a_1, \dots, a_C\}$ a fixed attribute vocabulary with C targets, and let $\mathbf{y} \in \{0, 1\}^C$ be the per-attribute ground-truth indicator vector for the source image x . The adversary recovers $\hat{\mathbf{y}}$ using three attribute-attack approaches that leverage the attack dual encoders $\text{enc}_A(\cdot)$ and the learned $\mathbf{W}$.

Zero-Shot Probing For each attribute a_c , the adversary constructs a positive and a negative prompt and embeds both with the text encoder $\text{enc}_{t,A}$ (Radford et al., 2021):

$$\begin{aligned} e_c^+ &= \text{enc}_{t,A}(\text{"a photo of a person with } a_c \text{"}), \\ e_c^- &= \text{enc}_{t,A}(\text{"a photo of a person without } a_c \text{"}). \end{aligned} \quad (9)$$

We assign a_c to the closer prompt under cosine similarity:

$$\hat{y}_c^{\text{zs}} = \mathbb{1}[\langle e_c^+ - e_c^-, e_{V \rightarrow A} \rangle > 0]. \quad (10)$$

Linear Probe The adversary also trains a linear classifier f_ψ on top of the attack-model image embeddings. Given a labeled public corpus $\mathcal{D}_{\text{pub}} = \{(x_j, y_j)\}$ with attribute labels $y_j \in \{0, 1\}^C$, f_ψ is fit on the attack-encoded inputs $\text{enc}_{i,A}(x)$ by masked binary cross-entropy²:

$$\mathcal{L}_{\text{cls}}(\psi) = \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{pub}}} \left[\sum_{c=1}^C m_c \text{BCE}(y_c, \sigma(z_c)) \right] \quad (11)$$

where z_c is the c -th logit of $f_\psi(\text{enc}_{i,A}(x))$ and $m_c = \mathbb{1}[y_c \neq -1]$ masks unlabeled attributes. During attack, the trained classifier is applied to the aligned victim embedding $e_{V \rightarrow A}$,

$$\hat{y}_c^{\text{lc}} = \mathbb{1}[(f_\psi(e_{V \rightarrow A}))_c > 0], \quad (12)$$

with the raw logits acting as continuous scores for ranking metrics. We use a linear classifier following Alain and Bengio (2017).

MLP Probe The adversary additionally trains a non-linear classifier, i.e., a MLP $f_\eta(\mathbf{x}) = \mathbf{W}_2 \cdot \sigma_{\text{ReLU}}(\mathbf{W}_1 \mathbf{x} + b_1) + b_2$ with hidden dim 512 and dropout rate of 0.3, on the attack space embeddings with softmax cross-entropy for the single-label task (ROCOv2 modality), and masked binary

²<https://docs.pytorch.org/docs/2.12/generated/torch.nn.BCELoss.html>

cross-entropy over all labels jointly for the multi-label tasks (**ROCOv2** CUI concepts and **CelebA** attributes).

5 Experimental Setup

We use the FLANT5 model as a backbone for training our attack generator, and CLIP-VIT-BASE-PATCH32 as our attack dual encoder. We attack embeddings from victim encoders CLIP (LARGE-PATCH14), NOMIC, GEMINI2, COHERE and NVIDIA. For training, we use batch size 256, learning rate $1e-3$, and up to 50 epochs with early stopping on validation loss. The full list of models and encoders is in Table 6, as are datasets and tasks in Table 7. For caption recovery attacks, we use COCO (Lin et al., 2014), Flickr30k (Young et al., 2014) for training/evaluating in-domain caption recovery attacks; and NoCaps (Agrawal et al., 2019) for cross-domain evaluation with COCO-trained f_θ .

Caption Recovery Baselines As our attack scenario has not been widely examined, we have one directly related baseline, **SLImE** (Chen et al., 2026a), which contrastively aligns image embeddings with text-tag embeddings and refines the output with a multimodal LM. We further compare with **ZeroCap**, a zero-shot CLIP-guided caption generator; we adapt it by feeding $e_{V \rightarrow A}$ in place of its CLIP image embedding. **Vec2Text** assumes access to paired victim embeddings and captions, which we provide directly as (e_V, t) pairs for training. **ALGEN** is a cross-encoder *text*-embedding inverter, which we adapt to images by aligning e_V to $e_{t,A}$ first and then running its text-trained generator on $e_{V \rightarrow A}$.

CLIP Rerank f_θ generates $K = 5$ candidates per image via beam search, we choose $K = 5$ to strike a balance of compute efficiency and performance (see the ablation study K in Table 8); the attacker reranks them by cosine similarity to $e_{V \rightarrow A}$ using the attacker CLIP text encoder $enc_{t,A}$. The argmax is reported as $\hat{t}$; the top- K pool is fed to CoT.

CoT Recovery g_{img} is the frozen KANDINSKY 2.2 (Razzhigaev et al., 2023) conditioned on a separately aligned CLIP-BIGG embedding with $\mathcal{D}_{\text{pair}}$ (cf. Table 6). The lo-fi output and the rerank pool are passed to GPT-4O (OpenAI et al., 2024) CoT prompt; the top-1 output is the refined $\hat{t}$.

Attribute Attacks We evaluate the attribute attack on **ROCOv2** (Rückert et al., 2024) (UMLS concept detection on radiology images) with linear probe and **CelebA** (Liu et al., 2015) (40 binary facial attributes) through both attacks of §4.4.; see Table 7.

Defense Setup We evaluate embedding-level defense approaches applied to e_V before use: Gaussian noise insertion and local Differential Privacy. Utility tasks are evaluated in the victim’s joint space: image-embedding-to-text retrieval $R@ \{1, 5, 10\}$ on the caption datasets and zero-shot classification accuracy on the attribute datasets.

Evaluation Metrics **Caption recovery:** multi-reference BLEU and Rouge- $\{1,2,L\}$ between $\hat{t}$ and t , and alignment $\cos(e_{V \rightarrow A}, e_{i,A})$. **Attribute attacks:** top-1 accuracy (modality), balanced accuracy, mean ROC-AUC (CelebA), and F1 with mean average precision (mAP) (multi-label UMLS concepts). **Defense and Utility:** cosine similarity between original and protected e_V ; image-embedding-to-text retrieval $R@ \{1,5,10\}$ on caption datasets and zero-shot classification accuracy on attribute datasets.

6 Attack Image Embeddings

6.1 Caption Recovery from Image Embeddings

Our attack directly recovers text from the aligned victim embedding $e_{V \rightarrow A}$ using the trained f_θ . Figure 2 reports the cosine similarity $\cos(e_{V \rightarrow A}, e_{i,A})$ on the alignment set (dotted) and test set (solid), and (bottom) the Rouge-L between the recovered text and the ground-truth captions, across alignment sizes N . With more alignment pairs, $e_{V \rightarrow A}$ converges to the attacker-space image embedding and the recovered text correspondingly improves. Even at $N=1$, reconstruction already reaches substantial ≈ 30 Rouge-L across all victim encoders, with COHERE particularly vulnerable in the extreme few-shot setting. Our threat model centers on the few-shot setting ($N \leq 10^3$), we compare against baselines at $N=10^3$, simulating the upper end of what a single intercepted batch realistically exposes (Table 1), where **CRAFT** achieves ≥ 50.99 in Rouge-L, indicating that most of the caption’s content words and their order have been reconstructed. $N=10^4$ is reported, showing the attack continues to scale outside the few-shot settings (Table 11). To test how sensitive the alignment is

Table 1: Caption Recovery from Black-Box COHERE Image Embeddings ($N = 1000$). The best results across metrics are **bolded**, and the ceiling results are in grey.

Methodology	Strategy	BLEU	Rouge-L	Rouge-1	Rouge-2
Vec2Text	Base	0.28	8.12	8.36	0.55
ALGEN	Base	4.33	28.05	33.05	11.09
	+ Rerank [T5]	4.38	28.04	33.11	11.00
ZeroCap	+ Align	1.37	19.03	20.93	3.39
SLImE	Deepseek v3.2	9.61	26.97	31.13	6.69
Ours	$e_{i,A}$	28.27	53.46	58.02	32.59
	Base	25.62	50.99	55.67	29.68
	+Rerank [CLIP]	26.32	51.92	56.29	30.69
	+CoT	28.40	53.01	57.37	31.88

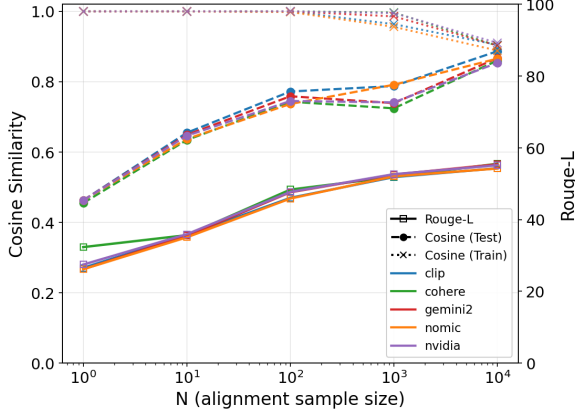


Figure 2: Caption Recovery (solid lines) and Cosine Similarity (dotted lines) across Embedding Models by Alignment Sample sizes, augmented with CLIP-Rerank.

to *which* alignment samples an attacker happens to obtain, we refit $\mathcal{W}$ on 10 independent random subsets for each victim embedder and each shot size N , and report alignment cosine and Rouge-L as mean $\pm$ 95% CI in Table 9. At each N , the results are stable across refits and statistically indistinguishable across embedders; the only axis along which they move is the attacker’s budget N itself.

Table 1 reports caption recovery from COHERE image embeddings at $N = 1,000$ alignment pairs. Vec2Text and ALGEN both require paired image-text data beyond the intercepted (image, embedding) pairs and are not directly applicable to our setting; we adapt each as faithfully as possible. Vec2Text fails to converge in the few-shot setting and remains subpar even with substantially larger victim data (Table 10, 11). ZeroCap, coupled with ALGEN alignment and QWEN as decoder, reaches 19.03 Rouge-L; SLImE with DeepSeek v3.2 reaches 26.97.

Our base attack already more than doubles BLEU over the strongest baseline (25.62 vs. 9.61) and improves Rouge-L by 19.9 absolute points (50.99 vs. 31.13), without any inference-time re-

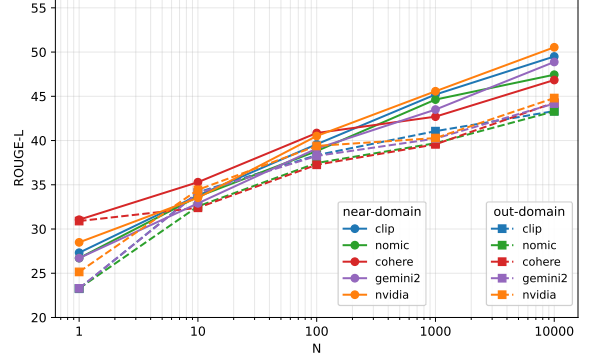


Figure 3: Near-Domain and Out-Domain Caption Recovery on NoCaps dataset with CLIP-Rerank.

finement. CLIP-Rerank consistently improves all four metrics (+0.70 BLEU, +0.93 Rouge-L); applying a T5 reranker to ALGEN, in contrast, yields essentially no gain (+0.05 BLEU), suggesting the benefit derives specifically from attack CLIP’s joint image-text space. CoT refinement adds a further +2.08 BLEU, bringing our best configuration to 28.40 BLEU / 53.01 Rouge-L, close to the ceiling obtained by decoding directly on the attack embedding $e_{i,A}$. The same pattern holds across all victim embedders (Table 11). See Table 18 for qualitative analysis of leveraging CLIP-Rerank and CoT reasoning for caption recovery. The small margin reflects the strength of the core attack rather than a weakness of CoT: alignment and reranking alone already approach the white-box reference, leaving little headroom for any refinement stage. The core attack uses no VLM or diffusion model and establishes our main results; we include CoT as an optional stage to show that an attacker willing to pay can push recovery further, which matters in a privacy setting where any additional recovered detail is a real gain.

