DE-CLIP: Few-Shot Anomaly Detection via Difference-Guided Embedding Editing

Yage Zhang, Yukun Jiang, Michael Backes, Yang Zhang*
CISPA Helmholtz Center for Information Security
{yage.zhang, yukun.jiang, director, zhang}@cispa.de

Abstract

Anomaly detection (AD) plays a critical role in applications such as automated industrial inspection and medical image analysis. Empowered by the strong pre-trained vision-language model, CLIP, recent years have witnessed the emergence of several CLIP-based few-shot AD methods. Due to the overlap between the embedding distributions of normal and anomalous samples, many existing approaches introduce additional model training for more discriminative text embeddings. However, we demonstrate that such training is not necessary. Specifically, we find that this embedding overlap can be separated by introducing a Difference-guided vector for embedding Editing (DiffEdit). Based on this finding, we propose DE-CLIP, a simple yet effective framework based on DiffEdit, which directly edits text embeddings based on the textual and visual differences between normal and anomalous samples, resulting in more discriminative embeddings for AD. Extensive experiments on industrial and medical datasets demonstrate the superiority of our proposed DE-CLIP compared with existing baselines. For instance, on the MVTec dataset, DE-CLIP achieves 96.6% and 96.7% AUROC on anomaly classification and segmentation, surpassing both training-based and training-free methods. In addition, we observe that introducing DiffEdit into other training-free baselines could also significantly improve their performance, highlighting the potential of DiffEdit to promote better AD.[1]

1 Introduction

Anomaly detection (AD) is a critical task in safety-sensitive domains such as industrial inspection and medical diagnostics, where rare but significant deviations from normality should be identified accurately. AD typically comprises two sub-tasks: anomaly classification (AC), which distinguishes abnormal from normal instances, and anomaly segmentation (AS), which localizes the abnormal regions at the pixel level (Bergmann et al., 2019; Fernando et al., 2021). However, these tasks remain highly challenging in practice due to the scarcity of anomalous samples, leading to long-tailed data distributions and insufficient training signals (Liang et al., 2022; Tyshchuk et al., 2023; Schrodi et al., 2024). This imbalance creates a barrier to both learning and generalization, especially under zero- or few-shot settings (Song et al., 2022; Li et al., 2024b; Zhang et al., 2024).

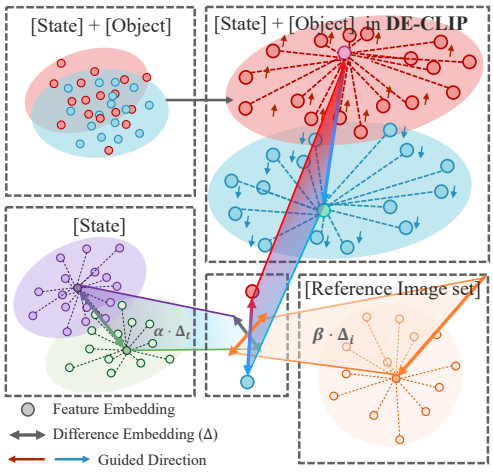


Figure 1: Overview of DE-CLIP's DiffEdit guided embedding mechanism. Feature embeddings are steered using difference vectors Δ derived from textual and visual domains.

Recent advances in vision-language models (VLMs) (Leng et al., 2026; Jiang et al., 2024; Zhu et al., 2023; Li et al., 2023), especially the contrastive language–image pretraining (CLIP) model (Radford et al., 2021), offer a promising solution to these challenges. CLIP (Radford et al., 2021), trained on approximately 400 million noisy image-text pairs crawled from the internet, exhibits remarkable zero-shot generalization capabilities,

*Corresponding author
[1]Our Code is available at https://github.com/TrustAIRLab/DE-CLIP.

particularly excelling in image classification and recognition tasks (Qu et al., 2024; Li et al., 2021, 2022, 2023; Radford et al., 2021). However, its performance on anomaly understanding remains limited. A key bottleneck lies in the semantic embedding space of CLIP's text encoder: prompts with semantically opposite meanings (e.g., *intact bottle* vs. *broken bottle*) are often mapped into nearby regions (Ma et al., 2025; Eslami and de Melo, 2024; Kang et al., 2025; Tong et al., 2024; Tang et al., 2023; Suo et al., 2023), resulting in decision ambiguity and blurred boundaries between normal and anomalous concepts.

WinCLIP (Jeong et al., 2023) is an early CLIP-based AD method that constructs extensive handcrafted prompts and performs multiple forward passes for zero-/few-shot classification and segmentation. VAND (Chen et al., 2023) extends this with projection learning from reference images. More deeply integrated approaches, such as AdaCLIP (Cao et al., 2024), AnomalyCLIP (Zhou et al., 2023), and AA-CLIP (Ma et al., 2025), rely on fine-tuning, specialized prompts, or added modules to improve performance. However, these approaches typically rely on large amounts of task-specific training data before being transferred to new datasets. This not only constrains their generalizability and applicability in zero-shot scenarios but also risks disrupting the internal representational space of CLIP.

In this work, we observe that structured antonymic adjective pairs naturally encode interpretable semantic directions within CLIP's embedding space. These directions define a cognitive opposition axis, which we find can be leveraged to enhance conceptual separability. Inspired by psycholinguistic theories of semantic differentials (Lakoff and Johnson, 1980; Langacker, 1987; Barsalou, 1999), we propose to inject directional difference vectors into CLIP's latent space. Surprisingly, by aligning these textual shifts with representative visual anchors, the resulting edited embeddings become more grounded and disentangled, leading to clearer semantic boundaries for AD tasks.

Building on this insight, we propose DE-CLIP: a plug-and-play, training-free framework for few-shot AC and AS. DE-CLIP operates by computing a semantic difference vector from antonymic state word pairs, and injecting it into state-object prompts. Notably, while DE-CLIP requires no parameter updates or fine-tuning, it does leverage a small number of normal samples and a predefined antonym lexicon to guide inference. Thus, "training-free" in our context refers strictly to the absence of gradient-based optimization. This process creates a perceptual-semantic axis that systematically transforms the embedding space. To further ground these edits, we align the semantic axis with visual reference embeddings from normal samples. As shown in Figure 1, this dual-modality difference guided editing introduces structured tension between normal and anomalous prompts. We call this adjustment method Difference-guided embedding Editing (DiffEdit). On industrial and medical datasets, DE-CLIP achieves the AUROC of 96.6%/96.7%, 87.1%/97.6%, and 75.2%/92.6% for AC/AS on MVTec, VisA, and Brain MRI datasets, respectively, outperforming existing baselines, including both training-based and training-free methods, and has great potential to enhance the performance of other training-free ones.

