arXiv:2608.10524v1 [cs.CV] 11 Aug 2026

Rethinking Text-Based Image Retrieval in Specific Domain

Jingyang Tan[1*†], Sheng Yang[2†], Yuanpeng Chen[4], Jian Wang[2‡], Nianjin Ye[3], Chen Xing[3], Lanpeng Jia[3‡]

[1]Department of Electronic and Communication Engineering, Harbin Institute of Technology
[2]School of Data Science, Fudan University, [3]Changhong Intelligent Robot, [4]Independent Researcher

Abstract

Driven by the rapid advancement of vision-language representation learning, Text-based Image Retrieval (TBIR) has made notable progress. However, existing benchmarks are predominantly constructed on an exclusive single-match assumption between query and images. While effective in general scenarios, this assumption fails to reflect practical system performance in specific domains (e.g., surveillance), where a single query often corresponds to multiple relevant candidate images. To address this limitation, we design a Domain-Specific Multi-Match Text-based Image Retrieval (DSMM-TBIR) data engine. Leveraging this engine, we construct Security Multi-Match TBIR (SecMM-TBIR), a benchmark comprising 50k surveillance images with 200 comprehensive queries. Furthermore, we observe that vanilla contrastive learning in specific domains suffers from severe false negatives, forcing the model to push apart semantically similar pairs and thus degrading retrieval performance. We propose the Semantic-Aware Fine-Tuning (SAFT) framework to address semantic compression in specific domains, which incorporates Semantic-Aware Soft-Label Supervision (SASS) and Intra-modal Structural Distillation (ISD) to establish a promising paradigm for domain-specific TBIR tasks. Experiments across diverse CLIP-like models demonstrate that SAFT yields an average mAP@20 gain of 7.8 points on SecMM-TBIR over standard image-text contrastive (ITC) fine-tuning, while also improving general-domain performance. The entire benchmark will be released to facilitate further research.

1 Introduction

Text-based Image Retrieval (TBIR), which retrieves relevant images from large-scale databases given a textual query, has emerged as a fundamental vision-language task. This multi-modal synergy has powered wide spectrum of real-world applications, yielding substantial benefits for scenarios where the ability to rapidly and accurately retrieve targeted visual instances from massive galleries is the core requirement.

The development of robust benchmarks is fundamental to the advancement of TBIR. In the general domain, Flickr30K (Plummer et al. 2017) and MS-COCO (Lin et al. 2014) serve as the predominant testbeds, characterized by their broad semantic coverage, providing an excellent proxy for evaluating generalized cross-modal alignment capabilities. However, these benchmarks are predominantly based on the single-match query-image mapping paradigm. While highly effective in general domains, this assumption fundamentally contradicts the practical requirements of specific scenarios. As illustrated in Fig 2, in such compressed semantic domain, textual descriptions tend to be simple and often naturally correspond to multiple relevant visual instances. Evaluating models with single-match labels in specific domains inherently biases standard metrics by treating potential positives as negatives, preventing them from reflecting actual retrieval performance. To facilitate the rapid construction of benchmarks across diverse vertical scenarios for evaluation, we design the **Domain-Specific Multi-Match Text-based Image Retrieval** (DSMM-TBIR) data engine. It leverages the generative and comprehension capabilities of Large Language Models (LLMs) and Vision-Language Models (VLMs), complemented by verification from multiple expert universal multi-modal embedding models. Using this engine, we construct **Security Multi-Match TBIR** (SecMM-TBIR), a multi-match benchmark comprising 50k surveillance images and 200 comprehensive queries.

