TokenCLIP: Token-wise Prompt Learning for Zero-shot Anomaly Detection

Qihang Zhou[1,2], **Bin-Bin Gao**[2], **Guansong Pang**[3], **Xin Wang**[4], **Jiming Chen**[1], **Shibo He**[1*]
[1]Zhejiang University [2] Tencent YouTu Lab
[3] Singapore Management University [4] Qilu University of Technology
[1] {zqhang, cjm, s18he}@zju.edu.cn [2] csgaobb@gmail.com [3]gspang@smu.edu.sg
[4]xinwang@qlu.edu.cn

Abstract

Adapting CLIP for anomaly detection on unseen objects has shown strong potential. Existing methods typically rely on a single textual space to align with visual semantics across diverse objects and domains. The indiscriminate alignment hinders the model from accurately capturing anomaly semantics. We propose TokenCLIP, a token-wise dynamic framework that aligns each visual token with learnt textual subspaces corresponding to its visual characteristics. However, explicitly assigning a unique learnable textual space to each token is computationally intractable and prone to insufficient optimization. We instead expand the token-agnostic textual space into a set of orthogonal subspaces, and then dynamically assign each token to a subspace combination guided by semantic affinity, which jointly supports customized and efficient token-wise adaptation. To this end, we formulate dynamic alignment as an optimal transport problem, where all visual tokens in an image are transported to textual subspaces under the cross-modal cost matrix. **The marginal constraint and minimal cost objective of OT ensure sufficient optimization across subspaces and encourage them to focus on different semantics.** Solving the problem yields a transport plan that adaptively assigns each token to semantically relevant subspaces. Extensive experiments show that the textual subspaces naturally specialize in different semantics, such as foreground and background regions, to promote fine-grained anomaly learning. The comparison between baselines shows the superiority of TokenCLIP.

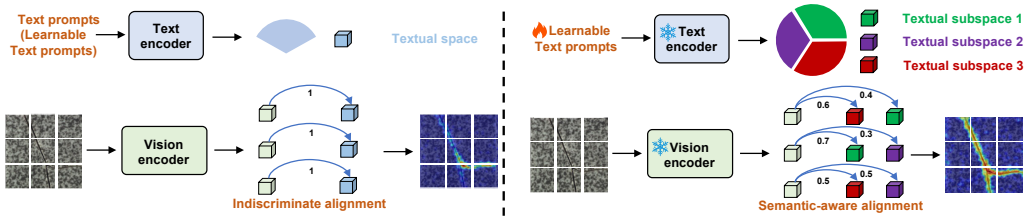
1 Introduction

Foundation Models (FMs) [30, 23, 29] have shown the potential to generalize to unseen class semantics and domains. This breakthrough has driven the rapid development of downstream tasks that explore zero-shot capabilities by adapting FMs [27, 36, 22, 18, 37, 14]. Anomaly detection has also evolved from specialized models toward more generalized detection frameworks [18, 37, 7, 5, 28, 21, 38, 11, 2]. A prominent line of research in this area involves adapting CLIP for zero-shot anomaly detection. These methods typically project either learnable [37, 14] or handcrafted text prompts [18, 7] into a shared embedding space and align them with diverse visual features, capturing both global and local abnormalities. As shown in Figure 1a, this coarse alignment makes it difficult for the detection model to capture generalized anomaly semantics, as the token-agnostic textual space is forced to make a trade-off between diverse semantic tokens.

A natural approach is to assign each visual patch token to its own textual embedding space. However, this design introduces two major challenges. **Challenge 1**: High computational cost. For instance, a 518×518 image typically yields 1,369 patch tokens. Assigning a unique textual embedding to each token requires encoding the same number of distinct text prompts through the text encoder.

*Corresponding authors

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Indiscriminate alignment in previous works. (b) Dynamic alignment in TokenCLIP.

Figure 1: Framework comparison. **(a)**: Previous works rely on a textual space to indiscriminately align the diverse visual patch tokens, which would compromise accurate alignment for diverse anomaly semantics. **(b)**: TokenCLIP introduces multiple orthogonal textual subspaces to dynamically align each visual patch token according to its visual semantics.

This leads to significant computational overhead. **Challenge 2**: Underfitting of textual space due to insufficient optimization. Since each token-specific textual embedding is updated only once during training, it results in severe underfitting and poor textual adaptation. To address these challenges, this paper proposes TokenCLIP, a fine-grained adaptation framework that dynamically aligns each visual patch token with a combinatorial set of orthogonal textual subspaces. Figure 1b shows that TokenCLIP enables token-level alignment based on visual semantics, thereby achieving granular modeling of anomaly semantics. TokenCLIP could circumvent the above challenges by 1) Leveraging weighted combinations of textual subspaces to provide token-level supervision avoids explicitly encoding individual textual embeddings for each token, thereby mitigating the computational burden; 2) Sharing orthogonal textual subspaces across all visual tokens enables sufficient optimization of each subspace and promotes their semantic specialization.

We formulate dynamic alignment as an optimal transport (OT) task, where all visual tokens in an image are transported to textual subspaces based on the cosine similarity between visual and textual representations. In this formulation, the visual patch tokens (source distribution) are required to be transported to the textual subspaces (target distribution), with minimal transport cost. The marginal constraint and minimal cost objective facilitate sufficient optimization and encourage semantic specialization across textual subspaces, respectively. Specifically, we first define a learnable text prompt to construct a base textual space, which captures global anomaly semantics through indiscriminate alignment. Building upon this space, we introduce a multi-head projection to derive multiple orthogonal textual subspaces, further regularized by an orthogonality constraint to promote semantic diversity. Unlike prior many-to-one alignment approaches, we model token-level alignment as a many-to-many correspondence between visual patch tokens and textual subspaces via OT. With visual-textual cosine similarity as the cost matrix, solving this OT problem yields a transport plan, where each entry quantifies the mass (weight) assigned from a visual token to a textual subspace. This formulation provides two key advantages: (1) the marginal constraints of OT ensure that each textual subspace is sufficiently optimized; (2) minimization of total transport cost encourages textual subspaces to specialize in distinct visual semantics. We further sparsify the transport plan by retaining only the topK textual subspaces for each visual token. The selected masses are then normalized to produce soft assignment weights for final alignment. In doing so, TokenCLIP adaptively selects the most semantically relevant combination of textual subspaces for each token, without requiring an explicitly tailored textual space for every token. The main contributions of this paper are as follows:

- We reveal that current methods rely on indiscriminate alignment, which limits the capacity of textual spaces to capture comprehensive anomaly semantics. To address this, we propose TokenCLIP, a novel dynamic alignment framework that enables semantics-aware textual supervision at the token level.
- We formulate the dynamic alignment between tokens and orthogonal textual subspaces as an OT problem, where the marginal constraint and minimal cost objective respectively ensure sufficient optimization and encourage semantic specialization across subspaces. The textual spaces are specialized to different semantics, such as foreground and background regions.
- We conduct extensive experiments across a wide range of object semantics to evaluate the effectiveness of TokenCLIP. Results on both industrial and medical benchmarks demonstrate its superiority in capturing diverse and comprehensive anomaly semantics.

2 Related Work

Zero-shot anomaly detection ZSAD is an emerging field that aims to detect anomalies in unseen object categories and even across domains [9, 24, 10]. Some methods, such as VAND [7] and AdaCLIP [5], focus on adapting CLIP's visual encoder to capture anomaly semantics. However, these approaches rely heavily on human-crafted text prompts to represent normality and abnormality [18]. Another branch, including methods like AnomalyCLIP [37], FiLo [13] takes the opposite approach: rather than adapting the visual space, they learn text prompts to adapt the textual space for modeling anomaly semantics. Additionally, methods such as AACLIP [25], BayesCLIP [28], and AdaptCLIP [11] aim to adapt both the visual and textual spaces for improved performance. Despite these advances, most existing methods rely on a single textual space to simultaneously align with diverse visual patch tokens. FAPrompt [39] ensembles multiple prompts to capture more abnormal patterns, but remains at the image level. In contrast, TokenCLIP introduces a dynamic alignment that provides token-level semantics-aware textual subspaces.

Prompt Learning Prompt learning was proposed to efficiently adapt CLIP for more accurate image classification with minimal computational overhead [36]. AnomalyCLIP [37] extends prompt learning to zero-shot anomaly detection, aiming to capture both local and global semantics for anomaly classification and localization [13]. They typically model a token-agnostic textual space and overlook the semantic differences among local regions. In contrast, TokenCLIP introduces an orthogonal textual subspace and adaptively combines these subspaces to align each token according to its visual semantics. We formulate this dynamic alignment between visual regions and textual prompts as an OT problem. While PLOT [6] also incorporates OT for prompt learning, it is primarily designed for image-level classification. In contrast, we leverage OT for pixel-level anomaly segmentation and image-level anomaly detection, achieving fine-grained and spatially aware alignment.

Optimal Transport OT has emerged as a powerful framework for comparing probability distributions, with wide applications in computer vision [34]. The entropic regularized OT proposed by Cuturi [8] significantly improved computational efficiency via the Sinkhorn algorithm, making OT feasible for large-scale learning. Subsequent works [12] have extended OT to stochastic, unbalanced, and mini-batch scenarios. In multi-modal learning, OT has been employed for fine-grained alignment between modalities (e.g., image patches and text tokens), enabling interpretable and structure-aware correspondence. In contrast to prior work [6], we are the first to introduce OT for fine-grained anomaly semantics learning, particularly in local visual anomaly detection.

3 Methodology

This paper introduces TokenCLIP, a fine-grained adaptation framework for accurate anomaly detection in Figure 2. The key insight of TokenCLIP is to move beyond indiscriminate visual-textual alignment by introducing a dynamic alignment mechanism, which provides token-level textual supervision for each visual patch token. The proposed TokenCLIP framework comprises two key modules: (1) a multi-head text prompt that projects the base textual space into multiple orthogonal subspaces; and (2) a dynamic alignment mechanism by solving the OT plan to assign each visual patch token to the semantically relevant textual subspace or their weighted combinations.