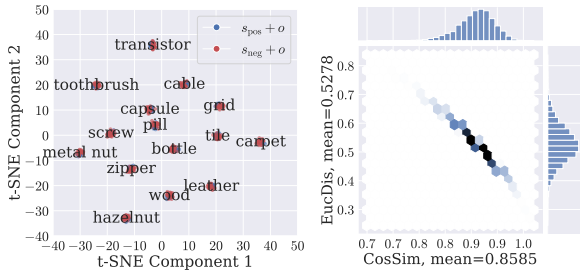
Overall, our contributions are four-fold.

- We identify semantic overlap in CLIP's text embedding space as a key bottleneck for anomaly detection and formulate a solution by introducing a Difference-guided vector for embedding Editing (DiffEdit).
- We propose DE-CLIP, a dual-guided editing framework that introduces semantic difference vectors from both textual and visual modalities to separate semantic overlap in the embedding space without training.
- We conduct extensive evaluations on multiple industrial and medical AD datasets and show that DE-CLIP consistently outperforms existing approaches in both AC/AS.
- DiffEdit can also improve their performance when combined with other baselines, demonstrating the potential of Difference-Guided Embedding Editing to promote better AD. DiffEdit yields up to a 30.1% increase in AUROC.

2 Problem Formulation

2.1 Zero/Few-Shot Anomaly Detection

Problem Set-Up. Zero-shot AD separates normal from anomalous samples without any defect labels, while few-shot AD augments each object class with K reference normals to improve robustness in scarce-data regimes.



(a) (b)

Figure 2: (a) t-SNE visualization of composed embeddings $E(S+O)$. (b) Cosine similarity and Euclidean distance between normal and anomalous $E(S+O)$ embeddings.

Given a test image $x \in \mathcal{X}$ and prompts $\{t_i\}_{i=1}^N$, CLIP embeddings are

$$E(I) := f_{\text{img}}(x), \qquad E(T) := f_{\text{text}}(t_i).$$

Zero-shot uses prompts alone, few-shot adapts $E(T)$ with features from the reference set $\mathcal{R}_o = \{x_1^{(o)}, \ldots, x_K^{(o)}\}$.

Binary Test in Joint Space. Define the hypotheses

$$\mathcal{H}_0 : x \sim \mathcal{D}_{\text{pos}} \quad \text{(normal)},$$
$$\mathcal{H}_1 : x \sim \mathcal{D}_{\text{neg}} \quad \text{(anomalous)},$$

and score an image via maximum cosine similarity:

$$\text{score}(x) = \max_i \cos\big(E(x), E(t_i)\big).$$

Geometry Requirement. Let $E_{\text{pos}} := E(t_{\text{pos}})$, $E_{\text{neg}} := E(t_{\text{neg}})$ be canonical normal/defect embeddings. Reliable AD requires

$$\cos(E_{\text{pos}}, E_{\text{neg}}) \ll 1 \text{ and}$$
$$\|E_{\text{pos}} - E(x_{\text{pos}})\| \ll \|E_{\text{neg}} - E(x_{\text{pos}})\|,$$

to ensure a wide inter-class angle and tight normal alignment.

2.2 Imperfect Text Embedding

We first analyze the structure of the embedding space $E(S+O)$ to evaluate how well CLIP separates normal and abnormal prompts. Figure 2(a) shows the t-SNE projection of $E(S+O)$ embeddings of 15 object classes on MVTec Dataset, where each dot represents either a positive or negative phrase. We observe that clusters of normal and anomalous state words with object are largely overlapping across all categories.

To quantify this semantic entanglement, we compute all pairwise cosine similarities and Euclidean distances between each normal and anomalous $E(S+O)$ vector. The distribution of cosine similarities and Euclidean distances in Figure 2b further confirms that the positive/negative phrases are densely mixed in the embedding space, showing no significant margin.

We note that many existing prompt-based CLIP adaptations adopt handcrafted templates such as "a photo of a damaged screw." Although such phrases may not exist verbatim in CLIP's pretraining corpus, we conservatively categorize them as belonging to the same class of $E(S+O)$ representations for experimental consistency.

3 Our Proposed DE-CLIP

3.1 Overview

We propose DE-CLIP, a few-shot AD/AS framework that edits prompts with modality-specific shifts: linguistic Δ_T and visual Δ_I (Figure 3). In classification, weights α, β blend the two directions, and the cosine to the query embedding yields the anomaly score. For segmentation, multi-scale query–reference comparisons generate similarity maps that are fused into the final mask, capturing both global and local defects.

3.2 Difference Based on Language Modality

Guidance Potential of $\Delta_{T(S+O)}$. We test whether text-space directions that join state and object tokens can stand in for explicit visual guidance. For object o with positive and negative state vocabularies $\mathcal{S}_{\text{pos}} = \{s_{\text{pos},m}\}_{m=1}^M$ and $\mathcal{S}_{\text{neg}} = \{s_{\text{neg},n}\}_{n=1}^N$, the composed semantic offset is

$$\Delta_{T(S+O)} = \frac{1}{M}\sum_m E(s_{\text{neg},m} + o) - \frac{1}{N}\sum_n E(s_{\text{pos},n} + o).$$

All difference vectors are ℓ_2-normalised to remove scale effects noted in Figure 2.

Figure 4 plots cosine similarities between $\Delta_{T(S+O)}$ and Δ_I. Strong diagonal values show that the composed text offsets reliably align with object-specific visual shifts, while several off-diagonal peaks indicate transferable semantic directions that generalise across categories.

Cross-Category Direction Transfer. To verify that the off-diagonal similarities in Figure 4 correspond to genuinely transferable semantic directions, we conduct a direction-replacement experiment. For each target category, we replace its own

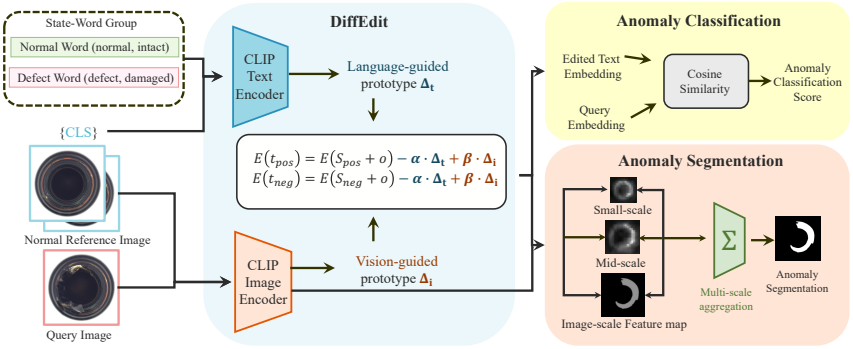


Figure 3: Workflow of our proposed DE-CLIP for few-shot anomaly detection and segmentation.