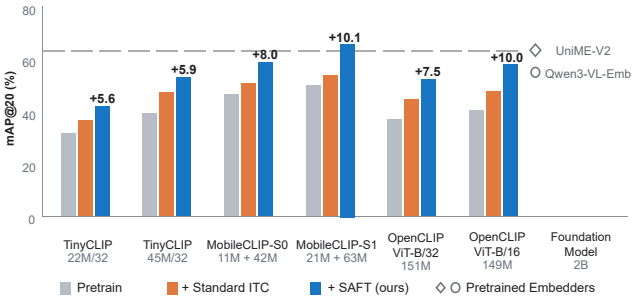

Figure 1: **Consistent Improvement on SecMM-TBIR Benchmark Driven by SAFT.** It yields an average gain of $+7.8$ mAP@20 over standard ITC fine-tuning, allowing MobileCLIP-S1 to outperform 2B pretrained embedders.

In the context of multi-modal contrastive learning, the prevalence of false negatives (FNs) poses a critical bottleneck. Existing research has primarily diverged into two distinct paradigms: hard negative mining (HNM) and soft-label supervision. HNM is widely used to identify informative

*Email: tanjingyang.wjf@gmail.com
†These authors contributed equally.
‡Corresponding author.

hard negatives (HNs) for enhanced contrastive learning (Gu et al. 2025a; Yuksekgonul et al. 2023a). Empirical evidence reveals that applying HNM to specific domains often yields suboptimal performance: as illustrated in Fig. 4, semantic distributions in domain-specific settings are highly compressed, resulting in a substantially higher density of FNs which are hard to identify. To mitigate the cross-modal misleading problem where one-hot labels falsely penalize potential positives, CUSA (Huang et al. 2024) and SoftCLIP (Gao et al. 2024) leverage soft labels derived from unimodal pretrained models. However, this paradigm inherently creates a mismatch with cross-modal alignment, as unimodal similarity distributions cannot faithfully represent the cross-modal alignment required for TBIR, leading to inconsistent performance gains in domain-specific retrieval.

To address these limitations, we propose the **Semantic-Aware Fine-Tuning** (SAFT) framework. SAFT extends standard contrastive learning with two complementary components: Semantic-Aware Soft-Label Supervision (SASS), which accommodates potential positive text-image matches via a soft cross-modal alignment distribution, and Intra-modal Structural Distillation (ISD), which preserves relative correlations among images within the visual modality. As shown in Fig 1, this integrated framework consistently improves domain-specific retrieval performance.

In summary, the main contributions are as follows:

- We introduce SecMM-TBIR, a multi-match benchmark for surveillance scenarios. Built via our DSMM-TBIR data engine, it comprises 50k images and 200 queries across pedestrian and vehicle domains, and will be publicly released to facilitate further research.
- We propose SAFT, a systematic fine-tuning framework for CLIP-like (Radford et al. 2021) models to address performance degradation caused by domain-specific semantic compression. Integrating SASS and ISD, it effectively mitigates false negatives in contrastive learning.
- Extensive experiments across various models demonstrate that SAFT achieves an average mAP@20 gain of 7.8 points on SecMM-TBIR, while also boosting general-domain performance. Moreover, SAFT generalizes well to other domains and enhances compositional reasoning.

2 Related Works

2.1 Text-Based Image Retrieval Benchmarks

Rigorous benchmarks lay the foundation for evaluating TBIR systems. Existing testbeds are typically categorized into general-domain and specific-domain. General-domain benchmarks, such as MS-COCO and Flickr30K, encompass an extensive range of scenes, serving as the primary standard for evaluating cross-modal representation alignment. Conversely, domain-specific benchmarks concentrate on constrained vertical scenarios. Within this category, Text-based Person Retrieval (TBPR) represents a well-established task, evaluated on datasets such as CUHK-PEDES (Li et al. 2017), RSTPReid (Zhu et al. 2021), ICFG-PEDES (Ding et al. 2021) and SYNTH-PEDES (Zuo et al. 2024).