Premiliary Given an auxiliary dataset $\mathcal{D} = \{x_1, x_2, \dots\}$, each image $x \in \mathbb{R}^{3\times H_{\text{image}}\times W_{\text{image}}}$ is accompanied by an image-level label $y \in \mathbb{R}$ and a pixel-level annotation $S \in \mathbb{R}^{H_{\text{image}}\times W_{\text{image}}}$. The image encoder of CLIP encodes x into a global image embedding $f \in \mathbb{R}^d$ and a set of visual patch tokens $V = \{v_i\}_{i=1}^N$. Given the text prompts corresponding to the normality class n and abnormality class a, the text encoder produces the associated textual embeddings g_n and $g_a \in \mathbb{R}^d$. The image-level anomaly score $P_a(g_a, f) \in \mathbb{R}$ and segmentation $S_a \in \mathbb{R}^N$ at index i are given by:

$$P_a(g_a, f) = \frac{\exp\left(\langle g_a, f\rangle/\tau\right)}{\sum_{c\in\{n,a\}}\exp\left(\langle g_c, f\rangle/\tau\right)}, S_a^{(i)} = S_a(g_a, v_i) = \frac{\exp\left(\langle g_a, v_i\rangle/\tau\right)}{\sum_{c\in\{n,a\}}\exp\left(\langle g_c, v_i\rangle/\tau\right)},$$

where τ is a temperature scaling factor, and $\langle\cdot,\cdot\rangle$ denotes cosine similarity.

3.1 Multi-head text prompt learning

As mentioned above, the indiscriminate alignment suffers from performance sacrifice from using one textual space to match all visual semantics, including local and global abnormality. Therefore, we first use separate learnable text prompts, i.e., pixel-level text prompt and image-level text prompt, to

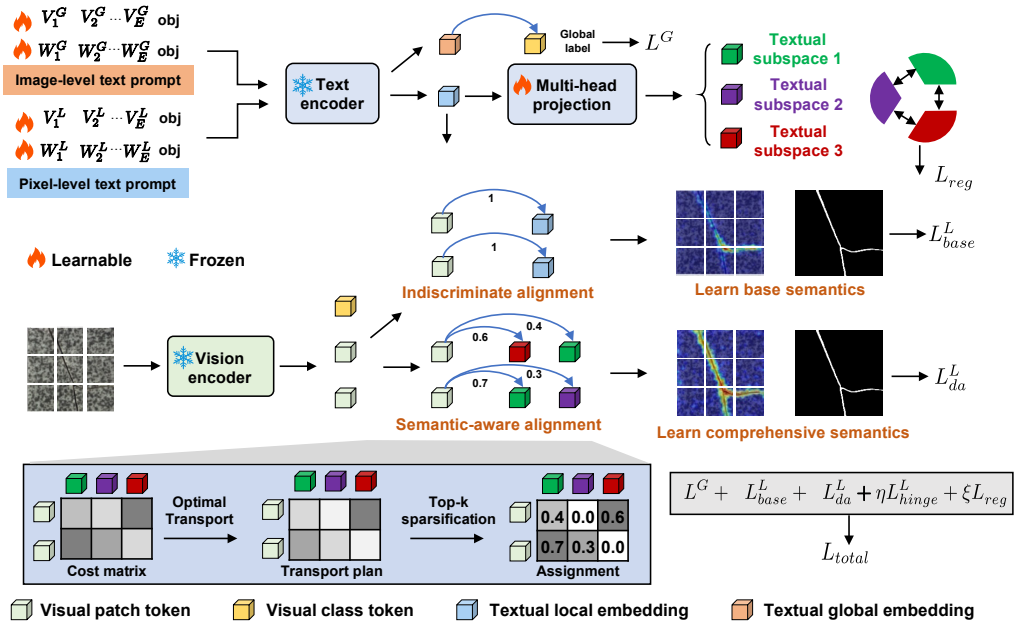


Figure 2: The Framework of TokenCLIP. TokenCLIP uses separate text prompts to learn global and local anomaly semantics. The global text embedding is aligned with the visual class token to detect image-level anomalies. In parallel, a corresponding local text prompt provides indiscriminate alignment with all visual patch tokens to capture base-level anomaly semantics. Building upon this, TokenCLIP employs a multi-head projection to map the base textual space into multiple orthogonal textual subspaces, regularized to encourage semantic diversity. The alignment between these subspaces and the visual patch tokens is then formulated as an OT problem.

model the local and global anomaly semantics. Formally, we define the pixel-/image-level prompt as $G = \{G_n, G_a\}$ and $L = \{L_n, L_a\}$. Each of them contains normality and abnormality prompts. Following AnomalyCLIP, we use the object-agnostic text prompt as follows:

$$\begin{aligned} G_n &= [V_1^G] \cdots [V_E^G][object], \quad G_a = [W_1^G] \cdots [W_E^G][damaged][object]. \\ L_n &= [V_1^L] \cdots [V_E^L][object], \quad L_a = [W_1^L] \cdots [W_E^L][damaged][object], \end{aligned} \tag{1}$$

where V and W are learnable word embeddings, and E is the length of the learnable embedding. Using separate text prompts can help decouple local and global anomaly semantics. However, this separation would not be absolute, as local anomaly semantics can enhance global anomaly recognition. Motivated by this, we incorporate local anomaly semantics into the global text prompt by concatenating $G_n \in \mathbb{R}^{E \times D}$ and $L_n \in \mathbb{R}^{E \times D}$ along the channel dimension to form the final image-level text prompts $\bar{G}_n \in \mathbb{R}^{E \times D}$: $\bar{G}_n = \text{MLP}([G_n, L_n])$. MLP denotes a multi-layer perceptron, and $[\cdot, \cdot]$ represents the concatenation. The same operation is applied to G_a: $\bar{G}_a = \text{MLP}([G_a, L_a])$.

The text encoder encodes $\bar{G}_n$ and $\bar{G}_a$ to derive both global embeddings $\bar{g}_n$ and $\bar{g}_a$, and encodes L_n and L_a to derive local embeddings l_n and l_a. We capture global anomaly semantics by matching the global textual embedding with the visual class token. The global loss L_g is computed via cross-entropy as follows:

$$L_g = \text{CrossEntropy}([P_n(\bar{g}_n, f), P_a(\bar{g}_a, f)], y). \tag{2}$$

Inspired by AnomalyCLIP, l_n and l_a are indiscriminately aligned with all visual patch tokens $\{v_i\}_{i=1}^N$. Although this indiscriminate alignment fails to model each patch token precisely, this coarse-grained modeling can provide a base anomaly semantics. To mitigate imbalance and promote the anomaly boundary, we combine Focal loss and Dice loss to learn the base anomaly semantics. Formally, the base local loss L_{base} is given by:

$$L_{base} = \text{Focal}(\text{Up}([S_n, S_a]), S) + \text{Dice}(\text{Up}(S_a), S) + \text{Dice}(\text{Up}(S_n), I - S), \tag{3}$$

where $S_n^{(i)} = S_n(l_n, v_i)$ and $S_a^{(i)} = S_a(l_a, v_i)$ represent the i-th element of S_n and S_a, respectively, and $\mathrm{Up}(\cdot)$ denotes bilinear interpolation for upsampling.

Building on the base textual space, we further construct more fine-grained textual spaces. Specifically, we apply a multi-head projection to project the base embeddings into multiple fine-grained embeddings, i.e., $O_n = \{o_n^1, \cdots, o_n^j, \cdots, o_n^Q\}$ and $O_a = \{o_a^1, \cdots, o_a^j, \cdots, o_a^Q\}$, where $o_n^j \in \mathbb{R}^d$ and $o_a^j \in \mathbb{R}^d$. The process can be formally defined as:

$$\{o_n^j\}_{j=1}^Q = \text{MultiHead}_n(l_n), \quad \{o_a^j\}_{j=1}^Q = \text{MultiHead}_a(l_a). \tag{4}$$

Each head is implemented as a single-layer multilayer perceptron (MLP). To encourage semantic diversity and minimize redundancy among subspaces, we impose an orthogonality regularization:

$$\mathcal{L}_{\text{reg}} = \left\| [\tilde{o}_n^1, \cdots, \tilde{o}_n^Q]^\top [\tilde{o}_n^1, \cdots, \tilde{o}_n^Q] - I \right\|^2 + \left\| [\tilde{o}_a^1, \cdots, \tilde{o}_a^Q]^\top [\tilde{o}_a^1, \cdots, \tilde{o}_a^Q] - I \right\|^2, \tag{5}$$

where $\tilde{o}_n^j = o_n^j / \|o_n^j\|$ and $\tilde{o}_a^j = o_a^j / \|o_a^j\|$ are the ℓ^2-normalized embeddings at j-th textual subspaces.

3.2 Dynamic Alignment via OT

Explicitly providing a unique textual embedding for each patch token is computationally prohibitive due to the high cost of text encoding. To circumvent this, we propose to implicitly provide token-level supervision through a semantic-aware combination of textual subspaces. In this paper, we formulate the dynamic alignment as an OT problem between all visual patch tokens and the textual subspaces. The minimal cost objective enables each subspace to specialize in distinct anomaly semantics, while the marginal constraints ensure that all subspaces are sufficiently optimized. We consider visual embedding set $\{v_i\}_{i=1}^N$ and **textual embedding set** $\{o_c^j\}_{j=1}^Q, c \in \{n, a\}$ in the discrete formulation of OT. We define two empirical probability distributions supported on $\mathcal{V}$ and $\mathcal{O}$: $u = \sum_{i=1}^N p_i \delta_{v_i}$ and $v = \sum_{j=1}^Q q_j \delta_{o_c^j}$, where δ_{v_i} and $\delta_{o_c^j}$ denote Dirac delta measures centered at $v_i \in \mathcal{V}$ and $o_c^j \in \mathcal{O}$, respectively. The weights p_i and q_j are non-negative and satisfy $\sum_i p_i = \sum_j q_j = 1$ for marginal normalization.