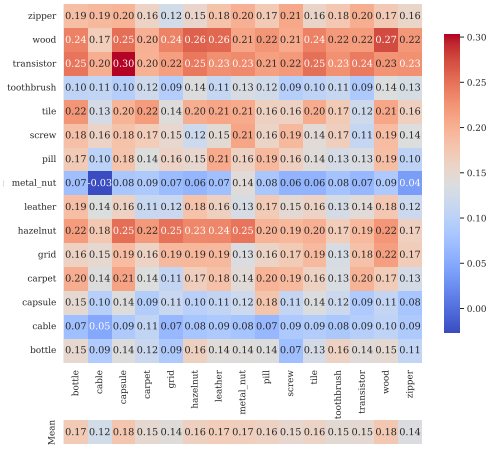
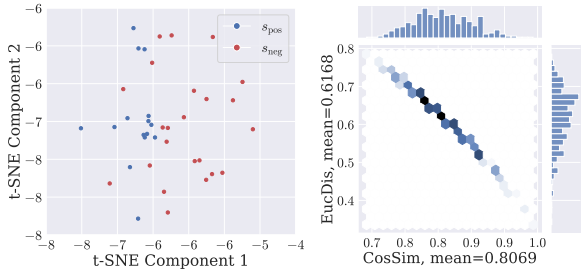


Figure 4: Cosine similarity between Δ_I and $\Delta_{T(S+O)}$ across all object categories. Diagonal values indicate intra-class alignment, off-diagonal values reflect cross-category generalization.



(a) (b)

Figure 5: (a) t-SNE visualization of state-only embeddings $E(S)$. (b) Cosine similarity and Euclidean distance between normal and anomalous $E(S)$ embeddings.

$\Delta_{T(S+O)}$ with the direction from a high-similarity category and a low-similarity category (as read from Figure 4), and measure the resulting AC-AUROC on MVTec (8-shot). Table 1 shows representative results. Replacing with a high-similarity direction preserves or even slightly improves performance, whereas using a low-similarity direction

Target	Source		AC-AUROC		
	High	Low	Self	High	Low
hazelnut	leather	cable	91.4	92.8 (+1.4)	64.9 (−26.5)
wood	leather	cable	99.2	99.2 (+0.0)	69.0 (−30.2)
carpet	grid	toothbrush	99.8	98.9 (−0.9)	54.0 (−45.8)

Table 1: Cross-category direction transfer on MVTec.

causes a substantial drop, confirming that the off-diagonal structure in Figure 4 reflects meaningful semantic transferability.

Lightweight Directional Editing via $\Delta_{T(S)}$. Motivated by cross-category generalizability, we explore a simplified, object-agnostic direction $\Delta_{T(S)}$, computed purely from state adjectives:

$$\Delta_{T(S)} = \frac{1}{M}\sum_{m=1}^{M} E(s_{\text{neg},m}) - \frac{1}{N}\sum_{n=1}^{N} E(s_{\text{pos},n}).$$

As shown in Figure 5, this direction already separates normal and abnormal concepts well, even without object context. Averaged across all categories, its cosine similarity with visual embedding differences reaches 0.14, modest but consistently positive, confirming $\Delta_{T(S)}$'s alignment with visual anomalies. Given its generality and efficiency, we adopt $\Delta_{T(S)}$ as a default editing vector, reducing both computational and annotation overhead.

3.3 Difference Based on Vision Modality

Because language differences capture high-level semantics that subsequently guide low-level visual refinements, we next turn from textual to visual cues to show how the two forms of guidance complement each other.

Similarly, let $\mathcal{I}_{\text{pos}} = \{x_{\text{pos},p}\}_{p=1}^{P}$ and $\mathcal{I}_{\text{neg}} = \{x_{\text{neg},q}\}_{q=1}^{Q}$ represent the collections of positive (normal) and negative (anomalous) image samples.

The average visual embedding difference is

$$\Delta_I = \frac{1}{Q}\sum_{q=1}^{Q} E(x_{\text{neg},q}) - \frac{1}{P}\sum_{p=1}^{P} E(x_{\text{pos},p}). \quad (1)$$

In principle, incorporating visual embedding differences between anomalous and normal samples Δ_I can provide strong semantic alignment between vision and language modalities. Such guidance helps steer the prompt embedding toward regions more consistent with the actual image distribution. However, in practical scenarios, abnormal samples are scarce or unavailable during deployment, limiting access to $\mathcal{I}_{\text{neg}}$.

To address this constraint, we approximate the visual difference using only the normal samples $\mathcal{I}_{\text{pos}}$. Specifically, we incorporate vision-based guidance through Δ_I, which captures the mean CLIP embedding of a small reference set of normal images:

$$\Delta_I = \frac{1}{P}\sum_{p=1}^{P} E(x_{\text{pos},p}),$$

This vector represents the prototypical visual concept of "normality" and acts as a perceptual grounding signal for guiding semantic contrast.

3.4 Theoretical Insights

Let the class-conditional image means be

$$\mu_+ = \tfrac{1}{P}\sum_{p=1}^{P} E(x_{\text{pos},p}), \qquad \mu_- = \tfrac{1}{Q}\sum_{q=1}^{Q} E(x_{\text{neg},q}),$$

so the visual shift defined in (1) can be written compactly as follows. Let $\Delta_T = \Delta_{T(S)}$, and let μ_+, μ_- be the class-mean image embeddings. Define

$$\boxed{\Delta_I^{\pm} = \mu_+ - \mu_-, \qquad \Delta_I^{+} = \mu_+.}$$

Ideal (Two-Sided) Case. CLIP alignment gives $\Delta_I^{\pm} \approx \delta\Delta_T \quad (\delta > 0)$; under the isotropic assumption, $w^\star \propto \Delta_I^{\pm}$ (Fisher LD optimum (Fisher, 1936)). DiffEdit shifts the positive anchor $\tilde{t}_{\text{pos}} = E(t_{\text{pos}}) + \beta\Delta_I^{\pm}$, hence

$$\tilde{t}_{\text{pos}} - E(t_{\text{neg}}) = \Delta_T + \beta\Delta_I^{\pm} \approx (1+\beta\delta)\Delta_T,$$

i.e. any β rescales but never rotates $w^\star$.

Practical (One-Sided) Case. When only normal images are available, use $\Delta_I^{+} = \mu_+$. Because μ_+ aligns with $E(t_{\text{pos}})$, $\Delta_I^{+} \approx \gamma\Delta_T \ (\gamma > 0)$, so the same derivation yields

$$\hat{t}_{\text{pos}} - E(t_{\text{neg}}) \approx (1+\beta\gamma)\Delta_T,$$

again preserving the FLD direction. Thus, DiffEdit maintains optimality even with a small normal reference set.

3.5 DiffEdit: Difference-Guided Vector for Embedding Editing

To mitigate semantic overlap between normal and anomalous prompts in CLIP's latent space, we introduce a directional editing strategy that enhances class separation.

Inspired by Semantic Differential Scales (SDS) (Osgood et al., 1957), which encode meaning along bipolar adjective axes (e.g., good–bad), we construct a task-specific Defect-Oriented Semantic Differential Lexicon comprising antonymic adjective pairs (e.g., intact–broken, stable–cracked). Unlike classical SDS, our lexicon is grounded in domain-relevant visual semantics, ensuring that semantic differences align with perceptually meaningful axes for anomaly discrimination.