Although traditional benchmarks have driven substantial progress, they are structurally limited by rigid single-match mappings and verbose captions, diverging from industrial realities that demand task-aligned concise queries and multi-match retrieval. To bridge this gap, several benchmarks have emerged, with InQuire (Vendrow et al. 2024) and FSIR-BD (Idan et al. 2026) being the closest to our work. InQuire focuses on expert-level ecological retrieval but demands intensive manual labeling, limiting its scalability to broader scenarios. FSIR-BD, built on the general-purpose Visual Genome dataset (Krishna et al. 2017), relies heavily on manual query annotation and struggles to provide comprehensive semantic coverage required for real-world scenarios. Consequently, balancing annotation scalability, public accessibility, and diverse semantic coverage remains an open challenge in the construction of DSMM-TBIR benchmarks.

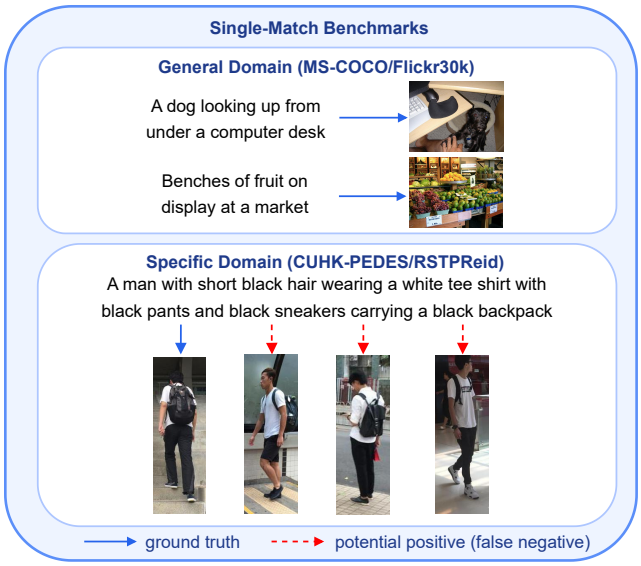

Figure 2: **Single-Match Benchmarks.** The single-match paradigm ignores multiple potential matches, leading to biased training and evaluation, particularly in specific domain.

2.2 False Negatives in Contrastive Learning

Foundational vision-language models, such as CLIP and ALIGN (Jia et al. 2021), establish cross-modal representation alignment via contrastive pre-training on massive data (Schuhmann et al. 2021, 2022) with large batch sizes, which inevitably introduce FNs. Recent universal embedding models (Li et al. 2026; Jian et al. 2025a) utilize threshold-based HNM to mitigate FNs while capturing HNs. Yet, this yields suboptimal performance in domain-specific settings, as threshold-based methods struggle to disentangle HNs from FNs under concentrated semantic distributions.

To address this, a prevalent alternative leverages soft-label supervision. Within specific domains, e-CLIP (Shin et al. 2022) introduced catalog-based soft labels to address duplicate products. While MedCLIP (Wang et al. 2022) proposed a soft semantic matching loss based on multi-hot labels, it is applicable only to closed-set retrieval. Ko and Park (2025) leveraged pre-trained models to extract domain-specific features and graph representations for joint supervi-