The goal is to find a transport plan $T_c \in \mathbb{R}^{N \times Q}$ that minimizes the total transport cost between the two distributions. To reflect the affinity between visual and textual space, we compute the cosine distance between all visual and textual tokens to construct a cost matrix $C \in \mathbb{R}^{N \times Q}$, where $C_{ij} = 1 - \frac{v_i^\top o_c^j}{\|v_i\| \cdot \|o_c^j\|}$. However, solving this original problem is time-consuming due to its large computational complexity. We employ the Sinkhorn-Knopp algorithm to accelerate the solution of OT problems through entropic regularization.

$$T_c^* = \min_{T_c \in \Pi(u,v)} \sum_{i=1}^N \sum_{j=1}^Q (T_c \odot C)_{ij} - \lambda E(T_c), \ \text{ subject to } \ T_c 1^Q = u, \ T_c 1^N = v, \tag{6}$$

where $\Pi(u, v) \in \mathbb{R}^{N \times Q}$ is the set of all joint distributions with marginals u and v. The second term $E(T_c) = \sum_{ij} (T_c \odot \log T_c)_{ij}$ is the entropy regularization, and $\lambda > 0$ is the regularization coefficient. This term encourages more numerically stable solutions. Moreover, the resulting optimization problem becomes strictly convex, enabling the OT plan to be computed in fewer iterations:

$$T_c^* = \mathrm{diag}(u^t) \exp\left(-\frac{C}{\lambda}\right) \mathrm{diag}(v^t), \tag{7}$$

where t is the iteration step, u^t and v^t are the scaling vectors updated via: $u^t = \mu / \left(\exp(-C/\lambda) v^{t-1}\right)$, $v^t = \nu / \left(\exp(-C/\lambda)^\top u^t\right)$, with the initialization $v^0 = \mathbf{1}$. T^* is the plan minimizing the total cost, and it provides an underlying mapping for visual tokens to the given textual subspaces. Given the transport plan T_c^*, we retain only the topK entries in each row (i.e., per token), applying a threshold ϵ to filter out small values. This is because some low mass would interrupt the specialization of textual subspace learning. The selected values are then row-wise normalized as the affinities to serve as the final assignment matrix $A_c \in \mathbb{R}^{N \times Q}$, where each row is a sparse soft selection of textual subspaces for a visual patch token. Since affinity reflects the matching extent of visual space and textual space, the assignment is semantic-aware.

$$A_c^{ij} = \begin{cases} (T_c^*)_{ij}, & \text{if } j \in \text{topK}\big((T_c^*)_{i,:}, k\big) \& (T_c^*)_{ij} > \epsilon \\ 0, & \text{otherwise,} \end{cases} \quad \bar{A}_c^{ij} = \begin{cases} \dfrac{A_c^{ij}}{\sum_l A_c^{il}}, & \text{if } A_c^{ij} \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Holding sparse assignment matrix $\bar{A} \in \mathbb{R}^{N \times Q}$, we can compute the logits for the class c. The final anomaly score for each visual patch token v_i is calculated as:

$$S_a^{\text{da}}(i) = \frac{\exp\left(z_a^i/\tau\right)}{\sum_{c\in\{n,a\}}\exp\left(z_c^i/\tau\right)},\ \ z_c^i = \left[\bar{A}_c^{i1} \quad \cdots \quad \bar{A}_c^{iQ}\right]\cdot\left[\langle o_c^1, v_i\rangle, \cdots, \langle o_c^Q, v_i\rangle\right]^{\mathrm{T}}. \tag{9}$$

This process allows each visual patch token to be dynamically aligned with a weighted combination of textual subspaces, enabling fine-grained and semantically aware cross-modal alignment. We introduce a dynamic alignment loss L_{da} to achieve fine-grained anomaly modeling.

$$L_{da} = \text{Focal}(\text{Up}([S_n^{da}, S_a^{da}]), S) + \text{Dice}(\text{Up}(S_a^{da}), S) + \text{Dice}(\text{Up}(S_n^{da}), I - S). \tag{10}$$

OT penalizes subspace mixture and induces specialization For ℓ_2-normalized visual features v_i and textual subspace o_j, we have $C_{ij} = 1 - \langle v_i, o_j\rangle = \frac{1}{2}\|v_i - o_j\|^2$, So the balanced OT objective becomes $\mathcal{L}_{\text{OT}} = \sum_{i=1}^{N}\sum_{j=1}^{Q} T_{ij}\,\|v_i - o_j\|^2, T \in \Pi(u, v)$. Let $\mathcal{C}_p$ and $\mathcal{C}_q$ be two visual clusters with centroids μ_p and μ_q, and variances $\sigma_p^2 = \frac{1}{|\mathcal{C}_p|}\sum_{i\in\mathcal{C}_p}\|v_i - \mu_p\|^2$ and $\sigma_q^2 = \frac{1}{|\mathcal{C}_q|}\sum_{i\in\mathcal{C}_q}\|v_i - \mu_q\|^2$. If a subspace o_j receives positive mass $\alpha_p > 0$ and $\beta_q > 0$ from $\mathcal{C}_p$ and $\mathcal{C}_q$, respectively, then:

$$\underbrace{\sum_{i\in\mathcal{C}_p} T_{ij}\|v_i - o_j\|^2 + \sum_{i\in\mathcal{C}_q} T_{ij}\|v_i - o_j\|^2}_{\text{Subspace mixture cost}} \geq \underbrace{\alpha_p\sigma_p^2 + \beta_q\sigma_q^2}_{\substack{\text{Subspace}\\ \text{specialization cost}}} + \underbrace{\frac{\alpha_p\beta_q}{\alpha_p + \beta_q}\|\mu_p - \mu_q\|^2}_{\text{Penalty}}. \tag{11}$$

Since $\|\mu_p - \mu_q\| > 0$ for any pair of distinct clusters, the inequality implies that mixing clusters in a single subspace always incurs a strictly higher OT cost. Equality holds only when $\mu_p = \mu_q$, i.e., $\mathcal{C}_p$ and $\mathcal{C}_q$ are the same cluster. Therefore, the OT minimizer naturally induces subspace specialization.

3.3 Training and inference

Training We train TokenCLIP in an end-to-end manner to capture global anomaly semantics L_g, local anomaly semantics L_{base} and L_{da}. In addition to the regularization term L_{reg}, we introduce a hinge loss to explicitly enforce separation between normal regions and anomalous regions. Let the normal and anomaly indices be defined as: $\mathbb{I}_n = \{i \mid S(i) = 0\}$ and $\mathbb{I}_a = \{i \mid S(i) = 1\}$, where $S(i)$ denotes the ground-truth pixel-level annotation. The hinge loss is formulated as:

$$L_{\text{hinge}} = \frac{1}{|\mathbb{I}_n|}\sum_{i\in\mathbb{I}_n}\max(S_n^{\text{da}}(i) - \delta^-,\, 0) + \frac{1}{|\mathbb{I}_a|}\sum_{i\in\mathbb{I}_a}\max(\delta^+ - S_a^{\text{da}}(i),\, 0), \tag{12}$$

where $S_a^{\text{da}}(i)$ is the predicted anomaly score for token i, and δ^-, δ^+ are thresholds for enforcing margin constraints. The total loss is:

$$L_{\text{total}} = L_g + L_{\text{base}} + L_{\text{da}} + \eta L_{\text{hinge}} + \xi L_{\text{reg}}, \tag{13}$$

where η, ξ are weighting coefficients for the global loss, regularization loss, base local loss, dynamic alignment loss, and hinge loss, respectively.

Inference Given an image x, TokenCLIP could simultaneously provides image-level anomaly score $\mathcal{A}_I(x)$ and pixel-level segmentation $\mathcal{A}_S(x)$. The pixel-level anomaly score combines anomaly segmentation from indiscriminate alignment and dynamic alignment $\mathcal{A}_S(x) = \frac{1}{2}(S_a^{da} + S_a)$. Considering the maximum anomaly score of local anomaly could reflect the image-level anomaly, the image-level anomaly score is given as $\mathcal{A}_I(x) = \frac{1}{2}(P_a(\bar{g}_a, f) + \frac{1}{2}max(\mathcal{A}_S(x))$.

4 Experiments

Dataset Details & Baselines We evaluate TokenCLIP in ZSAD through large-scale experiments. The evaluation covers two distinct domains: industrial inspection and medical diagnosis. In the industrial domain, we evaluate seven benchmarks: MVTec AD [3], VisA [40], MPDD [19], BTAD [26], SDD [32], DAGM [35], and DTD-Synthetic [1]. These datasets span various manufactured objects and defect types. In the medical domain, we assess tasks including skin lesion detection (ISIC [15]), colon polyp segmentation (CVC-ClinicDB [4], CVC-ColonDB [33], Kvasir [20], Endo [17]), and brain anomaly detection (HeadCT, BrainMRI [31], Br35H [16]). Detailed dataset, implementation, and evaluation metric are provided in Appendices A and B.

Implementation details We use publicly available CLIP (`ViT-L/14@336px`) as the backbone. For data preprocessing, we adopt the same pipeline as AnomalyCLIP to ensure fair comparison. All input images are resized to 518×518. We use the top feature as the set of visual patch tokens $\{v_i\}_{i=1}^{N}$. The length of the learnable text embedding is set to $E = 12$. For the textual subspace configuration, we use $Q = 3$ subspaces when training on the VisA dataset, and $Q = 4$ subspaces for MVTec AD.

Table 1: ZSAD performance on industrial domain datasets. Best: Red; Second-best: Blue. Image-level performance (AUROC, AP); Pixel-level performance (AUROC, PRO).