Default Difference Vectors. In this work, we use Δ_T and Δ_I^{+} by default to refer to the language- and vision-guided difference vectors, respectively. As discussed previously, Δ_T specifically denotes $\Delta_{T(S)}$. The textual difference vector Δ_T is computed from state word embeddings, while the visual difference vector Δ_I^{+} is derived from a small set of normal reference images.

Dual-Modality Difference-Guided Embedding Construction. Motivated by cognitive semantics (Lakoff and Johnson, 1980; Langacker, 1987; Barsalou, 1999), we apply a structured editing mechanism that integrates both language and vision differences, which we term the difference-guided vector for embedding Editing (DiffEdit). Specifically,

$$\text{DiffEdit} = \begin{cases} -\alpha\cdot\Delta_T + \beta\cdot\Delta_I^{+} & \text{if } t = t_{\text{pos}} \\ +\alpha\cdot\Delta_T - \beta\cdot\Delta_I^{+} & \text{if } t = t_{\text{neg}} \end{cases}$$

where $s_{\text{pos},m}$ and $s_{\text{neg},n}$ are adjective tokens describing normal and anomalous states, o denotes the object class, and α, β control the strength of semantic and visual guidance. This push-pull formulation ensures that normal prompts are drawn toward visual and linguistic prototypes, while anomalous prompts are repelled along both axes, resulting in an amplified semantic margin.

For clarity, we denote the editing vector as $\text{DiffEdit}(t)$, applied conditionally based on polarity.

3.6 AC in DE-CLIP

Prototype Construction. For object class o with normal adjectives $\{s_{\text{pos},m}\}_{m=1}^{M}$ and anomalous adjectives $\{s_{\text{neg},n}\}_{n=1}^{N}$, DiffEdit yields

$$E(t_{\text{pos}}) = \frac{1}{M}\sum_{m=1}^{M} E(s_{\text{pos},m} + o) + \text{DiffEdit}(t_{\text{pos}}),$$

$$E(t_{\text{neg}}) = \frac{1}{N}\sum_{n=1}^{N} E(s_{\text{neg},n} + o) + \text{DiffEdit}(t_{\text{neg}}),$$

two semantically disentangled prototypes anchored by both linguistic Δ_T and visual Δ_I guidance.

Directional Anomaly Score. Given query image x with embedding $E(x)$,

$$\begin{aligned}\text{score}(x) = &\cos\big(E(x), E(t_{\text{neg}})\big) \\ &- \cos\big(E(x), E(t_{\text{pos}})\big),\end{aligned}$$

where larger values imply stronger anomalous affinity. This soft differential metric preserves fine-grained angular cues that hard thresholds miss. The score is later fused with segmentation evidence for added robustness.

3.7 AS in DE-CLIP

Following few-shot AD practice (Jeong et al., 2023; Li et al., 2024b), the AS module aggregates multi-scale CLIP features. For scales $l \in \{\text{low}, \text{mid}, \text{high}\}$, extract patches $\{\mathbf{v}_i^{(l)}(x)\}_{i=1}^{N_l}$ and build a normal memory bank $\mathcal{R}^{(l)} = \{\mathbf{r}_j^{(l)}\}_{j=1}^{K_l}$. The per-patch anomaly score is the minimum cosine distance

$$A_i^{(l)} = \min_j \big(1 - \cos(\mathbf{v}_i^{(l)}(x), \mathbf{r}_j^{(l)})\big),$$

and the cross-scale fusion uses a harmonic mean

$$A_i = \Big(\sum_l \frac{1}{A_i^{(l)} + \epsilon}\Big)^{-1},$$

which is then upsampled to an image-level anomaly map. This multi-resolution scheme captures both local defects and global structural deviations, entirely without pixel-level labels.

4 Experiments

4.1 Experiment Setups

Datasets. Following prior AD literature (Huang et al., 2024; Cao et al., 2024; Chen et al., 2023; Jeong et al., 2023; Zhou et al., 2023), we benchmark on two industrial datasets, MVTec AD (Bergmann et al., 2019) and VisA (Zou et al., 2022), and one medical dataset, BMAD Brain MRI (Bao et al., 2023). Together they span natural photographs and MR images and provide both image- and pixel-level ground truth, allowing unified evaluation of detection and localisation.

Model and Training Protocol. DiffEdit is conceived as a training-free, plug-and-play module that drops into existing AD frameworks without any fine-tuning. For our main instantiation, DE-CLIP follows (Cao et al., 2024; Chen et al., 2023; Qu et al., 2024; Zhou et al., 2023) and uses OpenCLIP with a ViT-L/14 backbone as the visual encoder.

Metrics and Hyperparameters. Consistent with (Deng and Li, 2022; You et al., 2022; Defard et al., 2021; Roth et al., 2022; Huang et al., 2024; Cao et al., 2024; Chen et al., 2023; Jeong et al., 2023), we report Area Under the ROC Curve (AUROC), average precision (AP) and the maximum F1 score (max-F1) under the optimal threshold to evaluate both image-level and pixel-level AD performance. Unless stated otherwise, guidance weights are fixed to $\alpha = \beta = 0.5$ for every dataset, and Section 4.4 analyses the full sweep.

Implementation Details. Consistent with (Jeong et al., 2023; Chen et al., 2023; Li et al., 2024a,b; Zhang et al., 2024), we supply a small pool of normal reference images at test time to instantiate DiffEdit. Every experiment is run under three random seeds (10, 42, 59) and averaged. All computations are performed on a single NVIDIA DGX A100 GPU.

4.2 Comparison with Other Methods

We benchmark DE-CLIP against a diverse suite of CLIP-based anomaly detectors grouped by training depth.

Training-free baselines include (i) CLIP (Radford et al., 2021), which compares "A photo of a normal/anomalous [object]" prompts to query images; (ii) CLIP-AC (Radford et al., 2021), which averages normal and anomalous ImageNet-style templates; and (iii) WinCLIP(+) (Jeong et al., 2023), which uses reference images at inference but keeps both CLIP encoders frozen. *Shallow-training* is represented by VAND (Chen et al., 2023), which adds and trains a lightweight adapter at the head, leaving the backbone fixed. *In-depth training* includes AA-CLIP (Ma et al., 2025), which fits multiple projectors to align orthogonalised modalities;