We adopt CLIP-Rerank as the default decoding strategy for the rest of the paper, unless noted otherwise, to balance cost and performance.

6.2 Cross-Domain Caption Recovery

We test robustness to image distribution shift by applying the COCO-trained f_θ and COCO-fit alignments W unchanged to NoCaps, whose images are drawn from Open Images and split by object-vocabulary overlap with COCO into *near-domain* and *out-domain*. Recovery improves monotonically with N on both, as in the in-domain case (Fig. 3). At $N=10^3$, the near-domain split reaches mean BLEU 18.8 and Rouge-L 44 across the five

victims, within a few points of in-domain data, while the out-domain split reaches BLEU 13.1 and Rouge-L 40 (Table 13). The BLEU gap between the splits is 5-7 across N ; the Rouge-L gap is smaller (~ 4). Distribution shift therefore makes the attack measurably harder, but does not prevent it. See full results in Table 12, and qualitative analysis across victim encoders in Table 19.

6.3 Attribute Attacks on Image Embeddings with Sensitive Information

Image embeddings also reveal sensitive attributes of the original images; the privacy implications are stark. We attack the 40 binary facial attributes of CelebA via both zero-shot and linear probe (Table 2). Even without labeled training data, the zero-shot attack reaches ROC-AUC ~ 62 at $N=10^3$, within 4-5 points of the white-box reference 66.81. An adversary with a small labeled face corpus on the attack side reaches ROC-AUC ~ 80 via a linear probe at only $N=10^2$, near the the white-box reference ceiling of 82.71; the probe at $N=10$ already exceeds the zero-shot attack at $N=10^3$ (ROC-AUC > 70), recovering gender, expressions, facial characteristics, accessories with no query access to the victim encoder. Fig. 6 shows the result per attribute: *Male* is recovered near-perfectly, and *Wearing Lipstick*, *No Beard*, and *Heavy Makeup* all exceed ROC-AUC 90 on both victim embeddings. Results are consistent across the two victim encoders (CLIP and Nomic within ~ 2 AUC at every N).³ With MLP probe, for attribute attacks on CelebA across N , the performance is better or comparable across metrics compared to zero-shot and the linear probe (cf. Table 2); for attribute attacks on ROCov2 CUI, the performance is particularly strong in F1 score (≥ 60 at $N = 10^3$), while comparable in mAP in comparison to the linear probe (cf. Table 14).

Beyond face attributes, image embeddings expose the anatomical and modality structure of medical images. We attack UMLS concept detection on ROCov2 over the 19 most frequent concepts present in the eval data, using the linear probe (see details per-class in Fig. 5). The recovered concepts describe *how* an image is acquired and *what* is imaged, such as imaging modality (CT, MRI, plain X-ray, ultrasound, angiogram), anatomical region (skull, spine, pelvis, breast, lower extremity), and

³CelebA’s evaluation split is too large to extract through paid embedding APIs; we restrict this dataset to open-source victims.

Table 2: Attribute Attacks on CelebA across N .

	Embedder	N	ROC-AUC	ACC	Bal. ACC
Zero-shot	$e_{i,A}$	–	66.81	68.03	61.29
	CLIP	1	49.09	64.66	50.00
		10	50.90	67.49	51.15
		100	59.73	67.04	55.79
		1000	62.96	65.87	58.39
	Nomic	1	50.07	64.66	50.00
		10	52.38	67.95	52.01
		100	58.53	66.29	55.55
		1000	61.83	66.11	57.60
Linear Probe	$e_{i,A}$	–	82.71	83.99	54.28
	CLIP	1	50.34	79.91	50.00
		10	70.86	81.69	51.84
		100	79.82	83.64	53.94
		1000	79.70	83.71	53.97
	Nomic	1	50.19	79.91	50.00
		10	75.76	83.03	53.31
		100	79.52	83.78	54.09
		1000	77.90	83.72	54.00
MLP Probe	$e_{i,A}$	–	79.32	85.41	56.37
	CLIP	1	54.59	78.35	50.00
		10	71.44	82.66	53.16
		100	77.37	84.96	55.93
		1000	78.35	85.22	56.19
	Nomic	1	49.23	78.35	50.00
		10	74.98	83.99	54.57
		100	77.61	84.85	55.78
		1000	78.41	85.14	56.15

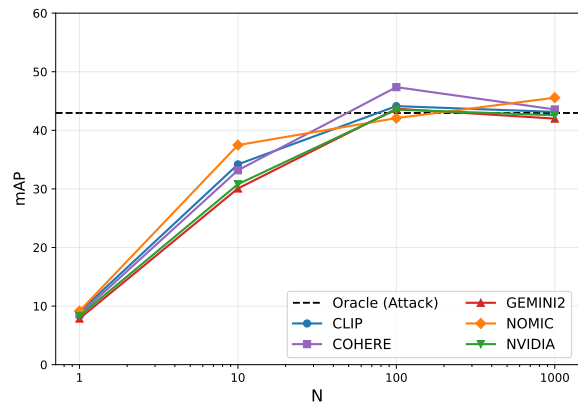


Figure 4: Attribute Attacks on ROCov2 CUI across N and Victim Encoders in mAP.

view orientation (anterior-posterior, sagittal). The attack matches or exceeds the the white-box reference ceiling (mAP 42.96) at $N=10^2$ across all five victim encoders, including the proprietary ones (COHERE 47.37, GEMINI2 43.68, NVIDIA 43.53) and the open-source CLIP 44.13 and NOMIC 42.08 (Table 14). Even at $N=10$, mAP reaches 30–37 across encoders. Precision remains at 100% across all $N \geq 10^2$: the probe predicts few positives per image (low recall), but every recovered concept discloses, with no error, what kind of medical exami-

Table 3: COHERE Embeddings, Utility and Defense, from the perspective of consumers and service providers, the retrieval performance the higher the better $\uparrow$, and the inversion performance the lower the better $\downarrow$. (N=1000)

Dataset	Defense	Parameter	R@1 $\uparrow$	R@5 $\uparrow$	R@10 $\uparrow$	COS $\downarrow$	BLEU $\downarrow$	Rouge-L $\downarrow$	Rouge-1 $\downarrow$	Rouge-2 $\downarrow$
Flickr30k	No Defense	-	95.8	99.8	100.0	0.6519	19.63	44.45	48.06	24.06
	Gaussian Noise	0.001	95.60	99.80	100.00	0.6515	20.21	45.06	48.74	24.33
		0.01	94.80	99.80	100.00	0.6563	19.52	45.47	48.84	24.87
		0.1	76.20	93.60	96.30	0.4358	8.47	32.99	36.20	12.93
		1	0.5	1.30	2.60	0.0418	5.90	28.87	31.85	9.66
	LapMech	0.1	10.90	26.50	36.50	0.2114	6.86	30.40	33.50	10.85
		0.25	77.00	93.70	96.10	0.4332	8.86	33.04	36.62	13.53
		0.5	92.40	99.60	99.90	0.5500	13.15	38.45	42.02	18.52
		1	94.30	99.60	99.90	0.6186	17.59	42.62	46.01	22.17
		2	94.70	99.60	100.00	0.6510	18.84	44.40	48.19	23.63
	PurMech	0.1	8.70	24.10	33.4	0.0174	6.22	29.30	32.03	9.79
		0.25	55.60	68.60	70.30	0.4310	7.92	32.69	36.16	12.54
		0.5	91.10	99.20	99.80	0.5473	13.08	37.86	41.62	18.15
		1	93.30	99.90	99.90	0.6167	17.03	41.75	45.59	21.65
		2	94.90	99.70	100.00	0.6510	19.42	44.91	48.58	24.67
NoCaps	No Defense	-	97.40	100.00	100.00	-	15.32	41.14	45.38	19.57
	Gaussian Noise	0.001	97.60	100.00	100.00	-	14.79	40.98	45.22	19.29
		0.01	97.40	100.00	100.00	-	13.99	40.20	44.29	19.05
		0.1	85.40	97.00	99.00	-	6.43	31.99	34.90	10.03
		1	1.00	3.00	6.40	-	4.79	30.77	33.71	8.83
	LapMech	0.1	13.60	35.80	45.60	-	5.55	31.23	33.89	9.43
		0.25	84.60	97.40	99.00	-	6.43	32.28	35.22	10.54
		0.5	94.80	99.80	100.00	-	8.01	33.72	36.62	12.25
		1	96.40	99.80	100.00	-	9.69	36.22	39.74	14.00
		2	97.00	100.00	100.00	-	13.04	39.75	43.28	17.79
	PurMech	0.1	12.80	29.20	41.80	-	5.51	31.29	34.13	9.43
		0.25	61.60	70.00	70.80	-	6.24	32.05	35.09	10.11
		0.5	94.20	99.80	100.00	-	7.69	33.30	36.47	11.64
		1	98.00	100.00	100.00	-	11.29	36.78	40.45	15.47
		2	97.60	100.00	100.00	-	13.40	39.74	43.66	18.21

nation the patient underwent, a class of metadata that conventional embedding-storage protocols assume to be hidden.

7 Limits of Perturbation-Based Embedding Sanitization

To explore mitigations against our methods, we evaluate perturbation-based mechanisms designed to protect systems from embedding-based attacks.

Gaussian Noise Insertion We add Gaussian noises to e_V with $(e_V + \lambda \cdot \epsilon) / \|e_V + \lambda \cdot \epsilon\|$, $\epsilon \sim \mathcal{N}(0, 1)$, with $\lambda \in [0.001, 0.01, 0.1, 1]$ (Morris et al., 2023; Chen et al., 2024, 2025b).

Local Differential Privacy (LDP) Purkayastha Mechanisms (PurMech) and Normalized Planner Laplace (LapMech) are leveraged to ensure metric LDP on textual embeddings (Du et al., 2023; Chen et al., 2025b). To compare the same operating point across our victims of dimension $d \in \{768, 1536, 2048, 3072\}$ (cf. Table 6), we apply a dimension-normalized budget $c := \epsilon/d$. Both

mechanisms then yield the same retained cosine similarity between unprotected and protected victim embeddings:

$$\rho(c) = \frac{c}{\sqrt{1+c^2}}, \quad (13)$$

independent of d . We evaluate $c \in \{0.1, 0.25, 0.5, 1, 2, 4, 8\}$. Equivalently, $\rho \in \{0.10, 0.24, 0.45, 0.71, 0.89, 0.97, 0.99\}$. The formal LDP guarantee at each c is $\epsilon = c \cdot d$.

7.1 Results

We show comprehensive utility and mitigation results for COCO (Table 15), Flickr30 (Table 16) and with cross-domain results on NoCaps (Table 17), across victim embedders. We present results for COHERE embeddings on Flickr30k and NoCaps in Table 3.

At the most aggressive LDP setting, $c=0.1$ for both LapMech and PurMech, retrieval R@1 falls to 10-14 across the three datasets while recovered captions still reach Rouge-L between 29.30 and 31.29 (COHERE: 30.40 on Flickr30K, 31.23 on NoCaps,

Table 4: Cost of each stage of the attack.