Task	Dataset	CoOp (IJCV'22)	WinCLIP (CVPR'23)	VAND (ARXIV'23)	AdaCLIP (ECCV'24)	AnomalyCLIP (ICLR'24)	FAprompt (ICCV'25)	TokenCLIP (Ours)
Image-level	MVTec AD	(88.8, 94.8)	(91.8, 96.5)†	(86.1, 93.5)†	(89.6, -)†	(91.5, 96.2)†	(90.8, 94.9)	(93.5, 96.7)
	VisA	(62.8, 68.1)	(78.1, 81.2)†	(78.0, 81.4)†	(83.8, -)†	(82.1, 85.4)†	(83.6, 85.6)	(85.8, 88.2)
	MPDD	(55.1, 64.2)	(63.6, 69.9)	(73.0, 80.2)	(76.8, -)†	(77.0, 82.0)†	(77.5, 82.2)	(80.0, 82.3)
	BTAD	(66.8, 77.4)	(68.2, 70.9)	(73.6, 68.6)	(88.6, -)†	(88.3, 87.3)†	(92.3, 93.0)	(91.0, 90.5)
	SDD	(74.9, 65.1)	(84.3, 77.4)	(79.8, 71.4)	(-, -)	(84.7, 80.0)†	(81.8, 77.5)	(88.1, 85.2)
	DAGM	(87.5, 74.6)	(91.8, 79.5)	(94.4, 83.8)	(98.3, -)†	(97.5, 92.3)†	(96.9, 90.6)	(98.7, 95.2)
	DTD	(-, -)	(93.2, 92.6)	(86.4, 95.0)	(95.5, -)†	(93.5, 97.0)†	(95.6, 97.4)	(95.8, 97.6)
Pixel-level	MVTec AD	(33.3, 6.7)	(85.1, 64.6)†	(87.6, 44.0)†	(90.3, -)†	(91.1, 81.4)†	(90.6, 81.6)	(92.2, 87.9)
	VisA	(24.2, 3.8)	(79.6, 56.8)†	(94.2, 86.8)†	(95.6, -)†	(95.5, 87.0)†	(95.6, 86.7)	(95.9, 88.5)
	MPDD	(15.4, 2.3)	(76.4, 48.9)	(94.1, 83.2)	(96.4, -)†	(96.5, 88.7)†	(95.7, 85.6)	(96.8, 89.3)
	BTAD	(28.6, 3.8)	(72.7, 27.3)	(60.8, 25.0)	(92.1, -)†	(94.2, 74.8)†	(94.8, 75.1)	(95.1, 78.3)
	SDD	(28.9, 7.1)	(68.8, 24.2)	(79.8, 65.1)	(-, -)	(90.6, 67.8)†	(92.5, 70.0)	(90.8, 70.3)
	DAGM	(17.5, 2.1)	(87.6, 65.7)	(82.4, 66.2)	(91.0, -)†	(95.6, 91.0)†	(98.1, 94.9)	(95.8, 91.6)
	DTD	(-, -)	(83.9, 57.8)	(95.3, 86.9)	(96.9, -)†	(97.9, 92.3)†	(98.0, 92.2)	(98.1, 93.7)

Table 2: Cross-domain ZSAD performance on medical analysis. Best: Red; Second-best: Blue. Image-level performance (AUROC, AP); Pixel-level performance (AUROC, PRO).

Task	Dataset	CoOp (IJCV'22)	WinCLIP (CVPR'23)	VAND (ARXIV'23)	AdaCLIP (ECCV'24)	AnomalyCLIP (ICLR'24)	FAprompt (ICCV'25)	TokenCLIP (Ours)
Image-level	HeadCT	(78.4, 78.8)	(81.8, 80.2)	(89.1, 89.4)	(91.5, -)	(93.4, 91.6)	(93.9, 92.6)	(96.0, 95.3)
	BrainMRI	(61.3, 44.9)	(86.6, 91.5)	(89.3, 90.9)	(94.8, -)	(90.3, 92.2)	(94.8, 93.7)	(95.3, 95.8)
	Br35H	(86.0, 87.5)	(80.5, 82.2)	(93.1, 92.9)	(97.7, -)	(94.6, 94.7)	(96.6, 95.6)	(97.8, 97.6)
Pixel-level	ISIC	(51.7, 15.9)	(83.3, 55.1)	(89.4, 77.2)	(88.3, -)	(89.7, 78.4)	(90.7, 80.3)	(91.6, 83.4)
	ColonDB	(40.5, 2.60)	(70.3, 32.5)	(78.4, 64.6)	(79.1, -)	(81.9, 71.3)	(84.1, 73.2)	(83.8, 72.8)
	ClinicDB	(34.8, 2.40)	(51.2, 13.8)	(80.5, 60.7)	(84.4, -)	(82.9, 67.8)	(83.9, 69.3)	(84.2, 69.7)
	Kvasir	(44.1, 3.50)	(69.7, 24.5)	(75.0, 36.2)	(-, -)	(78.9, 45.6)	(80.4, 46.5)	(80.6, 46.9)
	Endo	(40.6, 3.90)	(68.2, 28.3)	(81.9, 54.9)	(-, -)	(84.1, 63.6)	(85.9, 65.0)	(86.1, 66.2)

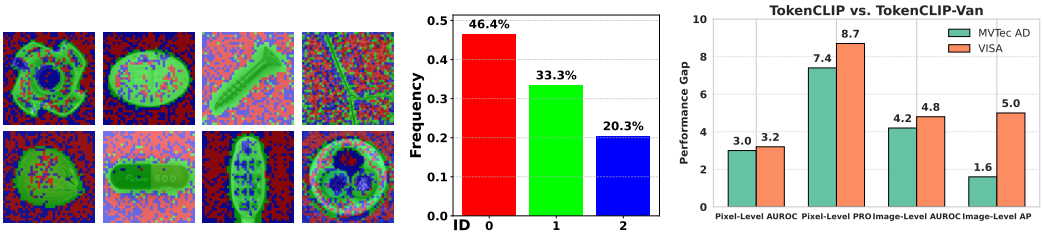
The OT problem is solved using the Sinkhorn-Knopp algorithm with 100 iterations, and the entropic regularization coefficient λ is set to 0.01. The marginal vectors $\mathbf{u}$ and $\mathbf{v}$ are initialized as uniform probability distributions. Accordingly, their weights are set to: $p_i = \frac{1}{N}$ and $q_j = \frac{1}{Q}$. We set the sparsification threshold parameter $\epsilon = 0.2$ and retain the top 2 entries per row in the transport plan. The loss weights η, ξ are set to 5 and 100, respectively. We use Adama optimizer with a learning rate of 1e-3 with batch size 8. The training epoch is 30. All experiments are conducted using PyTorch 2.0.0. † denotes results taken from original papers.

4.1 Main results

ZSAD performance on industrial defect detection We evaluate the effectiveness of TokenCLIP on ZSAD task across seven industrial datasets spanning diverse object categories. As presented in Table 1, TokenCLIP outperforms state-of-the-art baselines. On MVTec AD, it achieves a pixel-level performance of 92.2 AUROC and 87.9 PRO, surpassing AnomalyCLIP's 91.1 AUROC and 81.4 PRO. Notably, TokenCLIP demonstrates significant improvements in PRO, underscoring its superiority in detecting fine-grained and subtle anomalies. These improvements are primarily attributed to its dynamic token-level textual supervision, which enables more precise and semantics-aware alignment for anomaly modeling. In addition to pixel-level gains, TokenCLIP also shows obvious improvements in image-level anomaly detection. This stems from the decoupled modeling of global and local anomaly semantics. This shows that TokenCLIP captures fine-grained, generalized anomaly semantics. We observe that FAPrompt is a competitive model for pixel-level segmentation, but it requires higher computational overhead to learn multiple prompts. Additional visualizations see Appendix D.

ZSAD performance on cross-domain Medical analysis To further demonstrate the advantage of dynamic alignment over indiscriminate alignment, we use the checkpoint trained on MVTec AD to evaluate performance on medical datasets directly. As shown in Table 2, TokenCLIP consistently outperforms other methods at both the image and pixel levels. On ISIC, TokenCLIP achieves 91.6 AUROC and 83.4 PRO, compared to the second-best performance of 90.7 AUROC and 80.3 PRO.

In addition, image-level performance on HeadCT, BrainMRI, and Br35H demonstrates substantial improvements, further confirming the model's ability to capture generalized anomaly semantics.



(a) Token-level textual subspace. (b) Selection frequency. (c) TokenCLIP vs. TokenCLIP-Van.

Figure 3: (a) Visualization of the assigned textual subspaces for patch tokens; different colors denote different textual subspaces. (b) Selection frequency of textual subspaces across MVTec AD, where the color corresponds to that in (a). (c) Performance gap between TokenCLIP and TokenCLIP-Van.

4.2 Result analysis

Fine-grained alignment for token-level modeling In this section, we analyze the semantic patterns captured by the textual subspaces during dynamic alignment. As shown in Figure 3a, the green textual subspace is frequently assigned to foreground regions of objects such as nuts, pills, screws, and tiles. This indicates that it captures object-centric semantics. In contrast, the red and blue textual subspaces are predominantly distributed across background regions. This suggests that they primarily model contextual or low-variation areas. Furthermore, the green subspace tends to concentrate in regions with significant semantic variation, while the red and blue subspaces are more commonly associated with smooth or homogeneous textures. For example, in the tile image, the surface is largely uniform, but the crack introduces a distinct visual change, precisely where the green subspace becomes dominant. In cable, the missing wire appears as a flat black region with low-variance semantic nature and is typically assigned to the red or blue subspaces. We can conclude that the learned textual subspaces serve distinct semantic roles: one subspace captures object-level and variant semantics, while the others are more aligned with background and uniform regions. In addition, we analyze the selection frequency of each textual subspace. Figure 3b presents the frequency distribution of the selected textual subspaces across the MVTec AD dataset. We observe that the blue and green subspaces account for the majority of selections. This indicates that background regions or areas with low semantic variation occupy a large portion of the images.