Method	Training Required?	AC(AUROC/AP/max-F1)					AS(AUROC/AP/max-F1)				
		0-Shot	2-Shot	4-Shot	8-Shot	10-Shot	0-Shot	2-Shot	4-Shot	8-Shot	10-Shot
CLIP	No	76.8/75.4/76.7	–	–	–	–	47.5/12.6/8.7	–	–	–	–
CLIP-AC		61.8/60.5/62.2	–	–	–	–	51.3/14.2/9.5	–	–	–	–
WinCLIP		90.4/92.7/95.6	93.7/94.5/96.9	95.3/94.9/97.6	95.6/95.4/97.6	95.8/95.4/97.9	82.3/24.8/18.2	93.8/43.8/39.5	94.2/49.5/42.2	94.5/45.9/42.4	94.5/46.3/42.6
DE-CLIP (Ours)		85.9/90.2/93.4	91.2/93.4/94.5	94.0/93.7/95.5	96.6/94.7/96.6	95.1/94.8/96.2	94.2/50.6/48.1	95.8/51.9/50.0	96.2/54.8/53.2	96.6/55.7/54.2	96.7/55.2/53.9
VAND	Shallow Training	84.2/88.6/89.8	90.9/92.4/95.7	91.8/92.7/96	92.2/92.8/96.2	92.1/92.7/96.1	87.2/25.4/19.1	95.2/54.1/51.7	95.8/55.5/53.2	96/57.4/54.9	96.1/57.7/55.1
AA-CLIP	In-Depth Training	84.6/89.1/92.9 (Full-Shot)					91.4/46.1/44.8 (Full-Shot)				
AdaCLIP		84.8/90.3/92.0 (Full-Shot)					89.2/43.9/41.2 (Full-Shot)				
AnomalyCLIP		91.6/92.5/95.9 (Full-Shot)					91.0/45.9/44.2 (Full-Shot)				

Table 2: Average performance comparison of various methods and DE-CLIP on the **MVTec** dataset under different shot settings. The best, second-best, and third-best results on AUROC are highlighted as first, second, and third.

Method	Training Required?	AC(AUROC/AP/max-F1)					AS(AUROC/AP/max-F1)				
		0-Shot	2-Shot	4-Shot	8-Shot	10-Shot	0-Shot	2-Shot	4-Shot	8-Shot	10-Shot
CLIP	No	66.4/71.4/76.6	–	–	–	–	46.6/3.4/2.3	–	–	–	–
CLIP-AC		65.0/70.5/76.7	–	–	–	–	47.8/3.4/2.4	–	–	–	–
WinCLIP		75.5/78.2/78.7	83.4/82.5/84.6	84.1/82.4/85.8	86.0/83.8/87.1	86.4/83.9/87.8	73.2/9.0/5.4	95.1/23.9/16.6	95.2/25.3/17.7	95.3/25.8/18.2	95.2/25.8/17.8
DE-CLIP (Ours)		79.1/78.4/80.2	82.5/82.0/84.1	83.1/81.8/85.6	86.5/83.4/87.8	87.1/84.6/88.4	90.2/30.5/23.3	97.0/32.3/24.3	97.2/34.1/25.7	97.6/35.4/27.1	97.6/35.5/26.9
VAND	Shallow Training	78.0/76.9/80.3	80.9/79.6/83.5	85.9/82.5/89.3	86.5/83.7/89.8	87.0/83.5/90.2	87.2/18.5/14.3	95.3/34.4/27.3	95.4/34.4/27.9	95.7/34.9/28.3	95.7/35.2/28.5
AA-CLIP	In-Depth Training	82.7/81.5/83.4 (Full-Shot)					90.9/32.2/31.8 (Full-Shot)				
AdaCLIP		82.5/81.5/85.2 (Full-Shot)					95.6/36.4/32.0 (Full-Shot)				
AnomalyCLIP		82.1/81.8/83.5 (Full-Shot)					91.1/34.6/27.1 (Full-Shot)				

Table 3: Average performance comparison of various methods and DE-CLIP on the **VisA** dataset under different shot settings. The best, second-best, and third-best results on AUROC are highlighted as first, second, and third.

Method	Training Required?	AC(AUROC/AP/max-F1)					AS(AUROC/AP/max-F1)				
		0-Shot	2-Shot	4-Shot	8-Shot	10-Shot	0-Shot	2-Shot	4-Shot	8-Shot	10-Shot
CLIP	No	38.8/42.2/40.5	–	–	–	–	68.3/3.2/1.7	–	–	–	–
CLIP-AC		37.8/42.2/38.2	–	–	–	–	67.5/3.2/1.6	–	–	–	–
WinCLIP		60.7/90.6/87.8	62.2/90.8/89.7	65.7/90.8/89.4	65.1/90.7/89.1	64.3/90.7/88.8	87.7/9.1/4.8	92.1/15.1/8.8	91.4/13.7/7.8	91.4/13.6/7.7	91.3/13.5/7.5
DE-CLIP (Ours)		54.3/88.6/86.5	75.2/91.2/91.9	73.6/91.0/91.0	72.8/91/90.5	71.7/90.9/90.2	87.7/9.2/5.0	92.6/16.4/10	91.9/14.6/8.6	92.0/14.6/8.4	91.9/14.4/8.3
VAND	Shallow Training	49.9/80.6/72.7	65.7/86.3/86.7	65.1/86.1/87.1	67.4/86.1/87.3	67.8/86.1/87.5	84.7/14.2/8.4	85.9/11.0/13.7	85.9/10.9/13.8	86.0/11.4/14.0	85.5/11.5/14.2
AA-CLIP	In-Depth Training	44.9/52.8/40.9 (Full-Shot)					89.9/13.5/11.4 (Full-Shot)				
AdaCLIP		64.4/91.0/90.3 (Full-Shot)					91.3/14.3/10.6 (Full-Shot)				
AnomalyCLIP		68.9/91.1/90.1 (Full-Shot)					88.1/10.9/11.1 (Full-Shot)				

Table 4: Average performance comparison of various methods and DE-CLIP on the **Brain MRI** dataset under different shot settings. The best, second-best, and third-best results on AUROC are highlighted as first, second, and third.

Method w/ DiffEdit	Training Required?	AC					AS				
		0-Shot	2-Shot	4-Shot	8-Shot	10-Shot	0-Shot	2-Shot	4-Shot	8-Shot	10-Shot
CLIP	No	+9.2	+13.0	+13.6	+13.9	+14.1	+15.4	+22.5	+13.2	+9.0	+25.3
CLIP-AC		+24.6	+28.3	+28.8	+29.9	+30.1	+4.4	+9.3	+4.1	+22.4	+7.2
WinCLIP		+0.4	+0.1	+0.7	+2.1	+0.7	+0.0	+2.2	+2.0	+2.4	+2.2
VAND	Shallow Training	+0.2	+0.4	+0.2	+0.6	+0.2	+0.9	+0.0	+0.0	+0.9	+0.2
AA-CLIP	In-Depth Training	+0.4	+0.3	+0.4	+0.4	+0.4	+0.7	+0.6	+0.6	+0.6	+0.6
AdaCLIP		+0.2	+0.4	+0.2	+0.2	+0.4	+0.4	+0.4	+0.6	+0.2	+0.2
AnomalyCLIP		+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	+0.1	+0.0	+0.1	+0.1

Table 5: Average improvement in AUROC (%) on the **MVTec** dataset achieved by combining DiffEdit with each baseline method.