Stage	When paid	Cost	(V)LLM pass?
Decoder training	once, offline (shared by all victims)	3 h training	no
Alignment W	once per victim	2.6 ms ($N=1$) $\rightarrow$ 11.2 ms ($N=10^4$)	no
Recovery (no CoT)	per image embedding	≈ 10 ms (GPU) / ≈ 160 ms (CPU)	no
+ <i>CoT refinement (optional)</i>	per image embedding	≈ 10 –30 s (GPU)	yes

Table 5: Alignment fit time as a function of the number of paired samples N ($d = 768$).

N	1	10	10^2	10^3	10^4
Fit time	2.6 ms	2.6 ms	2.7 ms	3.4 ms	11.2 ms

30.37 on COCO). Gaussian noise at $\lambda=1$ produces a comparable result: R@1 falls to ≤ 2 while Rouge-L stays in the 28.87–30.77 range. Both mechanisms thus drive the attack to a Rouge-L of 30.

The LDP setting with $c=0.5$ retains usable retrieval (COHERE/Flickr30K R@1 92.4) at the cost of leaving the attack at Rouge-L 38.45 and BLEU 13.15, reducing inversion quality by approximately 6 Rouge-L points from the undefended baseline (44.45) but not eliminating it. The cross-domain NoCaps results are particularly alarming: even when the attacker holds no in-domain alignment and no defended alignment pairs, a weaker threat model than the in-domain setting, the attack still recovers Rouge-L in the 30–33 range under every defense in our setting. Lower c and higher λ values reduce both utility and attack quality simultaneously but never drive Rouge-L below 30.

8 Discussion and Conclusion

Image embeddings are not anonymous representations of the underlying images. Using only a small set of intercepted image-embedding pairs, as few as a single pair, and a publicly available dual-encoder, an adversary recovers natural-language captions and sensitive attributes from leaked embeddings across contemporary image encoders, including proprietary ones. The attacker trains an attack model on its own encoders once and amortizes the cost across all future victims via a linear alignment. None of the standard embedding-based defenses we evaluate eliminates the leakage. The assumption that storing embeddings rather than pixels preserves the privacy of the underlying images is called into question. We urge the design and evaluation of defenses that account for cross-encoder transferability and adaptive adversaries, and treat image embeddings in production systems as the sensitive, high-risk artifacts they are.

Computing Resources and Use of AI

We specify the cost of the attack per stage in Table 4. All training and evaluation were run on a single AMD Instinct MI250X GPU (ROCm 6.2.4 + PyTorch 2.7); each attack model takes approximately 3 hours to train. The cost of alignment depends on the hidden dimension size d of the victim embeddings; as d grows (we evaluated embeddings with $d \in [768, 1536, 2048, 3072]$), the cost is negligible. We show an example of the cost with $d = 758$ in Table 5. The core attack is cost-efficient: after a one-time decoder training (reused across all victims), calibrating to a new victim is a closed-form alignment fit costing about 10 ms and a handful of pairs, and recovering a caption costs about 10ms per image on GPU (beam search + CLIP rerank; no VLM). The CoT refinement is an optional stage that trades cost for performance; it adds a Kandinsky (open-source model) reconstruction plus an VLM reasoning pass per refinement iteration, about 10–30s per image. We report it separately and present the low-resource attack without it; CoT is only for an attacker willing to pay, where even marginal privacy gains are worthwhile.

We used a large language model assistant to provide programming suggestions and to check the draft’s grammar and fluency. We did not use AI to generate ideas, create images, or produce other substantive content for this work. The authors have reviewed the draft and take full responsibility for the manuscript.

Limitations

Our work does not propose a sufficient defense mechanism for **CRAFT**. Although we evaluated two of the most efficient existing defense mechanisms for textual embeddings, we found them ineffective against cross-modal semantic recovery attacks on image embeddings. The primary focus of this work is to expose the security vulnerabilities in embedding services providing image embeddings as commodities, and to inspire the development of future defense paradigms.

Ethical Considerations

CRAFT relies only on publicly available components and a small set of intercepted image-embedding pairs that any API user can construct independently; the vulnerability we expose pre-exists our publication, and we disclose it to motivate mitigation by the community. All datasets used are publicly available (MS-COCO, Flickr30K, NoCaps, CelebA, ROCOv2); CelebA contains identifiable faces, and we report only aggregate metrics. Commercial APIs were accessed under standard developer terms. This work is at risk of dual-use, as the methods that we develop can be used to infer sensitive information from exposed embeddings. We therefore open-source our code to enable reproducibility and defense development, but withhold the pre-fit alignments for the commercial victims and the intercepted alignment pair caches to limit direct misuse. We make no claim that defenses cannot exist; cryptographic protection at storage and per-deployment secret transforms remain promising directions our setup does not evaluate. Our intention is to support responsible risk assessment, and to encourage embedding providers and downstream users to treat image embeddings as sensitive data as the original images.

Acknowledgments

YC and JB are funded by the Carlsberg Foundation, under the Semper Ardens: Accelerate programme (project nr. CF21-0454) and by the Novo Nordisk Foundation, under the Ascending Data Science Investigator programme (grant no. NNF24OC0092972).

References

- Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. [no-caps: novel object captioning at scale](#). In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8947–8956.
- Guillaume Alain and Yoshua Bengio. 2017. [Understanding intermediate layers using linear classifier probes](#).
- Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. [Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks](#). *Preprint*, arXiv:2511.07025.
- Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2025. [The 2023 foundation model transparency index](#). *Transactions on Machine Learning Research*. Featured Certification.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. [A simple framework for contrastive learning of visual representations](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.
- Yiyi Chen, Russa Biswas, Heather Lent, and Johannes Bjerva. 2025a. [Against all odds: overcoming typography, script, and language confusion in multilingual embedding inversion attacks](#). In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’25/IAAI’25/EAAI’25*. AAAI Press.
- Yiyi Chen, Heather Lent, and Johannes Bjerva. 2024. [Text embedding inversion security for multilingual language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7808–7827, Bangkok, Thailand. Association for Computational Linguistics.
- Yiyi Chen, Qionghai Xu, and Johannes Bjerva. 2025b. [ALGEN: Few-shot inversion attacks on textual embeddings via cross-model alignment and generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24330–24348, Vienna, Austria. Association for Computational Linguistics.
- Yiyi Chen, Qionghai Xu, Desmond Elliott, Qionghai Li, and Johannes Bjerva. 2026a. [Semantic leakage from image embeddings](#). *Preprint*, arXiv:2601.22929.
- Yunhao Chen, Shujie Wang, Xin Wang, Ran He, Xingjun Ma, and Yu-Gang Jiang. 2026b. [Leaky-clip: Extracting training data from clip](#). *Preprint*, arXiv:2508.00756.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).
- Cohere. 2025. Cohere embed 4: Multimodal search and representation models. <https://cohere.com/blog/embed-4>. Accessed: 2026-05-26.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke

- Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*. *Preprint*, arXiv:2507.06261.
- Minxin Du, Xiang Yue, Sherman S. M. Chow, and Huan Sun. 2023. *Sanitizing sentence embeddings (and labels) for local differential privacy*. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2349–2359, New York, NY, USA. Association for Computing Machinery.
- Raj Hansini Khoiwal and Alan B. McMillan. 2024. *Embeddings are all you need! achieving high performance medical image classification through training-free embedding analysis*. *Preprint*, arXiv:2412.09445.
- Doohyun Kim, Donghwa Kang, Kyungjae Lee, Hyeonboo Baek, and Brent Byunghoon Kang. 2026. *Zero2text: Zero-training cross-domain inversion attacks on textual embeddings*. *Preprint*, arXiv:2602.01757.
- Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftexhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, and 28 others. 2025. *Gemini embedding: Generalizable embeddings from gemini*. *Preprint*, arXiv:2503.07891.
- Haoran Li, Mingshi Xu, and Yangqiu Song. 2023. *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*, pages 14022–14040, Toronto, Canada. Association for Computational Linguistics.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. *Microsoft coco: Common objects in context*. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.
- Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. *Deep learning face attributes in the wild*. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738.
- Ikram Messadi, Giulia Cervia, and Vincent Itier. 2025. *Image selective encryption analysis using mutual information in cnn based embedding space*. In *2025 13th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6.
- John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. *Text embeddings reveal (almost) as much as text*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. *Gpt-4 technical report*. *Preprint*, arXiv:2303.08774.
- Hatef Otroshi Shahreza, Laurent Colbois, and Sébastien Marcel. 2026. *On the generation of face morphs by inversion of optimal morph embeddings*. *IEEE Transactions on Information Forensics and Security*, 21:333–346.
- Hatef Otroshi Shahreza, Anjith George, and Sébastien Marcel. 2025. *Face reconstruction from face embeddings using adapter to a face foundation model*. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5584–5593.
- Kenneth Philbrick. 2026. Enhancing medical image discovery: similarity search in the imaging data commons archive. <https://developers.google.com/health-ai-developer-foundations/blogs/idc-image-search>. Accessed: 2026-05-19.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. *Learning transferable visual models from natural language supervision*. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. *Improving language understanding by generative pre-training*. *OpenAI*.
- Anton Razhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. 2023. *Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 286–295, Singapore. Association for Computational Linguistics.
- Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S. Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, Henning Müller, Peter A. Horn, Felix Nensa, and Christoph M. Friedrich. 2024. *Rocov2: Radiology objects in context version 2, an updated multimodal image dataset*. *Scientific Data*, 11(1).

- Congzheng Song and Ananth Raghunathan. 2020. [Information leakage in embedding models](#). In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, CCS '20*, page 377–390, New York, NY, USA. Association for Computing Machinery.
- Tyler Sorensen and Heidv Khlaaf. 2024. [Leftoverlocals: Listening to llm responses through leaked gpu local memory](#). *Preprint*, arXiv:2401.16603.
- Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. [ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17897–17907, Los Alamitos, CA, USA. IEEE Computer Society.
- The OWASP Foundation. 2025. [OWASP Top 10 for Large Language Model Applications](#).
- Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. 2018. [Split learning for health: Distributed deep learning without sharing raw patient data](#). *Preprint*, arXiv:1812.00564.
- Hanrui Wang, Shuo Wang, Chun-Shien Lu, and Isao Echizen. 2026. [Diffmi: Breaking face recognition privacy via diffusion-driven training-free model inversion](#). *IEEE Transactions on Information Forensics and Security*, 21:4275–4290.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Kedong Xiu and Sai Qian Zhang. 2025. [Caprecover: A cross-modality feature inversion attack framework on vision language models](#). In *Proceedings of the 33rd ACM International Conference on Multimedia, MM '25*, page 3808–3816, New York, NY, USA. Association for Computing Machinery.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. [From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions](#). *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Wenrui Yu, Yiyi Chen, Johannes Bjerva, Sokol Kosta, and Qiongxu Li. 2025. [Lago: Few-shot crosslingual embedding inversion attacks via language similarity-aware graph optimization](#). *Preprint*, arXiv:2505.16008.
- Collin Zhang, John X. Morris, and Vitaly Shmatikov. 2025. [Universal zero-shot embedding inversion](#). *Preprint*, arXiv:2504.00147.