OT is important to dynamic alignment We investigate the role of OT in dynamic alignment by comparing TokenCLIP with a variant called TokenCLIP-Van, which replaces OT with a simpler mechanism that directly selects the textual subspace that has the highest cosine similarity. As shown in Figure 3c, TokenCLIP consistently outperforms TokenCLIP-Van across all evaluation metrics. This improvement arises from the marginal constraints and minimal cost objective of the OT formulation, which globally optimizes the assignment between visual patch tokens and textual subspaces, allowing textual subspaces sufficient optimization and semantic specialization. In contrast, TokenCLIP-Van relies solely on local cosine similarity. Without global regularization, as illustrated in Figure 4, textual subspaces fail to collectively and distinctly capture different semantics, since each feature only greedily matches more visual tokens during training. When one subspace dominates the matching, others struggle with insufficient optimization.

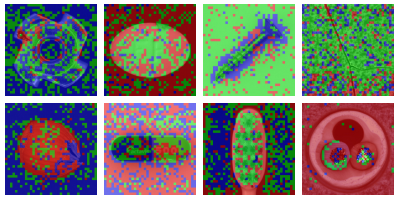
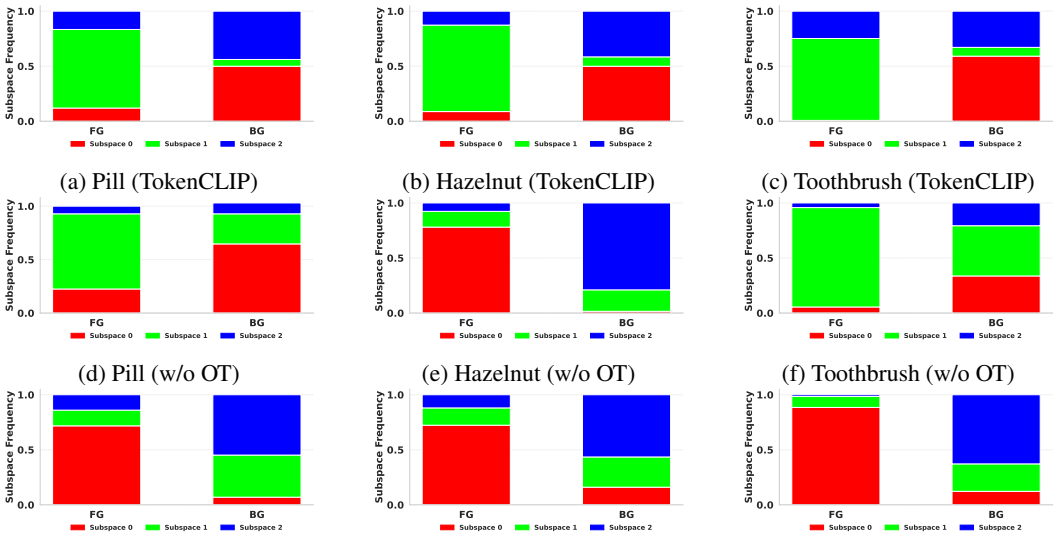


Figure 4: Token-level assignment.

What makes the textual specialization Here, we discuss the contributions of OT and orthogonality regularization to textual specialization. To this end, we visualize the subspace assignment frequency on background and foreground regions across different categories in Figure 5. The first row shows the results of TokenCLIP. The green subspace mainly focuses on object foregrounds (FG), while the blue and red subspaces focus on backgrounds (BG). This demonstrates that TokenCLIP successfully achieves subspace specialization, consistent with the observation in Figure 3a. The middle row shows TokenCLIP without OT. In this case, no clear textual specialization is observed. For example, in the Toothbrush category, the green subspace dominates both FG and BG assignments, while in Hazelnut it is rarely used. This indicates that OT plays a dominant role in enabling fine-grained semantic specialization. The bottom row shows TokenCLIP without orthogonality regularization. The red



(a) Pill (TokenCLIP) (b) Hazelnut (TokenCLIP) (c) Toothbrush (TokenCLIP)

(d) Pill (w/o OT) (e) Hazelnut (w/o OT) (f) Toothbrush (w/o OT)

(g) Pill (w/o Reg.) (h) Hazelnut (w/o Reg.) (i) Toothbrush (w/o Reg.)

Figure 5: Subspace assignment frequency on foreground (FG) and background (BG). Different colors correspond to different textual subspaces. **Top row:** Green subspace of TokenCLIP dominates FG across diverse categories, while blue and red subspaces dominate BG. **Middle row:** Without OT, no subspaces are specialized across categories. **Bottom row**: Without orthogonality regularization, subspaces exhibit some specialization but do not outperform TokenCLIP.

subspace dominates the FG regions, while the blue and green subspaces are mainly assigned to the BG regions. The specialization still emerges even without orthogonality regularization, but the extent of specialization is weaker than that of the full TokenCLIP model. Analysis on subspace textual orthogonality and specialization and subspace initialization are provided in Appendices E and F.

Computation overhead TokenCLIP efficiently provides token-level supervision by combining a limited number of textual subspaces, rather than assigning a dedicated textual space to each visual token. We measure GPU memory consumption during training and report inference speed in terms of frames per second (FPS). All experiments are conducted on an idle GPU with a batch size of 1. Table 3 shows that TokenCLIP incurs only a slight increase in inference time and GPU memory usage compared to AnomalyCLIP. In contrast, FAPrompt requires 1.9× more GPU memory and nearly double the inference time.

Table 3: Analysis of computation overhead.

Methods	Training per epoch	Inference Time	FPS	Peak GPU Memory	MVTec AD pixel-level	VisA pixel-level
AnomalyCLIP	126s	0.139s	7.19	2235MB	(91.1, 81,4)	(95.5, 87.0)
FAprompt	284s	0.276s	3.62	4238MB	(90.6, 81,6)	(95.6, 86.7)
TokenCLIP	149s	0.160s	6.25	3186MB	(92.2, 87.9)	(95.9, 88.5)

Table 4: Module Ablation.

Module T_1	T_2	T_3	T_4	MVTec AD Pixel-level	MVTec AD Image-level	VisA Pixel-level	VisA Image-level
				(80.3, 77.8)	(89.9, 95.4)	(86.6, 78.1)	(82.2, 84.9)
✓				(91.1, 81,4)	(91.5, 96.2)	(95.5, 87.0)	(82.1, 85.4)
✓	✓			(91.8, 87.0)	(93.1, 96.6)	(95.6, 87.8)	(84.5, 88.0)
✓	✓	✓		(91.7, 87.4)	(93.2, 96.8)	(95.8, 88.2)	(85.1, 87.9)
✓	✓		✓	(91.8, 87.2)	(92.9, 96.0)	(95.7, 88.0)	(85.2, 87.5)
✓		✓	✓	(91.5, 83.2)	(92.4, 96.3)	(95.2, 87.6)	(83.0, 85.8)
	✓	✓	✓	(83.2, 62.5)	(89.6, 94.8)	(95.2, 87.1)	(84.8, 86.7)
✓	✓	✓	✓	(92.2, 87.9)	(93.5, 96.7)	(95.9, 88.5)	(85.8, 88.2)

Table 5: Subspace number ablation.

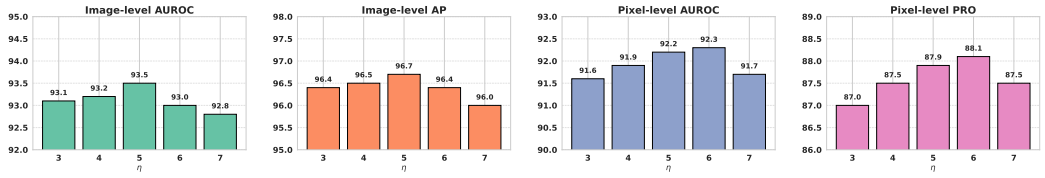
Q	MVTec AD Pixel-level	MVTec AD Image-level	VisA Pixel-level	VisA Image-level
1	(91.6, 85.3)	(92.1, 96.2)	(95.3, 87.2)	(83.6, 85.9)
2	(91.8, 86.8)	(93.0, 96.2)	(95.5, 87.4)	(84.2, 86.8)
3	(92.2, 87.9)	(93.5, 96.7)	(95.7, 87.8)	(85.3, 87.7)
4	(91.8, 86.8)	(93.2, 96.5)	(95.9, 88.5)	(85.8, 88.2)
5	(91.5, 86.2)	(93.3, 96.1)	(95.1, 87.2)	(85.0, 87.8)

4.3 Ablation study

Module Ablation In this section, we investigate the contributions of four key modules to the overall performance of TokenCLIP: base semantics learning (T_1), OT (T_2), decoupled local and global

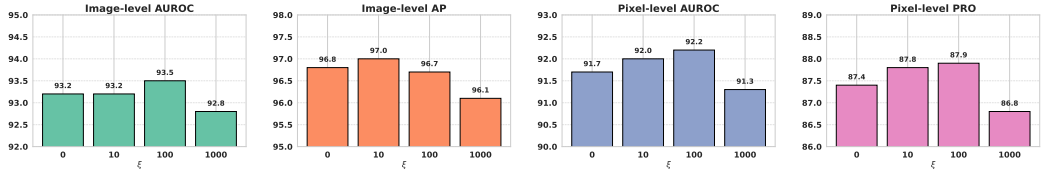
text prompts (T_3), and orthogonal regularization (T_4). The vanilla model, which includes only T_1, captures global anomaly semantics and serves as the basic baseline. Adding base local anomaly modeling (T_2) yields a notable performance boost, corresponding to AnomalyCLIP. This gain is primarily attributed to the indiscriminate alignment that enables learning a token-agnostic unified textual space. Decoupling global and local anomaly semantics (T_3) further improves both image-level and pixel-level performance by allowing each prompt to specialize. Building on this, we incorporate orthogonal regularization (T_4) to encourage diversity among the textual subspaces. Additional gains are observed due to finer semantic specialization. Notably, when we remove T_1, TokenCLIP exhibits a significant performance drop on MVTec AD. This suggests that base anomaly semantics promote learning stability of dynamic alignment.