Method w/ DiffEdit	AC					AS				
	0-Shot	2-Shot	4-Shot	8-Shot	10-Shot	0-Shot	2-Shot	4-Shot	8-Shot	10-Shot
WinCLIP	+3.5	+7.9	+1.7	+2.1	+1.7	+2.3	+2.2	+2.0	+2.4	+2.2
VAND	+9.0	+15.1	+14.2	+14.4	+14.9	+4.3	+9.3	+9.6	+9.5	+10.0

Table 6: Average improvement in AUROC (%) on the **Brain MRI** dataset achieved by combining DiffEdit with each baseline method.

AdaCLIP (Cao et al., 2024), which learns prompts end-to-end; and AnomalyCLIP (Zhou et al., 2023), which jointly fine-tunes both vision and text encoders.

Table 2, Table 3, and Table 4 show that DE-CLIP delivers state-of-the-art AUROC and other metrics for both AC and AS on MVTec, VisA, and Brain MRI across different shot counts. Despite being entirely training-free, it equals or surpasses shallowly and deep baselines, e.g., outperforms WinCLIP and rivals VAND/AA-CLIP on MVTec, while ranking top-tier in AC and first–second in AS on Brain MRI.

4.3 Visualization

Figure 6 highlights consistent localisation from DE-CLIP: defects are sharply isolated on both

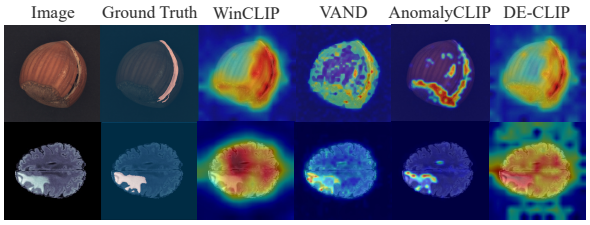


Figure 6: Visualization of anomaly localization results of different methods. DE-CLIP, WinCLIP, and VAND are evaluated under the 8-shot setting, while AnomalyCLIP uses the full-shot setting.

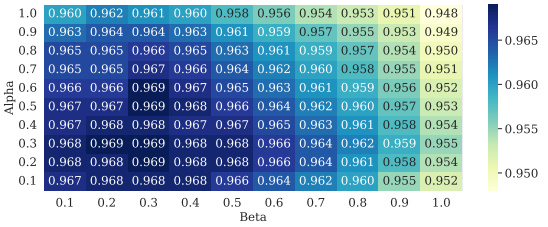


Figure 7: AUROC heatmap on MVTec across α and β.

MVTec-Hazelnut and Brain MRI. VAND misses anomalies entirely, WinCLIP fires in irrelevant areas, and AnomalyCLIP oscillates, over-detecting hazelnut regions but missing key brain lesions. DiffEdit's heat-maps, however, align closely with ground truth, evidencing robust suppression of spurious responses and faithful enhancement of true anomaly cues across domains.

4.4 Ablation Study

Impact of Hyperparameters. We grid-searched $\alpha, \beta \in \{0.1{:}0.1{:}1.0\}$ (100 pairs). As Figure 7 shows, AD AUROC on 8-shot stays above 0.95 across the plane, peaking at $\alpha \in [0.1, 0.3]$, $\beta = 0.2$ (0.969). Although the default $(0.5, 0.5)$ is not the global optimum on MVTec, it remains near-optimal and transfers unchanged to all datasets, confirming that a single fixed weighting delivers strong, domain-agnostic performance.

Impact of Prompt-Scale. Varying the number of antonym pairs K from 5 to 30 reveals a saturation curve. AUROC jumps from $91.2\% \pm 0.8$ at $K = 5$ to $93.2\% \pm 0.4$ at $K = 10$, rises only marginally to $93.4\% \pm 0.4$ at $K = 15$, and plateaus or dips slightly beyond $K \approx 20$. Thus, lexical diversity is beneficial up to a "sweet spot" of roughly 18 pairs, additional pairs introduce semantic noise that offsets the gains. Uniform and TF-IDF token weighting trace almost identical trajectories, indicating that once K is in the effective band, performance is governed more by the breadth of contrasts than by weighting heuristics.

Method	Avg. AUROC	
	AC	AS
Ours	96.6	96.6
w/o Δ_I	95.2 (−1.4)	89.2 (−7.4)
w/o $\Delta_{T(S)}$	92.3 (−4.3)	91.9 (−4.7)
w/o Δ	91.2 (−5.4)	87.8 (−8.8)

Table 7: Average AUROC (%) on MVTec in the 8-shot setting for the DE-CLIP and its three ablated variants.

Method	Avg. AUROC	
	AC	AS
DE-CLIP (w/ $\Delta_{T(S+O)}$)	96.8	96.3
DE-CLIP (w/ $\Delta_{T(S)}$)	96.6	96.6

Table 8: Average AUROC (%) on MVTec in the 8-shot setting for $\Delta_{T(S+O)}$ and $\Delta_{T(S)}$ of DE-CLIP.

Impact of Semantic Editing Components. The ablation results in Table 7 reveal that textual semantic editing and visual anchoring contribute orthogonally. The former enhances semantic separability, while the latter improves alignment with visual prototypes. Their combination introduces structured tension in the embedding space, enabling more robust anomaly detection without training.

$\Delta_{T(S+O)}$ vs $\Delta_{T(S)}$. We evaluate whether incorporating object-specific information in the difference vector is necessary by comparing DE-CLIP with $\Delta_{T(S+O)}$ and its simpler variant $\Delta_{T(S)}$, which uses only adjective-level semantics. As shown in Table 8, both variants achieve nearly identical performance on MVTec, with differences below 0.3% AUROC in both AC and AS tasks.

These results suggest that $\Delta_{T(S)}$ captures sufficient semantic contrast for guiding prompt editing, while offering a lightweight and generalizable alternative to object-conditioned editing.

4.5 Applying DiffEdit to Other Baselines

Table 5 assesses DiffEdit on MVTec by simply prepending its dual guidance $(\Delta_{T(S)}, \Delta_I)$ to each method's original prompt. A clear monotone emerges: the less a baseline has been fine-tuned, the more it gains. Training-free models jump by $> 30\%$ (AC) and $> 20\%$(AS), VAND, whose only learned component is a lightweight head—still improves, but deeply adapted variants (AA-CLIP, AdaCLIP, AnomalyCLIP) move by mere basis points. Extensive task-specific optimisation evidently solidifies CLIP's latent space, leaving little

	AC			AS		
Lexicon	**AUROC**	**AP**	**max-F1**	**AUROC**	**AP**	**max-F1**
LLM-Generated	96.2	95.1	96.9	96.7	56.3	55.0
Hand-Crafted	96.6	94.7	96.6	96.6	55.7	54.2

Table 9: DE-CLIP on MVTec with LLM-generated antonym lexicons.

plasticity for external edits.

On Brain MRI (Table 6), DiffEdit likewise lifts VAND by $10-15\%$ across all shot counts and yields consistent boosts for WinCLIP, confirming that the guidance transfers from industrial to medical imagery.