Table 6: Details of the Models

Model	Usage	Dimension	# Parameters	Platform	Access	Name
CLIP	Attack Dual Encoder	512	151M	Huggingface	Open-source	openai/clip-vit-base-patch32
FLAN-T5 (Chung et al., 2024)	Attack LM f_θ	512	77M	Huggingface	Open-source	google/flan-t5-small
CLIP	Victim Embedder	768	428M	Huggingface	Open-source	openai/clip-vit-large-patch14
NOMIC	Victim Embedder	768	92.9M	Huggingface	Open-Source	nomic-ai/nomic-embed-vision-v1.5
GEMINI2	Victim Embedder	3072	-	Vertex AI	Proprietary	google/gemini-embedding-2-preview
COHERE	Victim Embedder	1536	-	Cohere	Proprietary	embed-v4.0
NVIDIA (Babakhin et al., 2025)	Victim Embedder	2048	-	OpenRouter	Proprietary	nvidia/llama-nemotron-embed-v1-1b-v2:free
DEEPSEEK-V3.2	LLM	-	685 B	OpenRouter	Open-source	deepseek/deepseek-v3.2
GPT-4o	VLM	-	-	OpenRouter	Proprietary	openai/gpt-4o
KANDINSKY 2.2 (Razhigaev et al., 2023)	Stable Diffusion Model	-	271B	Huggingface	Open-Source	kandinsky-community/kandinsky-2-2-decoder
CLIP in KANDINSKY 2.2	Dual Encoder	1280	1.8B	Huggingface	Open-Source	ViT-bigG-14-laion2B-39B-b160

Table 7: Datasets and their Tasks. Caption-recovery uses a held-out set of 500 images for evaluation; attribute tasks use the dataset’s standard validation split. NoCaps has no training split; the COCO-trained generator f_θ is used for evaluating NoCaps. Alignment pairs $\mathcal{D}_N$ are sampled from the training split at sizes $N \in \{1, 10, 10^2, 10^3, 10^4\}$.

Dataset	Domain	Attack	Labels / classes	# Train	# Eval	Notes
COCO (Lin et al., 2014)	Natural	Caption recovery	5 Ref. Captions / Image	113,287	500	$\mathcal{D}_{pub}$, In-Domain Eval
Flickr30K (Young et al., 2014)	Natural	Caption recovery	5 Ref. Captions / Image	29,000	500	$\mathcal{D}_{pub}$, In-Domain Eval
NoCaps (Agrawal et al., 2019)	Natural (OOD)	Caption recovery	10 Ref. Captions / Image	-	500	Coss-domain Eval
ROCOv2 (Rückert et al., 2024)	Medical	Modality + CUI	6 modalities; top- K UMLS concepts	1000 / 59,962	1000 / 9,904	PMC open-access radiology
CelebA (Liu et al., 2015)	Face	Attribute	40 binary facial attributes	1000 / 162,770	1000 / 19,962	open-source victims only

Table 8: Ablation study of K using COCO, where $N = 1000$. The best value per victim embedder is in bold.

Victim	K	BLEU	Rouge-L	Rouge-1	Rouge-2
CLIP	1	24.11	51.43	56.35	29.14
	3	25.78	51.82	56.14	30.02
	5	25.28	51.85	56.22	29.84
	8	25.11	51.56	56.26	29.81
	10	25.24	51.66	56.40	29.99
NOMIC	1	25.15	51.71	56.61	30.32
	3	26.37	51.80	56.61	30.40
	5	26.33	52.05	57.00	30.89
	8	25.88	51.63	56.51	30.56
	10	25.28	51.17	56.07	30.04
COHERE	1	26.12	51.75	56.31	30.75
	3	26.63	51.74	56.23	30.74
	5	26.32	51.92	56.29	30.69
	8	26.93	52.14	56.70	30.83
	10	27.08	52.47	56.98	31.35
GEMINI2	1	26.23	51.63	56.06	30.79
	3	27.42	52.60	56.98	31.61
	5	26.68	52.56	57.14	31.51
	8	26.75	52.66	57.19	31.46
	10	26.59	52.31	56.96	30.83
NVIDIA	1	25.36	51.52	56.47	30.41
	3	27.02	52.87	57.70	32.02
	5	26.96	52.65	57.73	31.69
	8	27.41	52.91	57.78	31.74
	10	27.54	53.30	58.01	32.42

Table 9: Sensitivity of the attack to the alignment sample size N in the few-shot setting. For each victim and each N , the alignment $\mathbf{W}$ is refit on 10 independent random subsets of image-embedding pairs, and the caption-recovery attack is re-run for every victim embedding; we report the mean $\pm$ 95% confidence interval over the 10 runs.

Victim	$N=1$	$N=10$	$N=10^2$	$N=10^3$
(a) Alignment cosine on held-out validation pairs				
CLIP	0.441 $\pm$ 0.029	0.660 $\pm$ 0.007	0.771 $\pm$ 0.002	0.788 $\pm$ 0.000
NOMIC	0.441 $\pm$ 0.029	0.645 $\pm$ 0.008	0.734 $\pm$ 0.002	0.791 $\pm$ 0.000
COHERE	0.423 $\pm$ 0.034	0.636 $\pm$ 0.010	0.739 $\pm$ 0.002	0.724 $\pm$ 0.000
GEMINI2	0.441 $\pm$ 0.029	0.652 $\pm$ 0.007	0.758 $\pm$ 0.002	0.739 $\pm$ 0.000
NVIDIA	0.441 $\pm$ 0.029	0.650 $\pm$ 0.008	0.743 $\pm$ 0.002	0.741 $\pm$ 0.000
(b) Rouge-L of the recovered caption (CLIP rerank)				
CLIP	28.68 $\pm$ 1.73	33.80 $\pm$ 1.31	46.71 $\pm$ 0.51	51.85 $\pm$ 0.00
NOMIC	28.19 $\pm$ 2.31	33.74 $\pm$ 1.18	46.27 $\pm$ 0.53	52.07 $\pm$ 0.01
COHERE	29.68 $\pm$ 2.22	33.60 $\pm$ 1.22	48.11 $\pm$ 0.45	51.92 $\pm$ 0.00
GEMINI2	27.95 $\pm$ 2.16	33.17 $\pm$ 1.74	46.84 $\pm$ 0.52	52.56 $\pm$ 0.00
NVIDIA	30.25 $\pm$ 1.03	34.14 $\pm$ 1.37	48.25 $\pm$ 0.48	52.65 $\pm$ 0.00

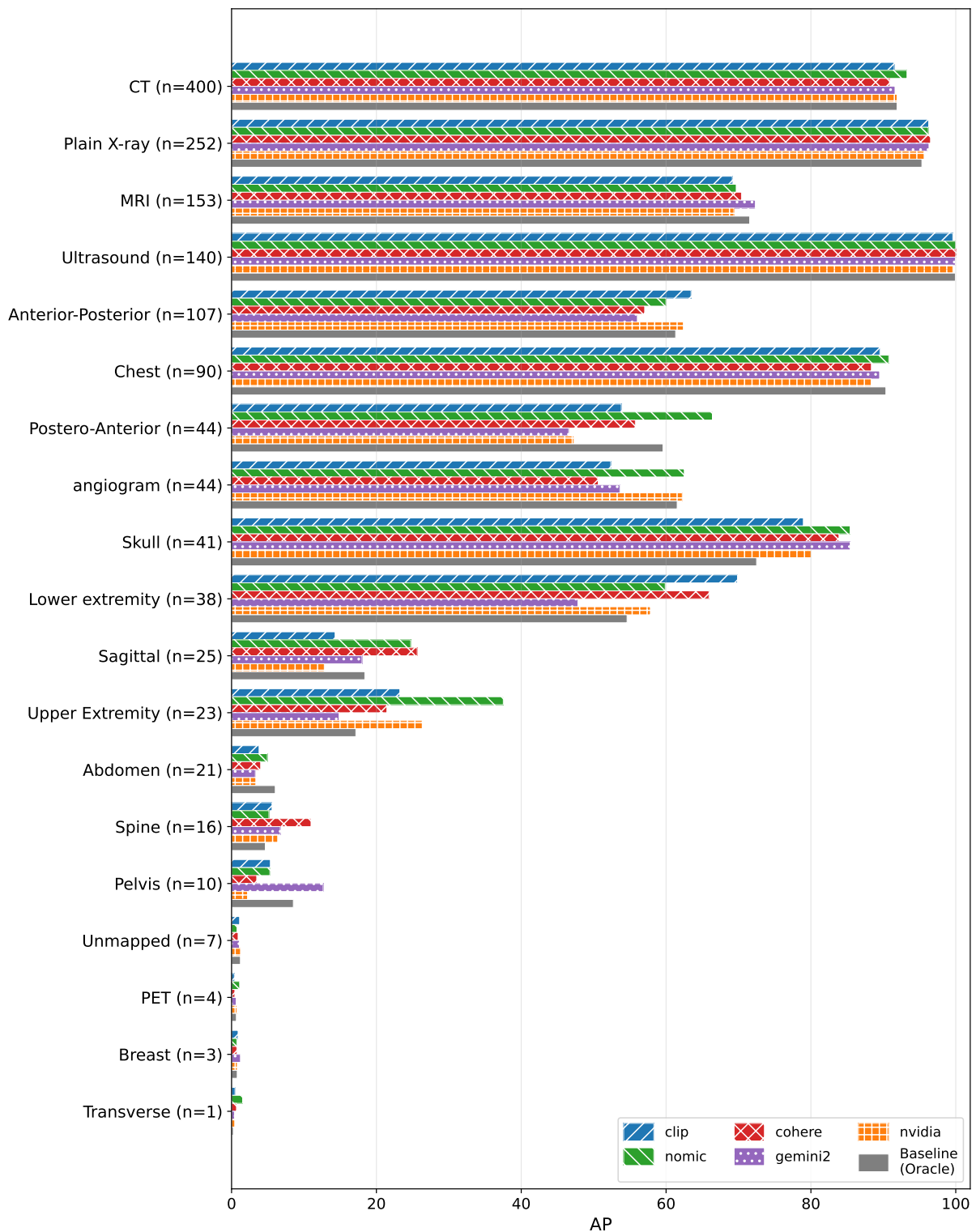


Figure 5: Attribute Attacks on RocoV2 CUI per Class in AP, results on a random subset of 1000 samples.