Number Ablation of Textual Subspaces. Textual subspaces enable TokenCLIP to perform fine-grained alignment by capturing diverse visual semantics. As shown in Table 5, using three and four subspaces yields the best overall results on MVTec AD and VisA, respectively. When the number of subspaces is too small (e.g., 1), the model fails to capture sufficient anomaly-related semantics. Increasing the number of subspaces from 1 to 3 leads to notable performance gains. However, using too many subspaces (e.g., 5) may result in suboptimal optimization due to fragmented semantic representations. Threshold Ablation and K Ablation are provided in Appendices 8 and 7.



(a) Image-level AUROC. (b) Image-level AP. (c) Pixel-level AUROC. (d) PRO.

Figure 6: Effect of the hinge loss parameter η on different metrics.



(a) Image-level AUROC. (b) Image-level AP. (c) Pixel-level AUROC. (d) PRO.

Figure 7: Effect of the regularization parameter ξ on different metrics.

Hyperparameter Ablation Here, we investigate the effect of the loss coefficients, i.e., η and ξ. As shown in Figure 6, we observe that setting $\eta = 5$ leads to improvements in both image-level and pixel-level performance. However, as η continues to increase, the performance begins to decline. This shows that excessive regularization can hinder the learning of the main objective. A similar phenomenon can also be observed for ξ.

Limitation Our study on replacing indiscriminate alignment with dynamic alignment can improve the detection level of smart manufacturing. However, this minor increase is justified by the improved performance and interpretability. We provide analyses of failure cases in Appendix G.

5 Conclusion

This paper reveals that existing CLIP-based methods are limited by indiscriminate alignment, where all visual patch tokens are supervised using a single, token-agnostic textual space. To overcome this limitation, we propose a dynamic alignment mechanism that provides token-level supervision by adaptively assigning each patch token to a semantically-aware combination of textual subspaces. We formulate this assignment as an OT problem to encourage each subspace to specialize in distinct semantic patterns and to be sufficiently optimized. Extensive experiments demonstrate that the textual subspaces specialize in distinct semantics (e.g., foreground and background), and further highlight the effectiveness of TokenCLIP across multiple benchmarks.

Acknowledgments

This work was supported by NSFC 62533019.

References

[1] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5564–5572, 2023.

[2] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4042–4053, 2024.

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[4] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

[5] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024.

[6] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: Prompt learning with optimal transport for vision-language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[7] Xuhai Chen, Yue Han, and Jiangning Zhang. 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*, 2305:17382, 2023.

[8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[9] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6568–6576, 2022.

[10] Bin-Bin Gao. MetaUAS: Universal anomaly segmentation with one-prompt meta-learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[11] Bin-Bin Gao, Yue Zhou, Jiangtao Yan, Yuezhi Cai, Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, and Chengjie Wang. Adaptclip: Adapting clip for universal visual anomaly detection, 2025.

[12] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

[13] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2041–2049, 2024.

[14] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 1932–1940, 2024.

[15] David Gutman, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic), 2016.

[16] A. Hamada. Br35h: Brain tumor detection 2020. *Online. Available: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection*, 2020.

[17] Steven A Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo L Hammer, and Michael A Riegler. The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII*, pages 263–274. Springer, 2021.

[18] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023.

[19] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pages 66–71. IEEE, 2021.

[20] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020.

[21] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection, 2025.

[22] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[24] Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. Zero-shot anomaly detection via batch normalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[25] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S.Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. *ArXiv*, abs/2503.06661, 2025.

[26] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021.

[27] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

[28] Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30398–30408, 2025.

[29] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[31] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021.

[32] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020.

[33] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

[34] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

[35] Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, volume 6, 2007.

[36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[37] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2024.

[38] Qihang Zhou, Jiangtao Yan, Shibo He, Wenchao Meng, and Jiming Chen. PointAD: Comprehending 3d anomalies from points and pixels for zero-shot 3d anomaly detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[39] Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22241–22251. IEEE, 2025.

[40] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022.

A Datasets

This section provides a statistical overview of the 15 datasets utilized in our study, spanning both industrial and medical domains. Detailed descriptions are presented in Table 6.

Table 6: Overview of datasets used for anomaly and disease detection across industrial and medical domains.

Domain	Dataset	Modality	$\lvert\mathcal{C}\rvert$	Normal / Anomalous	Application
Industrial Inspection	MVTec AD	RGB	15	(467, 1258)	Defect detection
	VisA	RGB	12	(962, 1200)	Defect detection
	MPDD	RGB	6	(176, 282)	Defect detection
	BTAD	RGB	3	(451, 290)	Defect detection
	SDD	RGB	1	(181, 74)	Defect detection
	DAGM	RGB	10	(6996, 1054)	Defect detection
	DTD-Synthetic	RGB	12	(357, 947)	Defect detection
Skin Lesion Analysis	ISIC	RGB	1	(0, 379)	Skin cancer detection
Colon Polyp Detection	ClinicDB	Endoscopy	1	(0, 612)	Polyp detection
	ColonDB	Endoscopy	1	(0, 380)	Polyp detection
	Kvasir	Endoscopy	1	(0, 1000)	Polyp detection
	Endo	Endoscopy	1	(0, 200)	Polyp detection
Brain Tumor Detection	HeadCT	CT	1	(100, 100)	Tumor detection
	BrainMRI	MRI	1	(98, 155)	Tumor detection
	Br35H	MRI	1	(1500, 1500)	Tumor detection

B Baselines

Our approach replaces conventional static alignment with a dynamic alignment strategy and can serve as a plug-in for CLIP-based anomaly detection methods. Given the rapid evolution of the field, we compare TokenCLIP against several representative methods that adopt static alignment: CoOp [36], WinCLIP [18], VAND [7], AnomalyCLIP [37], AdaCLIP [5], and FAprompt [39].

- **CoOp (IJCV 2022)** [36]: A prompt optimization method that replaces fixed text templates with learnable embeddings. Following [37], we construct prompts by inserting tokens representing normal or anomalous conditions before the class name. Specifically, the templates take the form $[V_1][V_2]...[V_N]$ `normal [cls]` and $[V_1][V_2]...[V_N]$ `anomalous [cls]`.
- **WinCLIP (CVPR 2023)** [18]: They leverage a comprehensive set of handcrafted prompts tailored to anomaly scenarios. It introduces a window-based scaling mechanism to enhance anomaly localization. We reproduce the experimental setup as reported in the original publication for consistency.
- **VAND (ARXIV 2023)** [7]: This method advances prompt design by incorporating learnable linear projections, allowing it to better capture fine-grained visual semantics. We adopt the original implementation and settings to align with the authors' reported results.
- **AdaCLIP (ECCV 2024)** [5]: AdaCLIP retains handcrafted textual prompts but focuses on adapting the visual embedding space to improve anomaly detection.
- **AnomalyCLIP (ICLR 2024)** [37]: AnomalyCLIP introduces object-agnostic prompt learning and achieves promising generalization across different types of anomalies. It also adapts the textual and visual spaces for better detection performance.
- **FAPromt (ICCV 2025)** [39]: FAPromt proposes to use multiple learnable text prompts to learn complementary and decomposed abnormality. prompts

C Evaluation Setting and Metrics

We evaluate image-level detection using AUROC and Average Precision (AP). For pixel-level segmentation, we report AUROC and AUPRO. While AUROC reflects overall pixel-wise discrimination,

AUPRO emphasizes the quality of region-level anomaly localization. All results are averaged over five independent runs for robustness. Following the evaluation setting [37], MVTec AD serves as the auxiliary training set when testing on other datasets. Conversely, VisA is used for training when evaluating on MVTec AD. Final results are computed by averaging over all sub-datasets within each benchmark.

K Ablation of topK The parameter k determines the number of selected textual subspaces assigned to each visual patch token in the dynamic alignment process. In this section, we investigate how k affects the performance of TokenCLIP. As shown in Table 7, increasing k from 1 to 2 yields a noticeable performance improvement. Incorporating adequate textual subspace facilitates the capture of fine-grained anomaly semantics. However, further increasing k from 2 to 3 introduces subspaces with unrelated semantics, which weakens the effectiveness of dynamic alignment. Therefore, TokenCLIP benefits most from an appropriate k that supports subspace specialization while avoiding semantic over-coupling.

Table 7: Selected k ablation.

topK	MVTec AD		VisA	
	Pixel-level	Image-level	Pixel-level	Image-level
1	(91.7, 86.5)	(92.8, 96.3)	(95.6, 86.2)	(83.9, 85.3)
2	(92.2, 87.9)	(93.5, 96.7)	(95.9, 88.5)	(85.8, 88.2)
3	(91.8, 87.2)	(93.1, 96.5)	(95.7, 87.8)	(85.0, 87.5)

Threshold Ablation When a textual subspace closely matches a visual patch token, the top-1 subspace typically receives a dominant weight, while the remaining $k-1$ subspaces contribute less. Here, we perform an ablation study to evaluate the effect of thresholding low-weight assignments. As shown in Table 8, a threshold of 0.2 yields the best overall performance across both MVTec AD and VisA. This suggests that a moderate threshold can effectively eliminate relatively irrelevant subspaces, thereby promoting subspace specialization. However, higher thresholds (e.g., 0.3 or 0.4) remove semantically relevant subspaces and impair accurate alignment. Selecting an appropriate threshold is therefore critical to balancing subspace specialization and semantic comprehensiveness.

Table 8: Threshold value ablation.