4.6 Robustness to Lexicon Construction

A common concern is whether DE-CLIP depends on a carefully curated antonym lexicon. To investigate this, we replace the hand-crafted lexicon with automatically generated alternatives. Specifically, we prompt GPT-5 to produce 50 antonym pairs starting from the root pair (*normal*, *defect*), without any domain-specific guidance. For each trial, we randomly sample $K{=}18$ pairs from the generated pool and evaluate on MVTec under the 8-shot setting. We repeat this process across 10 random seeds and report the results in Table 9.

As shown, the LLM-generated lexicons achieve performance comparable to the original hand-crafted one across all metrics. For instance, the average AC-AUROC is 96.2 ± 0.1, closely matching the original 96.6. Notably, the AS-AUROC even slightly improves (96.7 ± 0.3 vs. 96.2), suggesting that the broader lexical diversity introduced by LLM generation can be beneficial for segmentation. These results demonstrate that DE-CLIP does not require domain expertise for lexicon construction. A generic LLM can produce sufficiently effective antonym pairs, greatly enhancing the practical applicability of our method.

4.7 Generalization to Natural Image Classification

To evaluate whether DiffEdit generalizes beyond structured anomaly detection domains, we apply it to standard image classification on CIFAR-100 (Krizhevsky, 2009), a 100-way natural image benchmark. Unlike AD, this task has no notion of normal vs. anomalous states. Instead, the goal is to discriminate among 100 semantically distinct object categories.

Adapting DiffEdit to Multi-Class Classification. We reformulate DiffEdit for the classification setting as follows. For each target class c_i ($i \in \{1, \ldots, 100\}$), we treat c_i as the positive class and the remaining 99 classes $\{c_j\}_{j \neq i}$ as the negative set. The textual difference vector for class c_i is computed as

$$\Delta_T^{(i)} = \frac{1}{99} \sum_{j \neq i} E(c_j) - E(c_i),$$

which captures the semantic direction pointing away from class c_i toward the complement classes. In the few-shot setting, the visual anchor $\Delta_I^{+(i)}$ is the mean CLIP image embedding of K reference images from class c_i, identical to the normal-reference construction in DE-CLIP. Each class embedding is then edited via DiffEdit:

$$\bar{E}(c_i) = E(c_i) - \alpha \cdot \Delta_T^{(i)} + \beta \cdot \Delta_I^{+(i)},$$

which pushes $\bar{E}(c_i)$ away from the complement centroid and toward its own visual prototype. At inference, a test image x is classified as $\arg\max_i \cos\big(E(x),\, \bar{E}(c_i)\big)$.

Shots	**CLIP**	**0**	**2**	**4**	**6**	**8**	**10**
Accuracy	66.9	68.7	71.4	72.2	73.2	73.4	74.3

Table 10: Top-1 accuracy (%) of CLIP with DiffEdit on CIFAR-100 under different shot settings.

Results. Table 10 reports Top-1 accuracy on CIFAR-100. Even in the zero-shot setting, DiffEdit improves CLIP's accuracy from 66.9% to 68.7% by leveraging textual difference vectors alone. With increasing reference shots, accuracy rises steadily to 74.3% at 10-shot, a gain of $+7.3$ percentage points over the CLIP baseline. These results confirm that the semantic editing mechanism underlying DiffEdit is not specific to anomaly detection but extends to general visual recognition, highlighting the broad applicability of difference-guided embedding editing.

5 Conclusion

We propose DiffEdit, a dual-modality, plug-and-play embedding edit that sharpens CLIP's semantic boundaries, and built DE-CLIP, a training-free few-shot AD/AS framework. DiffEdit can be dropped into existing CLIP baselines to boost accuracy without retraining. Experiments on industrial and medical datasets show consistent gains in both classification and segmentation, demonstrating effectiveness and wide applicability in low-resource anomaly detection.

Limitations

Although DE-CLIP works without training, it presumes (i) a hand-crafted antonym lexicon for the textual vector Δ_T, which may be hard to define in domains such as satellite imagery or finance, and (ii) that the textual Δ_T and visual Δ_I point in similar directions, misalignment could weaken the edit, calling for adaptive or confidence-aware fusion. In addition, our anomaly-segmentation head reuses a classic sliding-window pipeline (Li et al., 2024b; Chen et al., 2023; Jeong et al., 2023; Ma et al., 2025) that overlooks long-range structure, hinting at future attention- or generation-based decoders.

Furthermore, many medical anomaly detection datasets provide only abnormal samples (Bao et al., 2023; Bernal et al., 2012, 2015), making them incompatible with the proposed DiffEdit mechanism. Nevertheless, we believe that in most real-world scenarios, normal (positive) samples are substantially easier to collect than defective ones, and thus DE-CLIP remains practical and broadly applicable in realistic deployment settings.

Ethical Considerations

The primary objective of this work is to advance the understanding and safety of anomaly detection (AD) systems based on vision–language models (VLMs). All experiments are conducted using publicly available datasets, including MVTec AD, VisA, and BMAD Brain MRI, which are standard benchmarks widely adopted in the anomaly detection community. These datasets do not contain personally identifiable information or sensitive medical data beyond anonymized, research-grade imagery. Our method, DE-CLIP, operates in a training-free manner and does not require the collection of new data, thereby minimizing potential privacy or environmental concerns associated with large-scale model training.

This research does not pose foreseeable ethical risks such as privacy violations, bias amplification, or misuse. The proposed technique focuses on improving interpretability and robustness in few-shot anomaly detection rather than generating or modifying sensitive visual content. We will publicly release our implementation and experimental configurations to ensure transparency, reproducibility, and the responsible advancement of research in secure and trustworthy AD systems.

References

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. 2023. BMAD: Benchmarks for Medical Anomaly Detection. *CoRR abs/2306.11876*.

Lawrence W. Barsalou. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–660.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9592–9600.

Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111.

Jorge Bernal, Javier Sánchez, and Fernando Vilariño. 2012. Towards Automatic Polyp Detection with a Polyp Appearance Model. *Pattern Recognition*, 45(9):3166–3182.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. 2024. AdaCLIP: Adapting CLIP with Hybrid Learnable Prompts for Zero-Shot Anomaly Detection. In *European Conference on Computer Vision (ECCV)*, pages 55–72. Springer.

Xuhai Chen, Yue Han, and Jiangning Zhang. 2023. APRIL-GAN: A Zero-/Few-Shot Anomaly Classification and Segmentation Method for CVPR 2023 VAND Workshop Challenge Tracks 1&2: 1st Place on Zero-shot AD and 4th Place on Few-shot AD. *CoRR abs/2305.17382*.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. 2021. PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization. In *International Conference on Pattern Recognition (ICPR)*, pages 475–489. Springer.

Hanqiu Deng and Xingyu Li. 2022. Anomaly Detection via Reverse Distillation from One-Class Embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9737–9746.