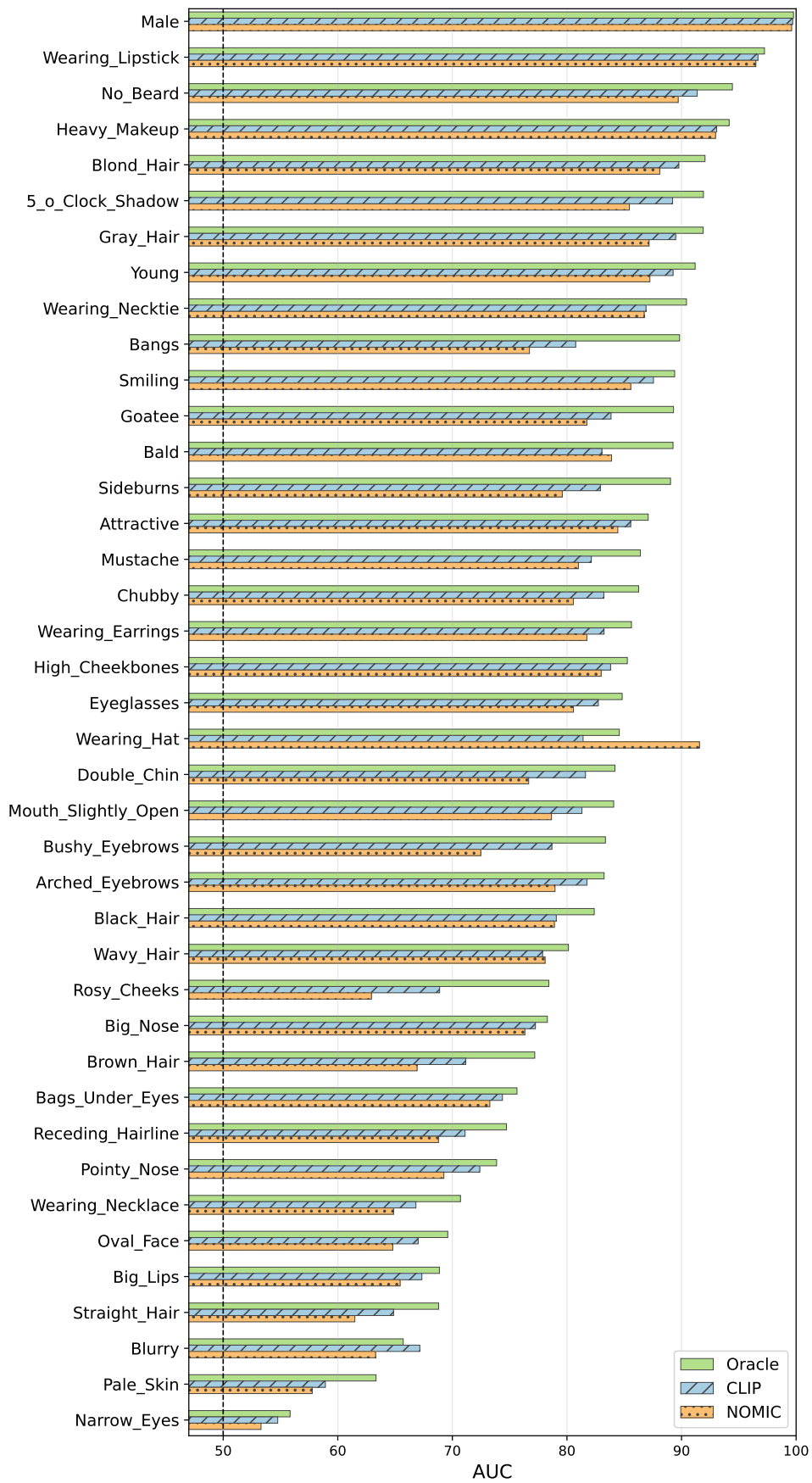


Figure 6: Attribute Attacks on CelebA per Class with ROC-AUC (AUC) with Linear Probe.



Figure 7: NoCaps source image; reference caption: “a man on a windy beach flying a kite”.

Table 19: Qualitative Cross-Domain Caption Recovery for Fig. 7 with CLIP-Rerank. At $N=10^4$, all five victim encoders produce the identical recovered caption.

Victim	$N = 1$	$N = 10$	$N = 10^2$	$N = 10^3$	$N = 10^4$
CLIP	A person riding a snowboard down a snow covered slope.	A man riding a skateboard down a snow covered slope.	A man standing on a beach flying a kite.	A man flying a kite on top of a beach.	A man standing on a beach flying a kite.
COHERE	A group of people sitting on top of a skateboard.	A man sitting on top of a skateboard in a park.	A man riding a skateboard on top of a sandy beach.	A man standing on a beach flying a kite.	A man standing on a beach flying a kite.
GEMINI2	A person riding a snowboard down a snow covered slope.	A man sitting on top of a wooden bench.	A person on a beach flying a kite.	A man flying a kite in the middle of a field.	A man standing on a beach flying a kite.
NOMIC	A person riding a snowboard down a snow covered slope.	A man riding a snowboard down a snow covered slope.	A man riding a board on top of a sandy beach.	A man standing on a beach flying a kite.	A man standing on a beach flying a kite.
NVIDIA	A man riding a snowboard down a snow covered slope.	A man standing on top of a wooden bench.	A man standing on top of a sandy beach.	A man standing on a beach flying a kite.	A man standing on a beach flying a kite.

Table 10: Vec2Text baseline results on the COCO dataset.

victim	N	BLEU	Rouge-L	Rouge-1	Rouge-2
RANDOM	1	0.04	4.61	4.79	0.09
	10	0.04	4.64	4.81	0.09
	100	0.04	4.44	4.56	0.05
	1000	0.15	8.15	8.39	0.60
	10000	1.40	18.74	19.06	1.89
CLIP	1	0.04	6.78	6.85	0.09
	10	0.08	6.88	6.96	0.17
	100	0.08	7.87	7.97	0.17
	1000	0.22	16.87	17.61	2.32
	10000	11.31	38.12	41.73	15.58
COHERE	1	0.06	3.43	3.64	0.13
	10	0.06	3.36	3.54	0.14
	100	0.03	3.18	3.36	0.09
	1000	0.28	8.12	8.36	0.55
	10000	12.40	40.54	44.60	17.36
GEMINI2	1	0.04	4.09	4.27	0.11
	10	0.04	4.04	4.19	0.08
	100	0.04	3.90	4.03	0.07
	1000	0.06	8.13	8.62	0.42
	10000	11.65	39.19	43.01	16.99
NOMIC	1	0.03	5.03	5.10	0.09
	10	0.03	5.01	5.06	0.10
	100	0.04	5.97	6.12	0.13
	1000	1.27	18.17	18.36	1.83
	10000	2.40	24.19	25.44	5.76
NVIDIA	1	0.03	5.21	5.38	0.03
	10	0.03	5.57	5.74	0.03
	100	0.04	5.95	6.12	0.06
	1000	0.17	7.77	8.14	0.38
	10000	11.02	37.74	41.26	15.41

Table 11: Text Reconstruction from Image Embeddings Across Victim Embedders on the COCO dataset. The best performances for each victim embedding are **bolded** across metrics .

Victim	N	Decoding	BLEU	Rouge-L	Rouge-1	Rouge-2
CLIP	1	Base	2.61	26.69	27.04	3.86
		+Rerank [CLIP]	2.28	26.56	26.92	3.24
	10	Base	8.80	35.98	39.63	13.75
		+Rerank [CLIP]	8.79	35.61	39.22	13.13
	100	Base	20.09	46.00	50.57	24.17
		+Rerank [CLIP]	18.98	46.05	50.50	24.16
	1000	Base	24.03	50.52	55.11	28.82
		+Rerank [CLIP]	25.28	51.85	56.22	29.84
		+CoT	27.20	53.08	57.20	31.42
	10000	Base	27.67	53.14	57.70	31.94
		+Rerank [CLIP]	29.82	54.30	58.94	33.37
		+CoT	30.42	54.95	59.25	33.86
COHERE	1	Base	7.44	32.86	35.53	10.96
		+Rerank [CLIP]	7.48	32.33	34.96	10.29
	10	Base	9.93	35.72	39.32	14.44
		+Rerank [CLIP]	9.59	35.66	39.23	14.11
	100	Base	20.89	46.71	51.05	25.10
		+Rerank [CLIP]	22.4	48.36	52.86	27.02
	1000	Base	25.62	50.99	55.67	29.68
		+Rerank [CLIP]	26.32	51.92	56.29	30.69
		+CoT	28.40	53.01	57.37	31.88
	10000	Base	29.06	54.05	58.79	33.64
		+Rerank [CLIP]	30.12	55.55	60.39	35.05
		+CoT	32.07	56.09	60.51	35.42
GEMINI2	1	Base	2.50	24.75	25.03	3.81
		+Rerank [CLIP]	2.11	26.17	26.57	2.72
	10	Base	8.55	36.08	39.50	13.55
		+Rerank [CLIP]	8.06	35.49	39.15	12.82
	100	Base	19.28	45.85	50.87	24.33
		+Rerank [CLIP]	20.88	47.62	52.68	25.86
	1000	Base	25.56	51.30	55.87	29.98
		+Rerank [CLIP]	26.68	52.56	57.14	31.51
		+CoT	27.92	53.06	57.57	31.92
	10000	Base	28.53	53.63	58.75	32.90
		+Rerank [CLIP]	31.05	55.40	60.46	35.01
		+CoT	31.97	55.88	60.57	35.33
NOMIC	1	Base	2.73	24.70	24.98	3.84
		+Rerank [CLIP]	2.11	26.17	26.57	2.72
	10	Base	8.94	35.42	39.07	13.38
		+Rerank [CLIP]	8.31	35.06	38.55	12.58
	100	Base	19.06	45.33	49.78	23.07
		+Rerank [CLIP]	19.6	45.87	49.98	23.47
	1000	Base	24.96	50.64	55.43	29.29
		+Rerank [CLIP]	26.33	52.05	57.00	30.89
		+CoT	27.64	52.43	57.00	31.39
	10000	Base	26.79	52.39	57.22	31.69
		+Rerank [CLIP]	29.00	54.24	59.11	33.64
		+CoT	29.92	54.92	59.40	33.59
NVIDIA	1	Base	3.08	27.81	28.36	4.47
		+Rerank [CLIP]	2.69	27.46	28.01	4.09
	10	Base	9.25	36.23	39.68	13.98
		+Rerank [CLIP]	9.04	35.92	39.25	13.38
	100	Base	20.75	47.26	51.66	25.37
		+Rerank [CLIP]	21.38	47.72	52.30	25.83
	1000	Base	26.71	51.85	56.47	30.96
		+Rerank [CLIP]	26.96	52.65	57.73	31.69
		+CoT	28.92	53.57	58.11	32.54
	10000	Base	27.95	53.01	58.22	32.66
		+Rerank [CLIP]	29.65	55.15	60.21	34.31
		+CoT	31.79	56.02	60.75	35.59

Table 12: Cross-Domain Image Embedding Inversion Across Victim Embedders on NoCaps.

victim	N	Decoding	BLEU	Rouge-L	Rouge-1	Rouge-2
CLIP	1	Base	0.95	25.25	25.71	2.57
		+Rerank [CLIP]	0.54	25.28	25.61	2.40
	10	Base	6.94	33.61	37.42	12.15
		+Rerank [CLIP]	7.23	33.96	37.69	12.38
	100	Base	12.25	38.79	42.02	16.65
		+Rerank [CLIP]	11.97	38.96	42.50	16.89
	1000	Base	15.26	42.20	45.81	20.39
		+Rerank [CLIP]	16.84	43.13	47.14	21.60
	10000	Base	19.45	45.12	48.42	23.79
		+Rerank [CLIP]	20.95	46.42	49.96	24.88
COHERE	1	Base	6.85	31.14	33.78	10.28
		+Rerank [CLIP]	6.94	30.98	33.21	9.35
	10	Base	7.99	34.00	37.70	12.54
		+Rerank [CLIP]	7.56	33.85	37.11	11.85
	100	Base	11.91	38.57	41.79	16.73
		+Rerank [CLIP]	13.21	39.05	42.66	17.51
	1000	Base	14.44	40.72	44.35	18.64
		+Rerank [CLIP]	15.32	41.14	45.38	19.57
	10000	Base	17.26	44.26	48.04	22.12
		+Rerank [CLIP]	20.04	45.56	49.22	24.07
GEMINI2	1	Base	0.67	23.62	24.00	2.48
		+Rerank [CLIP]	0.52	25.00	25.33	2.17
	10	Base	6.63	34.01	37.9	13.01
		+Rerank [CLIP]	6.49	33.48	37.03	12.34
	100	Base	10.89	38.17	41.39	16.04
		+Rerank [CLIP]	12.21	38.67	42.19	16.80
	1000	Base	14.47	41.35	44.45	18.95
		+Rerank [CLIP]	14.96	41.86	45.23	20.06
	10000	Base	18.92	44.70	48.05	23.22
		+Rerank [CLIP]	20.93	46.53	50.19	25.02
NOMIC	1	Base	0.52	23.62	23.96	2.44
		+Rerank [CLIP]	0.52	25.00	25.33	2.17
	10	Base	6.76	33.61	37.24	11.66
		+Rerank [CLIP]	6.56	33.11	36.46	11.07
	100	Base	11.68	38.17	41.38	16.40
		+Rerank [CLIP]	11.27	38.14	41.22	16.24
	1000	Base	15.14	41.35	44.71	19.30
		+Rerank [CLIP]	15.88	42.16	45.69	20.30
	10000	Base	17.52	44.08	47.26	21.84
		+Rerank [CLIP]	19.13	45.37	48.78	23.15
NVIDIA	1	Base	1.23	27.41	28.91	4.27
		+Rerank [CLIP]	1.49	26.82	27.93	3.76
	10	Base	7.41	34.60	38.19	12.81
		+Rerank [CLIP]	7.08	33.98	37.26	12.41
	100	Base	12.58	38.97	42.44	17.19
		+Rerank [CLIP]	13.08	39.95	43.79	18.11
	1000	Base	16.57	42.45	45.91	20.62
		+Rerank [CLIP]	17.13	42.92	46.68	21.73
	10000	Base	20.42	46.01	49.38	24.12
		+Rerank [CLIP]	22.73	47.67	51.24	25.99

Table 13: Near-Domain and Out-Domain Caption Recovery on NoCaps across Victim Embeddings with CLIP-Rerank.