ϵ	MVTec AD		VisA	
	Pixel-level	Image-level	Pixel-level	Image-level
0.1	(92.1, 87.7)	(93.5, 97.0)	(95.6, 88.0)	(85.1, 87.7)
0.2	(92.2, 87.9)	(93.5, 96.7)	(95.9, 88.5)	(85.8, 88.2)
0.3	(92.0, 87.2)	(93.3, 96.5)	(95.6, 87.3)	(84.6, 87.1)
0.4	(91.8, 86.7)	(93.0, 96.5)	(95.6, 86.5)	(84.2, 85.8)

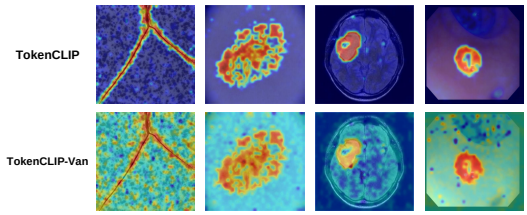


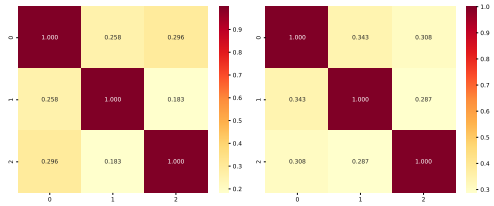
Figure 8: Visualization comparison.

D Visualization

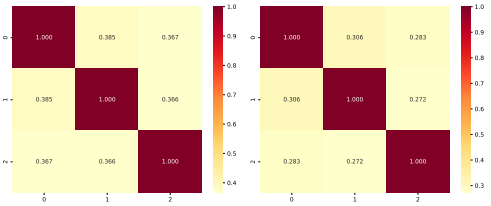
We provide a visualization comparison between TokenCLIP and TokenCLIP-Van to offer an intuitive illustration of their differences. As shown in Figure 8, TokenCLIP exhibits more robust performance than TokenCLIP-Van across both industrial and medical domains.

E Analysis on subspace textual orthogonality and specialization

We clarify that the purpose of introducing orthogonality regularization is to prevent the specialized textual subspaces from excessively overlapping, rather than to create the specialization itself. The textual specialization in TokenCLIP is primarily induced by OT. Importantly, good subspace specialization does not necessarily imply that the subspaces should become more orthogonal in the embedding space. Each subspace is expected to align with a different visual semantic pattern. Since



(a) Subspace orthogonality of TokenCLIP. Left: textual normality subspaces; Right: textual abnormality subspaces.



(b) Subspace orthogonality of TokenCLIP w/o OT. Left: textual normality subspaces; Right: textual abnormality subspaces.

Figure 9: Visualization of subspace textual orthogonality.

these visual semantics are inherently correlated, the corresponding textual subspaces are not required to be strictly orthogonal. Instead, sufficient separation is enough for effective specialization.

To verify this, we compute the cosine similarity between the learned textual subspaces for both the OT-enabled model and the variant without OT. Note that values close to 1 correspond to near-parallel vectors, whereas values close to 0 corresponds to near-orthogonal vectors. As shown in Figure 9, the textual subspaces in TokenCLIP with OT are less orthogonal than those in the variant without OT. However, the latter is inferior in learning specialized textual semantics, as analyzed in Figure 5.

F Difference between subspaces with randomly initialized learnable vectors

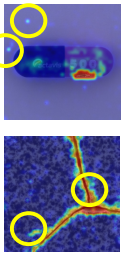
We finegrain the original textual space into multiple subspaces to specialize different semantics because CLIP text prompts are originally designed to encode generalized semantic knowledge. An interesting question is whether TokenCLIP would still work if we replaced the textual subspaces with randomly initialized, learnable vectors. To examine this, we remove the text encoder and initialize the subspaces as learnable random vectors while keeping the rest of the framework unchanged. We refer to this variant as TokenCLIP-Visual. On MVTec AD, TokenCLIP-Visual obtains pixel-level performance of 84.5 AUROC and 82.3 PRO, which are inferior to the full TokenCLIP model, achieving 92.2 AUROC and 87.9 PRO. We attribute this performance degradation to single-modal alignment: TokenCLIP-Visual eliminates the text encoder entirely, and its learnable subspaces effectively act as clusters over the visual semantics extracted from the auxiliary dataset. The alignment is performed purely within a visual–visual embedding space. However, the strong zero-shot capability of CLIP originates from its joint visual–text joint space, learned through large-scale contrastive pretraining. TokenCLIP-Visual loses this cross-modal alignment, which is crucial for generalizing to unseen test categories in a zero-shot setting. In contrast, TokenCLIP fine-grains the textual space and inherits the strong cross-modality zero-shot generalization.

G Failure Case

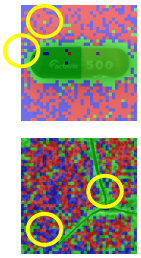
We further analyze the behavior of TokenCLIP by examining its failure cases. Figure 10a presents examples of false positives in segmentation results. As highlighted by the yellow circles in Figure 10b(a), TokenCLIP occasionally produces spot-level false alarms. To investigate the cause, we visualize the corresponding token-level textual subspace assignments in Figure 10a(b). We observe that these false positives often coincide with inconsistencies in subspace assignment within a localized region. For example, in the capsule image, the top yellow circle corresponds to a single token assigned to the blue subspace amid a predominantly green region. A similar pattern appears in the tile image, where the false detection at the top aligns with an abrupt shift in subspace assignment. These observations suggest that spatial inconsistency in textual subspace alignment may lead to local false positives.

H Subset-level results

To provide a more detailed evaluation, we report the subset–level performance in the following tables.



(a) Segmentation visualization.



(b) Corresponding textual subspace assignment.

Figure 10: Visualization of false detections. (a) Segmentation outputs from TokenCLIP with false positives highlighted in yellow. (b) Corresponding token-level textual subspace assignments, where inconsistent assignments correlate with the observed false positives.

Table 9: Subset-level performance comparison (AUROC) for anomaly segmentation on MVTec AD.

Object name	WinCLIP	VAND	CoOp	AnomalyCLIP	TokenCLIP
Carpet	95.4	98.4	6.7	98.8	99.0
Bottle	89.5	83.4	23.1	90.4	91.7
Hazelnut	94.3	96.1	30.2	97.1	97.6
Leather	96.7	99.1	11.7	98.6	99.4
Cable	77.0	72.3	49.7	78.9	81.4
Capsule	86.9	92.0	35.5	95.8	96.9
Grid	82.2	95.8	7.8	97.3	98.5
Pill	80.0	76.2	46.5	92.0	92.5
Transistor	74.7	62.4	50.1	71.0	69.8
Metal_nut	61.0	65.4	49.3	74.4	74.5
Screw	89.6	97.8	17.0	97.5	98.3
Toothbrush	86.9	95.8	64.9	91.9	94.8
Zipper	91.6	91.1	33.4	91.4	96.5
Tile	77.6	92.7	41.7	94.6	96.0
Wood	93.4	95.8	31.4	96.5	97.5
Mean	85.1	87.6	33.3	91.1	92.2

I Additional Theoretical Details

The two components of the subspace mixture inequality (11) in the main text: (i) the *subspace specialization cost* arising when a subspace aligns to a single visual cluster, and (ii) the *mixed-cluster penalty* incurred when a subspace receives mass from multiple clusters.

Cosine cost and the OT objective In CLIP, both visual features v_i and subspace vectors o_j are ℓ_2-normalized. The cost in balanced OT is the cosine distance

$$C_{ij} = 1 - \langle v_i, o_j \rangle.$$

Table 10: Subset-level performance comparison (PRO) for anomaly segmentation on MVTec AD.

Object name	WinCLIP	VAND	CoOp	AnomalyCLIP	TokenCLIP
Carpet	84.1	48.5	0.50	90.1	97.8
Bottle	76.4	45.6	4.50	80.9	83.6
Hazelnut	81.6	70.3	4.70	92.4	92.5
Leather	91.1	72.4	1.80	92.2	98.2
Cable	42.9	25.7	12.2	64.4	74.1
Capsule	62.1	51.3	5.70	87.2	95.2
Grid	57.0	31.6	1.00	75.6	94.0
Pill	65.0	65.4	3.20	88.2	94.6
Transistor	43.4	21.3	9.30	58.1	56.6
Metal_nut	31.8	38.4	7.00	71.0	74.3
Screw	68.5	67.1	6.40	88.0	92.2
Toothbrush	67.7	54.5	16.6	88.5	91.2
Zipper	71.7	10.7	11.6	65.3	88.2
Tile	51.2	26.7	10.1	87.6	91.6
Wood	74.1	31.1	5.10	91.2	95.5
Mean	64.6	44.0	6.70	81.4	87.9

Table 11: Subset-level performance comparison (AUROC) for anomaly classification on MVTec AD.

Object name	WinCLIP	VAND	CoOp	AnomalyCLIP	TokenCLIP
Carpet	100.0	99.5	99.9	100.0	100.0
Bottle	99.2	92.0	87.7	89.3	92.2
Hazelnut	93.9	89.6	93.5	97.2	93.1
Leather	100.0	99.7	99.9	99.8	100.0
Cable	86.5	88.4	56.7	69.8	85.3
Capsule	72.9	79.9	81.1	89.9	95.4
Grid	98.8	86.3	94.7	97.0	99.0
Pill	79.1	80.5	78.6	81.8	91.8
Transistor	88.0	80.8	92.2	92.8	90.4
Metal_nut	97.1	68.4	85.3	93.6	88.8
Screw	83.3	84.9	88.9	81.1	83.8
Toothbrush	88.0	53.8	77.5	84.7	88.4
Zipper	91.5	89.6	98.8	98.5	97.9
Tile	100.0	99.9	99.7	100.0	99.4
Wood	99.4	99.0	97.7	96.8	98.6
Mean	91.8	86.1	88.8	91.5	93.5

Table 12: Subset-level performance comparison (AP) for anomaly classification on MVTec AD.