Sedigheh Eslami and Gerard de Melo. 2024. Mitigate the Gap: Investigating Approaches for Improving Cross-Modal Alignment in CLIP. *CoRR abs/2406.17639*.

Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2021. Deep Learning for Medical Anomaly Detection–A Survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37.

Ronald A. Fisher. 1936. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. 2024. Adapting Visual-Language Models for Generalizable Anomaly Detection in Medical Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11375–11385.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. 2023. WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19606–19616.

Yukun Jiang, Zheng Li, Xinyue Shen, Yugeng Liu, Michael Backes, and Yang Zhang. 2024. ModSCAN: Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12814–12845. ACL.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. 2025. Is CLIP ideal? No. Can we fix it? Yes! *CoRR abs/2503.08723*.

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

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Ronald W. Langacker. 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Stanford University Press.

Ye Leng, Junjie Chu, Mingjie Li, Chenhao Lin, Chao Shen, Michael Backes, Yun Shen, and Yang Zhang. 2026. When Understanding Becomes a Risk: Authenticity and Safety Risks in the Emerging Image Generation Paradigm. *CoRR abs/2603.24079*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning (ICML)*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *CoRR abs/2201.12086*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation. *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 34:9694–9705.

Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. 2024a. MUSC: Zero-Shot Industrial Anomaly Classification and Segmentation with Mutual Scoring of the Unlabeled Images. In *International Conference on Learning Representations (ICLR)*.

Yuanwei Li, Elizaveta Ivanova, and Martins Bruveris. 2024b. FADE: Few-shot/zero-shot Anomaly Detection Engine using Large Vision-Language Model. *CoRR abs/2409.00556*.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the Gap: Understanding the Modality Gap in Multi-Modal Contrastive Representation Learning. In *Annual Conference on Neural Information Processing Systems (NIPS)*, volume 35, pages 17612–17625. Curran Associates, Inc.

Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. 2025. AA-CLIP: Enhancing Zero-Shot Anomaly Detection via Anomaly-Aware CLIP. *CoRR abs/2503.06661*.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. University of Illinois Press, Urbana, IL.

Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. 2024. VCP-CLIP: A Visual Context Prompting Model for Zero-Shot Anomaly Segmentation. *CoRR abs/2407.12276*.

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 *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. 2022. Towards Total Recall in Industrial Anomaly Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14318–14328.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. 2024. Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Representation Learning. In *International Conference on Learning Representations (ICLR)*.

Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. 2022. CLIP Models Are Few-Shot Learners: Empirical Studies on VQA and Visual Entailment. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6088–6100. ACL.

Yucheng Suo, Linchao Zhu, and Yi Yang. 2023. Text Augmented Spatial Aware Zero-Shot Referring Image Segmentation. In *Findings of the Association for Computational Linguistics: Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1032–1043. ACL.

Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and Ilker Yildirim. 2023. When Are Lemons Purple? The Concept Association Bias of Vision-Language Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14333–14348. ACL.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578.

Kirill Tyshchuk, Polina Karpikova, Andrew Spiridonov, Anastasiia Prutianova, Anton Razzhigaev, and Alexander Panchenko. 2023. On Isotropy of Multimodal Embeddings. *Information*, 14(7):392.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. 2022. A Unified Model for Multi-Class Anomaly Detection. *Advances in Neural Information Processing Systems*, 35:4571–4584.

Zhaoxiang Zhang, Hanqiu Deng, Jinan Bao, and Xingyu Li. 2024. Dual-Image Enhanced CLIP for Zero-Shot Anomaly Detection. *CoRR abs/2405.04782*.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. 2023. AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection. *CoRR abs/2310.18961*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *CoRR abs/2304.10592*.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. 2022. SPot-the-Difference Self-Supervised Pre-training for Anomaly Detection and Segmentation. In *European Conference on Computer Vision (ECCV)*, pages 392–408. Springer.

A Linearizing the Hyperspherical Problem in VLM Embeddings

A.1 Ideal Two-Sided Visual Difference

Define the full visual shift

$$\Delta_I^{\pm} \;=\; \mu_+ - \mu_-.$$

Empirical Collinearity. Because CLIP maximises cross-modal alignment, $\cos(\Delta_I^{\pm}, \Delta_T) \approx 1$. Hence there exists a constant $\delta > 0$ such that

$$\Delta_I^{\pm} \approx \delta\, \Delta_T. \tag{2}$$

Fisher Linear Discriminant. With the isotropy assumption $\Sigma_+ \approx \Sigma_- \approx \sigma^2 I$ (Fisher, 1936), the Rayleigh quotient is maximised by

$$w^\star \;\propto\; \mu_+ - \mu_- \;=\; \Delta_I^{\pm}.$$

Impact of DiffEdit. DiffEdit injects a scaled visual shift:

$$\tilde{t}_{\text{pos}} = E(t_{\text{pos}}) + \beta \Delta_I^{\pm}, \qquad \beta \in [0, 1].$$

Because both anchors move by the same vector $\frac{\beta}{2}\Delta_I^{\pm}$,

$$\tilde{t}_{\text{pos}} - E(t_{\text{neg}}) = \Delta_T + \beta \Delta_I^{\pm} \overset{(2)}{\approx} (1 + \beta\delta)\, \Delta_T.$$

Thus any β rescales but never rotates the optimal axis $w^\star$.

A.2 Practical One-Sided Visual Anchor (Normals Only)

In real applications, anomalous images are scarce, so we use the *one-sided* visual anchor

$$\Delta_I^{+} \;=\; \mu_+.$$

Collinearity Still Holds. CLIP alignment yields $\cos(\mu_+, E(t_{\text{pos}})) \approx 1$, and $E(t_{\text{pos}})$ is one endpoint of Δ_T. Therefore

$$\cos(\Delta_I^{+}, \Delta_T) \;\geq\; \cos(\theta_{\max}) \approx 0.9,$$

i.e. the angle between Δ_I^{+} and Δ_T is at most $25°$. Hence there exists a constant $\gamma > 0$ such that

$$\Delta_I^{+} \approx \gamma\, \Delta_T. \tag{3}$$

DiffEdit with One-Sided Guidance. DiffEdit now becomes

$$\hat{t}_{\text{pos}} = E(t_{\text{pos}}) + \beta \Delta_I^{+}.$$

Its difference with respect to the negative anchor is

$$\hat{t}_{\text{pos}} - E(t_{\text{neg}}) = \Delta_T + \beta \Delta_I^{+} \overset{(3)}{\approx} (1 + \beta\gamma)\, \Delta_T.$$

Thus, exactly as in the ideal case, injecting $\beta\Delta_I^{+}$ *only* scales the discriminative axis and never rotates it. Experiments show that performance is most stable for $\beta \in [0.1, 0.3]$.

Both full-sided ($\Delta_I^{\pm}$) and one-sided (Δ_I^{+}) visual guidance satisfy the same collinearity property with Δ_T. Therefore, DiffEdit preserves the optimal FLD direction even when only a small set of normal reference images is available.