Victim	N	BLEU	Rouge-L	Rouge-1	Rouge-2
Near-Domain					
CLIP	1	0.90	27.32	27.77	4.27
	10	6.61	33.78	37.24	11.56
	100	12.25	39.58	42.98	17.47
	1000	19.43	45.20	49.16	24.10
	10000	25.18	49.50	52.76	28.47
NOMIC	1	0.87	26.71	27.15	3.80
	10	7.35	33.66	36.75	11.83
	100	12.36	38.81	41.70	16.67
	1000	19.36	44.63	48.00	22.70
	10000	22.65	47.44	50.92	26.01
COHERE	1	7.79	31.05	33.22	10.12
	10	9.16	35.29	38.35	13.53
	100	15.50	40.85	43.99	19.23
	1000	17.89	42.70	46.81	21.47
	10000	22.41	46.84	50.43	26.01
GEMINI2	1	0.87	26.71	27.15	3.80
	10	5.60	32.90	36.53	11.18
	100	14.11	39.09	42.66	17.79
	1000	17.03	43.50	47.06	21.76
	10000	24.85	48.88	52.42	27.67
NVIDIA	1	0.98	28.49	29.37	5.31
	10	6.39	33.53	36.99	11.76
	100	13.53	40.51	44.23	18.32
	1000	20.35	45.57	49.46	25.12
	10000	26.59	50.54	53.98	29.58
Out-Domain					
CLIP	1	0.25	23.25	23.46	0.53
	10	7.82	34.14	38.15	13.21
	100	11.68	38.35	42.02	16.31
	1000	14.08	41.06	45.13	19.10
	10000	16.39	43.33	47.16	21.29
NOMIC	1	0.26	23.29	23.5	0.53
	10	5.72	32.56	36.17	10.31
	100	10.18	37.46	40.73	15.80
	1000	12.11	39.70	43.37	17.90
	10000	15.39	43.30	46.65	20.29
COHERE	1	6.04	30.9	33.2	8.57
	10	5.79	32.40	35.87	10.18
	100	10.75	37.26	41.33	15.78
	1000	12.53	39.57	43.94	17.68
	10000	17.50	44.28	48.01	22.13
GEMINI2	1	0.26	23.29	23.50	0.53
	10	7.33	34.06	37.52	13.50
	100	10.12	38.26	41.73	15.80
	1000	12.80	40.22	43.4	18.36
	10000	16.73	44.18	47.96	22.38
NVIDIA	1	1.70	25.15	26.49	2.22
	10	7.74	34.43	37.54	13.05
	100	12.61	39.39	43.35	17.91
	1000	13.75	40.28	43.90	18.33
	10000	18.56	44.80	48.51	22.40

Table 14: Attribute Attacks on RocoV2 CUI.

Embedder	N	mAP	F1
Linear Probe			
$e_{i,A}$	-	42.96	3.19
CLIP	1	8.99	0.00
	10	34.19	0.00
	100	44.13	2.51
	1000	43.14	4.14
COHERE	1	8.57	0.00
	10	33.18	7.85
	100	47.37	3.60
	1000	43.56	3.33
GEMINI2	1	7.87	0.00
	10	30.07	0.00
	100	43.68	1.82
	1000	42.00	2.78
NOMIC	1	9.15	0.00
	10	37.48	2.37
	100	42.08	3.60
	1000	45.57	2.51
NVIDIA	1	8.31	0.00
	10	30.80	5.62
	100	43.53	3.19
	1000	42.57	3.87
MLP Probe			
$e_{i,A}$	-	38.35	67.00
CLIP	1	6.75	33.07
	10	29.46	32.66
	100	37.39	64.94
	1000	36.92	65.72
COHERE	1	11.73	33.07
	10	33.49	50.99
	100	39.12	67.53
	1000	35.67	64.79
GEMINI2	1	6.42	33.07
	10	31.23	33.88
	100	37.26	60.90
	1000	34.61	61.02
NOMIC	1	6.85	33.07
	10	35.2	45.04
	100	39.11	65.88
	1000	38.78	66.52
NVIDIA	1	14.51	33.07
	10	31.95	36.41
	100	37.38	64.96
	1000	36.03	63.51

Table 15: Dataset COCO, Utility and Defense, from the perspective of consumers and service providers, the retrieval performance the higher the better $\uparrow$, and the inversion performance the lower the better $\downarrow$. (N=1000)

Embedder	Defense	Parameter	R@1 $\uparrow$	R@5 $\uparrow$	R@10 $\uparrow$	COS $\downarrow$	BLEU $\downarrow$	Rouge-L $\downarrow$	Rouge-1 $\downarrow$	Rouge-2 $\downarrow$
CLIP	No Defense	-	84.00	98.00	99.40	-	25.28	51.85	56.22	29.84
	Gaussian Noise	0.001	84.00	98.00	99.40	0.7874	25.51	51.65	56.09	29.90
		0.01	83.40	97.60	99.20	0.7363	24.52	50.53	54.75	28.88
		0.1	20.60	39.20	52.60	0.4334	7.26	32.53	35.54	10.63
		1	0.00	0.80	2.20	0.0926	5.04	30.56	33.02	7.93
	LapMech	0.1	1.40	4.80	7.40	0.2371	4.67	30.31	32.64	7.76
		0.25	8.40	19.20	28.40	0.3704	4.77	31.17	33.90	8.18
		0.5	34.60	64.80	77.60	0.4771	7.67	34.26	37.35	11.90
		1	68.00	92.80	95.80	0.5905	14.21	41.54	45.57	19.02
		2	80.00	96.00	99.20	0.6811	20.02	47.40	52.28	25.10
		4	82.20	97.40	99.40	0.7433	24.45	50.85	55.38	28.71
		8	84.00	97.60	99.20	0.7754	25.28	51.85	56.30	29.85
	PurMech	0.1	0.60	3.80	7.00	0.1639	5.31	30.98	33.47	8.46
		0.25	6.80	19.60	28.40	0.3845	6.52	32.37	34.96	9.96
		0.5	35.40	56.60	68.60	0.4825	8.03	34.31	37.66	12.09
		1	70.20	93.00	96.60	0.5882	14.51	41.50	45.76	19.46
		2	81.00	96.80	99.00	0.6807	20.25	47.37	51.97	25.23
		4	83.60	97.60	99.20	0.7435	24.38	50.49	54.50	28.57
		8	84.00	98.00	99.20	0.7754	25.68	51.67	55.90	30.11
COHERE	No Defense	-	94.40	99.60	100.00	-	26.32	51.92	56.29	30.69
	Gaussian Noise	0.001	94.40	99.60	100.00	0.7241	26.67	52.04	56.56	30.95
		0.01	93.80	99.60	100.00	0.6994	24.76	51.20	55.85	29.67
		0.1	79.40	96.60	99.20	0.4289	6.45	34.10	37.23	11.16
		1	1.40	2.40	3.40	0.0378	4.15	29.76	32.52	7.35
	LapMech	0.1	14.00	33.80	44.80	0.1866	4.95	30.37	33.13	7.63
		0.25	80.00	96.00	99.20	0.4163	8.22	34.17	36.96	11.55
		0.5	91.40	99.20	100.00	0.5560	14.76	41.92	46.09	19.43
		1	94.00	99.60	100.00	0.6431	22.11	48.50	53.14	26.78
		2	94.40	99.60	100.00	0.6902	25.51	51.51	56.14	30.06
		4	94.40	99.60	100.00	0.7134	25.22	51.23	55.81	29.73
		8	94.40	99.60	100.00	0.7212	27.03	52.63	56.85	31.39
	PurMech	0.1	11.80	29.40	40.00	0.0167	4.48	30.37	33.10	8.06
		0.25	57.00	69.80	71.80	0.4254	6.73	34.03	37.17	10.95
		0.5	91.40	99.40	100.00	0.5566	14.50	41.94	46.04	19.68
		1	93.40	99.80	100.00	0.6400	22.56	49.32	54.00	27.55
		2	94.60	99.80	100.00	0.6886	24.62	51.39	56.22	30.08
		4	94.00	99.60	100.00	0.7125	26.28	52.22	56.91	30.86
		8	94.20	99.60	100.00	0.7222	26.99	52.29	56.77	31.13
GEMINI2	No Defense	-	93.80	100.00	100.00	-	26.68	52.56	57.14	31.51
	Gaussian Noise	0.001	94.00	100.00	100.00	0.7468	27.28	52.80	57.41	31.25
		0.01	94.40	99.80	100.00	0.7796	26.93	52.68	57.28	30.75
		0.1	31.80	63.20	75.20	0.5717	5.92	31.86	34.84	9.30
		1	0.80	1.60	2.80	0.0881	4.76	30.88	33.04	7.94
	LapMech	0.1	8.60	21.60	29.60	0.4973	5.27	30.86	33.65	8.22
		0.25	63.20	89.40	94.20	0.6006	7.98	34.24	37.34	11.72
		0.5	88.60	99.00	100.00	0.6865	15.09	41.99	46.17	19.88
		1	92.40	99.60	100.00	0.7530	24.39	50.12	54.80	28.68
		2	92.80	99.80	100.00	0.7809	27.59	52.77	57.41	31.46
		4	93.80	100.00	100.00	0.7787	28.79	53.93	58.67	32.85
		8	93.80	100.00	100.00	0.7622	28.43	53.65	58.32	32.57
	PurMech	0.1	6.60	15.80	23.00	0.1022	8.25	33.35	36.10	10.87
		0.25	58.80	87.60	92.00	0.6077	7.65	33.87	37.15	11.74
		0.5	86.60	98.20	100.00	0.6906	14.57	41.82	46.11	19.58
		1	92.80	99.60	100.00	0.7544	24.00	50.12	54.68	28.24
		2	94.00	100.00	100.00	0.7815	27.80	53.13	57.77	31.43
		4	94.20	100.00	100.00	0.7797	28.12	53.69	58.46	32.04
		8	94.20	100.00	100.00	0.7595	27.78	52.95	57.53	31.72
NOMIC	No Defense	-	73.20	92.40	97.00	-	26.33	52.05	57.00	30.89
	Gaussian Noise	0.001	73.80	92.60	97.00	0.7899	25.36	51.52	56.93	30.25
		0.01	64.40	90.20	95.00	0.6997	24.28	50.43	54.83	28.67
		0.05	7.40	20.80	31.40	0.4778	8.84	35.72	39.55	13.10
		0.1	1.00	4.60	8.40	0.3985	5.84	31.68	34.77	9.53
		1	0.00	1.20	1.40	0.1115	4.74	30.99	33.23	7.92
	LapMech	0.1	1.00	1.60	3.40	0.2867	4.11	30.47	33.01	7.52
		0.25	1.20	3.60	7.40	0.3537	5.27	31.06	33.64	8.46
		0.5	4.00	11.60	16.60	0.4207	7.22	32.95	35.66	10.30
		1	18.20	41.40	52.60	0.5174	10.36	37.37	40.71	15.19
		2	53.60	78.60	85.20	0.6221	18.22	45.88	50.27	23.02
		4	65.60	90.00	95.20	0.7113	22.74	49.90	54.17	27.55
		8	72.20	92.80	97.00	0.7638	24.67	51.01	55.76	28.79
	PurMech	0.1	0.20	1.60	3.20	0.1361	5.27	31.21	33.42	8.39
		0.25	0.40	3.40	4.80	0.3656	4.80	30.91	33.58	8.30
		0.5	2.60	8.80	14.00	0.4330	7.15	33.38	36.65	11.11
		1	17.00	38.00	51.20	0.5212	10.75	38.38	42.31	15.82
		2	49.40	79.00	88.20	0.6220	18.62	45.62	50.22	23.57
		4	66.80	90.80	95.60	0.7109	23.87	50.37	54.87	28.41
		8	71.40	92.40	97.20	0.7632	25.66	52.00	56.75	30.49
NVIDIA	No Defense	-	79.80	97.40	98.60	-	26.96	52.65	57.73	31.69
	Gaussian Noise	0.001	79.60	97.60	98.60	0.7421	27.29	52.76	57.89	32.09
		0.01	79.00	97.00	98.80	0.7350	26.34	51.75	56.52	30.69
		0.05	70.60	91.80	97.40	0.6321	16.57	43.25	47.20	20.58
		0.1	41.60	72.00	82.80	0.5446	9.57	35.92	39.37	13.47
		1	0.40	2.20	4.20	0.0433	4.46	30.57	33.27	7.61
	LapMech	0.1	7.40	20.20	30.20	0.3829	5.77	31.59	34.31	9.03
		0.25	50.20	80.40	89.40	0.5523	10.62	36.62	40.24	14.16
		0.5	72.40	92.60	97.60	0.6413	17.93	44.89	48.87	22.77
		1	78.60	96.20	98.60	0.6994	23.46	49.77	54.33	28.39
		2	79.40	96.80	98.60	0.7300	25.95	51.64	56.44	30.40
		4	79.80	97.20	98.60	0.7423	27.54	53.18	57.77	31.99
		8	79.80	97.00	98.60	0.7435	27.30	52.90	57.90	32.08
	PurMech	0.1	5.00	14.40	23.60	0.0002	5.73	32.04	34.78	9.50
		0.25	26.80	46.40	55.00	0.5594	9.91	36.60	40.26	14.34
		0.5	72.40	93.40	97.60	0.6443	17.67	44.57	48.99	22.28
		1	78.00	96.40	98.60	0.7015	24.26	49.63	54.56	28.08
		2	79.00	97.00	98.80	0.7323	26.37	51.79	56.50	30.75
		4	78.80	97.00	98.60	0.7437	27.49	52.67	57.31	31.67
		8	79.60	97.40	98.60	0.7441	27.42	52.88	57.69	32.04