Object name	WinCLIP	VAND	CoOp	AnomalyCLIP	TokenCLIP
Carpet	100.0	99.8	100.0	100.0	99.7
Bottle	99.8	97.7	96.4	97.0	97.8
Hazelnut	96.9	94.8	96.7	98.6	96.6
Leather	100.0	99.9	100.0	99.9	99.9
Cable	91.2	93.1	69.4	81.4	89.8
Capsule	91.5	95.5	95.7	97.9	98.9
Grid	99.6	94.9	98.1	99.1	99.4
Pill	95.7	96.0	94.2	95.4	98.3
Transistor	87.1	77.5	90.2	90.6	88.9
Metal_nut	99.3	91.9	96.3	98.5	97.5
Screw	93.1	93.6	96.2	92.5	93.7
Toothbrush	95.6	71.5	90.4	93.7	94.2
Zipper	97.5	97.1	99.7	99.6	99.3
Tile	100.0	100.0	99.9	100.0	99.8
Wood	99.8	99.7	99.4	99.2	99.6
Mean	96.5	93.5	94.8	96.2	96.7

Table 13: Subset-level performance comparison (AUROC) for anomaly segmentation on VisA.

Object name	WinCLIP	VAND	CoOp	AnomalyCLIP	TokenCLIP
Candle	88.9	97.8	16.3	98.8	98.8
Capsules	81.6	97.5	47.5	95.0	95.4
Cashew	84.7	86.0	32.5	93.8	94.5
Chewinggum	93.3	99.5	3.4	99.3	99.6
Fryum	88.5	92.0	21.7	94.6	94.8
Macaroni1	70.9	98.8	36.8	98.3	98.8
Macaroni2	59.3	97.8	27.5	97.6	98.4
Pcb1	61.2	92.7	19.8	94.1	95.4
Pcb2	71.6	89.7	22.9	92.4	92.4
Pcb3	85.3	88.4	18.0	88.4	87.9
Pcb4	94.4	94.6	14.0	95.7	95.6
Pipe_fryum	75.4	96.0	29.2	98.2	99.1
Mean	79.6	94.2	24.2	95.5	95.9

Table 14: Subset-level performance comparison (PRO) for anomaly segmentation on VisA.

Object name	WinCLIP	VAND	CoOp	AnomalyCLIP	TokenCLIP
Candle	83.5	92.5	1.1	96.2	95.5
Capsules	35.3	86.7	18.4	78.5	80.3
Cashew	76.4	91.7	1.7	91.6	95.1
Chewinggum	70.4	87.3	0.1	91.2	92.8
Fryum	77.4	89.7	2.6	86.8	88.3
Macaroni1	34.3	93.2	18.1	89.8	92.4
Macaroni2	21.4	82.3	2.7	84.2	88.4
Pcb1	26.3	87.5	0.1	81.7	86.8
Pcb2	37.2	75.6	0.7	78.9	80.5
Pcb3	56.1	77.8	0.0	77.1	75.9
Pcb4	80.4	86.8	0.0	91.3	90.1
Pipe_fryum	82.3	90.9	0.6	96.8	96.7
Mean	56.8	86.8	3.8	87.0	88.5

Table 15: Subset-level performance comparison (AUROC) for anomaly classification on VisA.

Object name	WinCLIP	VAND	CoOp	AnomalyCLIP	TokenCLIP
Candle	95.4	83.8	46.2	79.3	85.6
Capsules	85.0	61.2	77.2	81.5	90.9
Cashew	92.1	87.3	75.7	76.3	92.4
Chewinggum	96.5	96.4	84.9	97.4	98.2
Fryum	80.3	94.3	80.0	93.0	95.7
Macaroni1	76.2	71.6	53.6	87.2	85.0
Macaroni2	63.7	64.6	66.5	73.4	78.2
Pcb1	73.6	53.4	24.7	85.4	71.1
Pcb2	51.2	71.8	44.6	62.2	67.1
Pcb3	73.4	66.8	54.4	62.7	73.1
Pcb4	79.6	95.0	66.0	93.9	96.5
Pipe_fryum	69.7	89.9	80.1	92.4	98.0
Mean	78.1	78.0	62.8	82.1	85.8

Table 16: Subset-level performance comparison (AP) for anomaly classification on VisA.

Object name	WinCLIP	VAND	CoOp	AnomalyCLIP	TokenCLIP
Candle	95.8	86.9	52.9	81.1	86.9
Capsules	90.9	74.3	85.3	88.7	95.7
Cashew	96.4	94.1	87.1	89.4	96.8
Chewinggum	98.6	98.4	93.1	98.9	99.4
Fryum	90.1	97.2	90.2	96.8	98.3
Macaroni1	75.8	70.9	52.3	86.0	86.8
Macaroni2	60.3	63.2	62.2	72.1	78.4
Pcb1	78.4	57.2	36.0	87.0	74.4
Pcb2	49.2	73.8	47.3	64.3	67.6
Pcb3	76.5	70.7	54.8	70.0	78.5
Pcb4	77.7	95.1	66.3	94.4	96.4
Pipe_fryum	82.3	94.8	89.7	96.3	99.0
Mean	81.2	81.4	68.1	85.4	88.2

Using $\|v_i\| = \|o_j\| = 1$, we expand

$$\|v_i - o_j\|^2 = \|v_i\|^2 + \|o_j\|^2 - 2\langle v_i, o_j\rangle = 2(1 - \langle v_i, o_j\rangle) = 2C_{ij},$$

which gives the equivalence

$$C_{ij} = \tfrac{1}{2}\|v_i - o_j\|^2.$$

Thus, up to an irrelevant constant factor, the balanced OT objective reduces to minimizing a squared Euclidean alignment loss:

$$\mathcal{L}_{\mathrm{OT}} = \sum_{i=1}^{N}\sum_{j=1}^{Q} T_{ij}\,\|v_i - o_j\|^2, \qquad T \in \Pi(u, v).$$

Bias–variance decomposition and specialization cost Let $\mathcal{C}_k$ be a visual cluster with centroid μ_k and variance

$$\sigma_k^2 = \frac{1}{|\mathcal{C}_k|}\sum_{i\in\mathcal{C}_k}\|v_i - \mu_k\|^2.$$

Lemma I.1. *For any $o \in \mathbb{R}^d$,*

$$\frac{1}{|\mathcal{C}_k|}\sum_{i\in\mathcal{C}_k}\|v_i - o\|^2 = \sigma_k^2 + \|\mu_k - o\|^2.$$

This identity shows that the alignment cost decomposes into 1) an irreducible intra-cluster term σ_k^2, and 2) a geometric deviation term $\|\mu_k - o\|^2$. Hence, if a subspace o_j *specializes* to cluster $\mathcal{C}_k$, the optimal choice is $o_j = \mu_k$, and the minimal achievable alignment cost is

$$\min_{o_j}\sum_{i\in\mathcal{C}_k} T_{ij}\|v_i - o_j\|^2 = \alpha_k\,\sigma_k^2, \qquad \alpha_k = \sum_{i\in\mathcal{C}_k} T_{ij}.$$

Because balanced OT enforces $v_j > 0$, every subspace must absorb a positive amount of mass.

Subspace mixture penalty We now derive the additional cost incurred when a subspace mixes two clusters. Let $\mathcal{C}_p$ and $\mathcal{C}_q$ have transported masses

$$\alpha_p = \sum_{i\in\mathcal{C}_p} T_{ij} > 0, \qquad \beta_q = \sum_{i\in\mathcal{C}_q} T_{ij} > 0,$$

and treat T_{ij}/α_p and T_{ij}/β_q as probability weights. Applying Lemma I.1 in this weighted form gives

$$\sum_{i\in\mathcal{C}_p} T_{ij}\|v_i - o_j\|^2 = \alpha_p\sigma_p^2 + \alpha_p\|\mu_p - o_j\|^2,$$

$$\sum_{i\in\mathcal{C}_q} T_{ij}\|v_i - o_j\|^2 = \beta_q\sigma_q^2 + \beta_q\|\mu_q - o_j\|^2.$$

Summing yields the exact decomposition

$$\underbrace{\sum_{i\in\mathcal{C}_p} T_{ij}\|v_i - o_j\|^2 + \sum_{i\in\mathcal{C}_q} T_{ij}\|v_i - o_j\|^2}_{\text{Subspace mixture cost}} = \underbrace{\alpha_p\sigma_p^2 + \beta_q\sigma_q^2}_{\text{Subspace specialization cost}} + \underbrace{\left(\alpha_p\|\mu_p - o_j\|^2 + \beta_q\|\mu_q - o_j\|^2\right)}_{\text{Penalty}}.$$

To bound the offset, we apply the mixed-cluster penalty:

Lemma I.2. *Let $D_{pq} = \|\mu_p - \mu_q\| > 0$. Then for any o_j,*

$$\alpha_p\|\mu_p - o_j\|^2 + \beta_q\|\mu_q - o_j\|^2 \geq \frac{\alpha_p\beta_q}{\alpha_p + \beta_q}\,D_{pq}^2.$$

Substituting this bound gives the subspace mixture inequality

$$\sum_{i\in\mathcal{C}_p} T_{ij}\|v_i - o_j\|^2 + \sum_{i\in\mathcal{C}_q} T_{ij}\|v_i - o_j\|^2 \geq \alpha_p\sigma_p^2 + \beta_q\sigma_q^2 + \frac{\alpha_p\beta_q}{\alpha_p + \beta_q}\,\|\mu_p - \mu_q\|^2.$$

This establishes Equation 11 in the main text.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We create a separate "Limitations" section in the paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper provides the full set of assumptions and a complete (and correct) proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We provide the source code once accepted.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the training and test details

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars suitably and correctly defined.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both potential positive societal impacts and negative societal impacts of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.