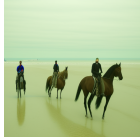
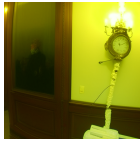
Table 16: Dataset Flickr30k, Utility and Defense, from the perspective of consumers and service providers, the retrieval performance the higher the better $\uparrow$, and the inversion performance the lower the better $\downarrow$. (N=1000)

Embedder	Defense	Parameter	R@1 $\uparrow$	R@5 $\uparrow$	R@10 $\uparrow$	COS $\downarrow$	BLEU $\downarrow$	Rouge-L $\downarrow$	Rouge-1 $\downarrow$	Rouge-2 $\downarrow$
CLIP	No Defense	-	83.10	97.00	99.00	-	20.82	46.66	50.55	26.15
	Gaussian Noise	0.001	83.00	97.00	99.00	0.7323	20.28	45.68	49.49	25.65
		0.01	82.90	97.00	99.00	0.6469	16.13	41.23	45.16	20.79
		0.1	13.80	30.60	40.30	0.4149	7.68	31.05	34.52	11.73
		1	0.00	0.70	2.20	0.0972	6.41	29.84	32.64	9.40
	LapMech	0.1	0.30	2.20	4.00	0.2355	5.61	29.23	32.44	9.72
		0.25	4.00	13.80	20.40	0.3618	5.58	29.43	32.74	9.83
		0.5	29.30	53.50	64.40	0.4512	7.78	32.20	35.49	12.15
		1	66.50	89.20	94.50	0.5433	11.72	35.87	39.40	16.48
		2	79.50	96.20	98.10	0.6102	14.43	39.77	43.78	19.91
		4	82.80	97.20	99.20	0.6516	16.64	41.64	45.22	21.36
		8	83.90	97.20	98.90	0.6832	18.46	43.58	47.46	23.23
	PurMech	0.1	0.20	2.20	3.50	0.1637	5.36	28.63	31.64	8.69
		0.25	3.60	12.80	20.00	0.3704	6.28	29.79	33.13	10.43
		0.5	25.50	48.50	58.70	0.4547	8.00	31.88	35.08	12.15
		1	66.00	89.40	94.50	0.5372	10.13	34.70	38.22	14.87
		2	79.90	96.70	98.50	0.6061	14.38	39.09	43.14	19.19
		4	83.10	97.20	99.00	0.6517	16.57	41.49	45.51	21.38
		8	83.10	97.40	98.80	0.6838	17.86	43.38	47.32	22.74
COHERE	No Defense	-	95.80	99.80	100.00	0.6519	19.63	44.45	48.06	24.06
	Gaussian Noise	0.001	95.60	99.80	100.00	0.6515	20.21	45.06	48.74	24.33
		0.01	94.80	99.80	100.00	0.6563	19.52	45.47	48.84	24.87
		0.1	76.20	93.60	96.30	0.4358	8.47	32.99	36.20	12.93
		1	0.50	1.30	2.60	0.0418	5.90	28.87	31.85	9.66
	LapMech	0.1	10.90	26.50	36.50	0.2114	6.86	30.40	33.50	10.85
		0.25	77.00	93.70	96.10	0.4332	8.86	33.04	36.62	13.53
		0.5	92.40	99.60	99.90	0.5500	13.15	38.45	42.02	18.52
		1	94.30	99.60	99.90	0.6186	17.59	42.62	46.01	22.17
		2	94.70	99.60	100.00	0.6510	18.84	44.40	48.19	23.63
		4	95.40	99.70	100.00	0.6584	19.78	44.79	48.28	24.47
		8	95.50	99.70	100.00	0.6551	20.11	44.82	48.11	24.28
	PurMech	0.1	8.70	24.10	33.40	0.0174	6.22	29.30	32.03	9.79
		0.25	55.60	68.60	70.30	0.4310	7.92	32.69	36.16	12.54
		0.5	91.10	99.20	99.80	0.5473	13.08	37.86	41.62	18.15
		1	93.30	99.90	99.90	0.6167	17.03	41.75	45.59	21.65
		2	94.90	99.70	100.00	0.6510	19.42	44.91	48.58	24.67
		4	94.90	99.80	100.00	0.6585	20.06	45.11	48.92	25.07
		8	95.40	99.80	100.00	0.6538	20.16	45.13	48.85	24.75
GEMINI2	No Defense	-	96.30	100.00	100.00	0.6065	16.17	41.66	45.40	21.03
	Gaussian Noise	0.001	96.30	100.00	100.00	0.6629	18.88	44.14	47.95	23.82
		0.01	95.80	99.80	100.00	0.7740	23.15	47.63	51.39	27.17
		0.1	27.90	50.80	63.90	0.5956	7.16	30.92	34.27	11.24
		1	0.20	1.50	2.70	0.1005	7.02	30.32	33.25	10.71
	LapMech	0.1	6.20	14.90	22.30	0.5325	7.13	30.31	33.46	10.72
		0.25	61.00	82.90	90.60	0.6198	8.19	32.77	36.26	12.76
		0.5	93.00	99.00	99.70	0.6928	13.12	37.74	41.26	17.56
		1	95.80	99.80	100.00	0.7507	19.52	44.30	48.02	24.16
		2	96.30	100.00	100.00	0.7720	24.18	48.94	52.55	28.82
		4	96.50	100.00	100.00	0.7591	23.82	48.36	52.33	28.32
		8	96.50	100.00	100.00	0.7196	21.30	46.70	50.41	26.22
	PurMech	0.1	4.50	10.70	14.80	0.0928	8.83	32.09	34.96	12.51
		0.25	55.60	81.20	88.70	0.6256	8.71	32.69	36.01	12.80
		0.5	90.20	99.10	99.90	0.6961	14.05	38.34	41.90	18.16
		1	95.00	100.00	100.00	0.7529	20.10	44.74	48.53	24.46
		2	95.80	100.00	100.00	0.7748	23.38	47.88	51.67	27.60
		4	96.10	100.00	100.00	0.7618	23.84	48.73	52.44	27.99
		8	96.20	100.00	100.00	0.7113	21.07	46.48	50.12	25.91
NOMIC	No Defense	-	64.60	89.60	94.50	0.6987	18.03	44.43	48.03	23.54
	Gaussian Noise	0.001	64.60	89.40	94.70	0.6943	18.52	44.33	47.63	23.50
		0.01	54.30	84.10	91.20	0.5896	13.88	39.63	43.18	19.30
		0.05	4.40	11.70	17.40	0.4411	6.67	31.35	34.88	11.17
		0.1	1.00	3.00	5.70	0.3800	5.71	29.33	32.43	9.74
		1	0.00	0.40	0.70	0.1157	5.42	28.80	31.59	8.65
	LapMech	0.1	0.10	0.70	1.80	0.2606	5.89	29.00	32.06	9.75
		0.25	0.40	2.30	3.00	0.3425	5.82	29.37	32.43	10.13
		0.5	1.10	6.60	9.60	0.4017	6.97	30.61	33.80	11.07
		1	11.70	26.70	34.70	0.4745	7.97	33.02	36.43	12.73
		2	37.70	65.70	77.20	0.5422	11.15	36.64	40.27	16.33
		4	56.90	83.10	90.70	0.5986	14.30	39.83	43.64	19.45
		8	62.10	87.60	94.10	0.6479	17.05	42.63	46.25	22.03
	PurMech	0.1	0.10	0.60	1.30	0.1319	5.92	29.02	31.80	9.15
		0.25	0.30	1.40	3.40	0.3503	5.97	29.42	32.57	9.90
		0.5	1.30	5.40	8.40	0.4037	6.21	30.11	33.41	10.41
		1	10.30	24.50	34.80	0.4716	8.31	32.76	36.38	12.95
		2	37.10	67.60	77.20	0.5399	10.62	36.42	40.14	16.21
		4	56.30	85.00	91.40	0.5976	14.50	39.95	44.00	19.81
		8	62.40	89.10	94.30	0.6481	17.29	42.36	46.43	22.06
NVIDIA	No Defense	-	84.00	96.70	98.80	0.6886	20.19	44.72	48.51	24.49
	Gaussian Noise	0.001	84.00	97.00	98.70	0.6926	20.91	45.72	49.53	25.30
		0.01	83.00	96.50	98.40	0.7198	20.36	45.29	49.08	25.13
		0.1	35.00	60.00	69.80	0.5493	8.64	32.41	35.78	12.83
		1	0.40	1.30	1.90	0.0485	5.99	29.05	32.13	9.99
	LapMech	0.1	3.90	12.00	18.10	0.4109	6.52	30.21	33.13	10.60
		0.25	39.40	66.90	75.80	0.5617	9.10	33.42	36.84	13.83
		0.5	72.70	92.50	96.30	0.6408	13.30	38.64	42.26	18.41
		1	80.50	95.60	98.90	0.6909	18.39	43.64	47.39	22.94
		2	83.40	96.50	99.00	0.7135	20.85	45.08	48.83	24.91
		4	83.70	97.00	99.00	0.7143	21.71	46.33	49.89	26.01
		8	83.80	96.90	98.80	0.7025	20.58	45.48	49.33	25.11
	PurMech	0.1	3.90	10.90	15.30	0.0080	6.44	29.39	32.31	10.07
		0.25	24.20	40.20	47.10	0.5634	9.96	33.79	36.97	14.01
		0.5	71.60	91.10	95.60	0.6416	14.71	38.82	42.31	18.96
		1	81.20	95.20	98.00	0.6942	18.27	43.02	46.61	22.31
		2	82.50	96.50	98.40	0.7183	19.78	44.78	48.50	24.59
		4	83.30	96.80	98.60	0.7188	21.43	46.27	50.14	26.30
		8	84.00	97.10	98.70	0.7033	21.21	45.66	49.47	25.94

Table 17: Dataset NoCaps, Utility and Defense, from the perspective of consumers and service providers, the retrieval performance the higher the better $\uparrow$, and the inversion performance the lower the better $\downarrow$. (N=1000)

Embedder	Defense	Parameter	R@1 $\uparrow$	R@5 $\uparrow$	R@10 $\uparrow$	BLEU $\downarrow$	Rouge-L $\downarrow$	Rouge-1 $\downarrow$	Rouge-2 $\downarrow$
CLIP	No Defense	-	90.60	99.80	99.80	16.84	43.13	47.14	21.60
	Gaussian Noise	0.001	90.60	99.80	99.80	17.00	43.27	47.04	21.80
		0.01	90.80	99.80	99.80	14.53	42.02	46.15	20.13
		0.1	20.00	43.20	54.00	6.97	32.61	35.57	10.90
		1	0.20	2.00	2.60	5.27	30.83	33.94	9.51
	LapMech	0.1	2.00	5.20	7.00	5.42	31.31	33.97	9.09
		0.25	11.00	22.60	31.80	6.51	32.20	35.11	10.16
		0.5	42.40	68.20	76.40	7.48	33.43	36.77	11.91
		1	79.20	95.60	98.40	10.51	37.13	40.62	15.16
		2	89.20	98.40	99.40	13.83	40.76	45.10	19.06
		4	91.00	99.60	99.80	15.44	42.47	46.29	20.74
		8	90.80	99.80	99.80	16.62	43.09	47.15	21.48
	PurMech	0.1	0.60	4.20	8.20	5.31	31.19	34.30	9.73
		0.25	7.80	21.20	29.60	6.54	32.40	35.32	10.55
		0.5	36.20	62.80	70.20	7.67	33.08	36.45	11.54
		1	75.80	96.00	98.80	10.60	37.22	40.68	15.78
		2	87.80	99.40	99.80	12.64	40.24	44.31	18.37
		4	90.60	99.80	99.80	14.95	42.15	46.25	20.42
		8	91.20	99.80	99.80	15.75	42.37	46.39	20.85
COHERE	No Defense	-	97.40	100.00	100.00	15.32	41.14	45.38	19.57
	Gaussian Noise	0.001	97.60	100.00	100.00	14.79	40.98	45.22	19.29
		0.01	97.40	100.00	100.00	13.99	40.20	44.29	19.05
		0.1	85.40	97.00	99.00	6.43	31.99	34.90	10.03
		1	1.00	3.00	6.40	4.79	30.77	33.71	8.83
	LapMech	0.1	13.60	35.80	45.60	5.55	31.23	33.89	9.43
		0.25	84.60	97.40	99.00	6.43	32.28	35.22	10.54
		0.5	94.80	99.80	100.00	8.01	33.72	36.62	12.25
		1	96.40	99.80	100.00	9.69	36.22	39.74	14.00
		2	97.00	100.00	100.00	13.04	39.75	43.28	17.79
		4	97.20	100.00	100.00	13.81	40.62	44.71	18.88
		8	97.20	100.00	100.00	14.04	40.93	45.18	19.10
	PurMech	0.1	12.80	29.20	41.80	5.51	31.29	34.13	9.43
		0.25	61.60	70.00	70.80	6.24	32.05	35.09	10.11
		0.5	94.20	99.80	100.00	7.69	33.30	36.47	11.64
		1	98.00	100.00	100.00	11.29	36.78	40.45	15.47
		2	97.60	100.00	100.00	13.40	39.74	43.66	18.21
		4	97.60	100.00	100.00	14.26	40.66	44.55	18.81
		8	97.60	100.00	100.00	14.13	40.97	45.38	19.13
GEMINI2	No Defense	-	98.20	100.00	100.00	14.96	41.86	45.23	20.06
	Gaussian Noise	0.001	98.20	100.00	100.00	14.96	41.66	45.01	19.75
		0.01	98.20	100.00	100.00	13.96	40.68	44.36	19.17
		0.1	32.40	59.00	69.00	6.77	32.10	34.94	10.90
		1	0.00	1.60	1.80	5.90	31.36	34.16	9.73
	LapMech	0.1	7.80	19.00	28.00	6.00	31.18	33.56	9.38
		0.25	62.40	83.60	90.00	6.62	32.07	34.86	10.36
		0.5	93.40	99.80	100.00	7.73	33.90	37.17	12.14
		1	97.40	100.00	100.00	11.23	37.30	41.29	15.99
		2	97.60	100.00	100.00	13.75	40.18	43.76	18.41
		4	98.00	100.00	100.00	14.28	41.07	44.83	19.39
		8	98.20	100.00	100.00	15.22	41.64	45.09	19.91
	PurMech	0.1	4.20	14.60	19.20	5.77	31.12	33.68	9.34
		0.25	62.60	85.40	91.00	6.72	32.43	35.02	10.89
		0.5	94.80	99.80	100.00	8.42	34.69	37.63	12.95
		1	98.00	100.00	100.00	11.27	37.81	41.37	16.22
		2	98.20	100.00	100.00	14.34	40.95	44.57	19.42
		4	98.40	100.00	100.00	15.15	41.49	45.07	20.19
		8	98.20	100.00	100.00	15.49	41.91	45.43	20.38
NOMIC	No Defense	-	84.20	98.20	99.20	15.88	42.16	45.69	20.30
	Gaussian Noise	0.001	84.40	98.20	99.00	15.92	42.19	45.79	20.54
		0.01	76.60	95.40	98.20	13.82	40.37	43.94	18.56
		0.1	2.40	6.40	9.80	5.68	32.11	35.00	10.24
		1	0.80	1.80	2.60	5.78	31.46	34.40	10.18
	LapMech	0.1	0.40	1.60	4.20	5.10	31.40	34.27	9.46
		0.25	0.60	4.80	7.40	5.80	31.62	34.36	9.56
		0.5	3.80	12.00	18.20	5.78	31.59	34.68	10.07
		1	22.20	43.60	56.20	7.78	33.78	36.98	12.19
		2	63.80	87.40	92.00	11.61	37.41	41.17	16.40
		4	79.20	96.00	97.80	13.51	39.98	43.99	18.32
		8	81.40	97.20	99.20	14.66	41.51	45.33	19.76
	PurMech	0.1	1.60	2.40	3.00	5.78	31.50	34.39	10.07
		0.25	2.20	4.00	6.00	5.61	31.46	34.58	9.90
		0.5	4.00	10.20	15.00	6.22	32.35	35.10	10.64
		1	19.40	40.20	50.60	8.44	33.81	36.98	12.26
		2	57.80	84.00	91.60	11.55	37.59	41.31	15.94
		4	77.80	95.80	98.20	13.71	40.21	43.90	18.29
		8	82.00	97.60	98.60	14.79	41.30	45.21	19.22
NVIDIA	No Defense	-	91.60	99.20	100.00	17.13	42.92	46.68	21.73
	Gaussian Noise	0.001	91.40	99.20	100.00	17.27	42.98	46.79	21.82
		0.01	91.40	99.20	100.00	16.80	42.32	45.95	20.86
		0.05	83.20	98.60	99.80	9.10	34.05	37.00	12.48
		0.1	53.20	79.40	87.20	6.47	31.92	34.81	10.21
		1	0.40	1.80	4.00	5.77	31.27	33.92	9.35
	LapMech	0.1	6.20	25.20	34.20	5.55	31.10	33.93	9.41
		0.25	59.20	88.00	94.20	6.52	32.13	35.18	10.50
		0.5	86.40	97.40	99.20	8.18	33.78	36.95	12.15
		1	90.80	98.80	99.80	12.22	37.75	41.48	16.50
		2	91.20	99.20	100.00	15.49	41.56	45.19	20.14
		4	91.60	99.00	100.00	16.47	42.61	46.54	21.58
		8	91.80	99.00	100.00	17.08	43.10	46.71	21.83
	PurMech	0.1	8.60	19.60	29.20	5.59	31.34	33.89	9.34
		0.25	36.80	55.40	60.40	6.91	32.13	35.13	10.54
		0.5	84.60	99.00	99.80	9.35	34.47	37.38	12.92
		1	90.60	99.20	99.80	13.56	38.75	42.06	17.10
		2	91.60	99.20	100.00	16.59	42.22	45.77	20.85
		4	92.20	99.20	100.00	17.33	42.99	46.70	21.81
		8	91.60	99.20	100.00	17.11	43.04	46.82	21.63

Table 18: Effect of CoT refinement on COCO val (COHERE victim, $N=10^3$). For each row we show the original image x , the low-fidelity diffusion reconstruction $\hat{x}^{\text{lo-fi}} = g_{\text{img}}(e_{V \rightarrow A})$ that GPT-4o reasons over, the reference caption, and the captions produced by CLIP-Rerank and by the CoT reasoning. Rows 1-4: CoT lifts recovery quality by grounding the candidate in the diffusion artifact (action verbs, spatial relations). Row 5: CoT regresses by hallucinating from a low-fidelity artifact.

ID	Original x	Lo-fi $\hat{x}^{\text{lo-fi}}$	Reference	CLIP-Rerank	CoT (GPT-4o)
396754			A group of people horse back riding on a sandy beach.	A couple of men riding horses down a beach.	A group of people riding horses on a beach.
127669			A man on a surfboard jumping over a wave.	A person on a surfboard in the water.	A person on a surfboard riding a wave.
125472			A man flying through the air while riding a skateboard.	A young man riding a skateboard on top of a wooden ramp.	A man flying through the air on a skateboard.
343642			A large metal clock in the middle of a room.	A clock sitting in the middle of a room.	A clock mounted to the side of a building. (<i>regression</i>)