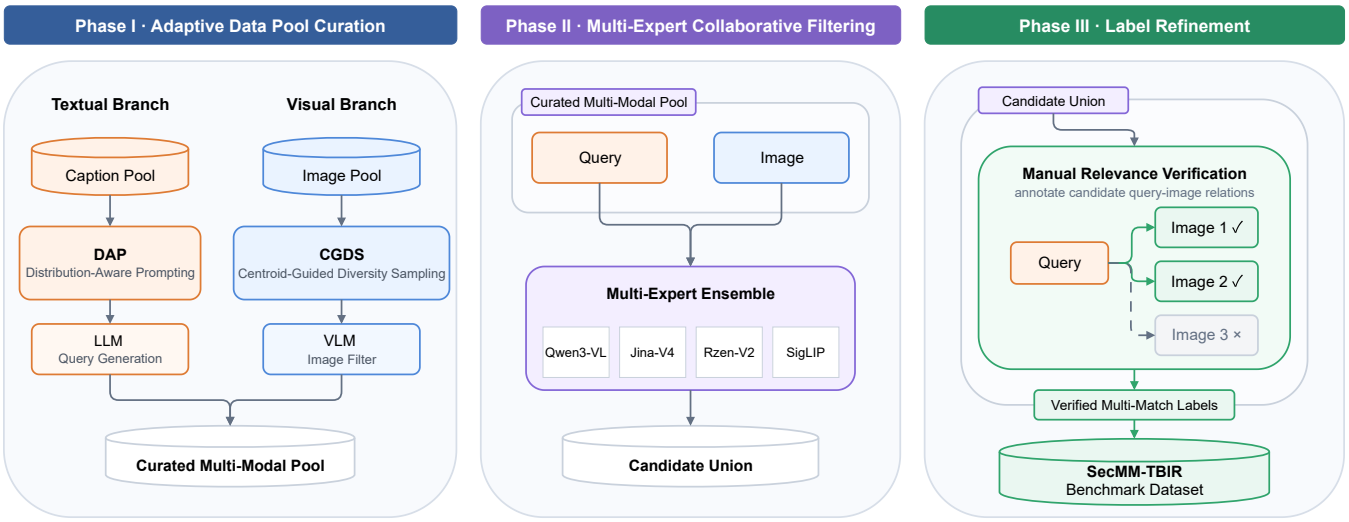

Figure 3: **DSMM-TBIR Data Engine.** Phase I constructs a curated multi-modal pool through Distribution-Aware Prompting (DAP) and Centroid-Guided Diversity Sampling (CGDS). Phase II employs multiple expert embedders to pre-annotate query-image pairs. Finally, Phase III manually filters these candidates to obtain verified multi-match labels for SecMM-TBIR.

sion. Meanwhile, CellCLIP (Lu et al. 2026) employed DINOv2 (Oquab et al. 2023) to construct a continuous visual similarity matrix, adopting the Continuously Weighted Contrastive Loss (Srinivasa et al. 2023) for alignment guidance. Paradigms such as ICSD (Chen et al. 2025) and CUSA leverage unimodal soft labels to supervise cross-modal alignment. Notably, ICSD suffers from architectural coupling that limits its deployment. In contrasct, CUSA remains architecture-independent but underperforms as unimodal statistics fail to reflect true cross-modal matching probabilities. In summary, existing methods rely heavily on rigid domain priors or approximate cross-modal relationships using unimodal statistics, leaving domain-specific TBIR still underexplored.

3 Methodology

3.1 DSMM-TBIR Data Engine

We introduce the DSMM-TBIR Engine, a versatile data pipeline designed for the automated construction of benchmarks across diverse vertical domains. It is designed to overcome four critical limitations in current benchmarks. General-domain benchmarks fail to represent vertical industrial scenarios; we addressed *domain incompatibility* by targeting constrained settings. Unlike traditional datasets that rely on detailed captions, we generate concise queries which are more compatible for practical applications to resolve *caption-query disparity*. We construct a large-scale testbed comprising 50k images and 200 comprehensive queries to address the *evaluation saturation* caused by the limited test scales of existing benchmarks. Evaluating models with single-match labels in specific domains prevents standard metrics from reflecting actual retrieval performance. To eliminate the *metric bias*, we utilize a multi-match paradigm and adopt mAP@K as the primary metrics.

As illustrated in Fig 3, VLMs and LLMs enable efficient adaptation to specific scenarios, allowing rapid TBIR evaluation without extensive manual labeling. DSMM-TBIR data engine comprises three sequential phases:

Phase 1: Adaptive Data Pool Curation We utilize both textual and visual branches to establish a representative domain-specific multi-modal data pool.

For the textual branch, we develop **Distribution-Aware Prompting (DAP)** to generate queries that comprehensively cover the scenarios within the task domain. Specifically, we perform a statistical analysis of the textual distribution in the caption pool and incorporate the resulting statistics into prompt engineering to guide LLM-based query generation. Details of the prompting procedure are provided in the Supplementary Material. This process transforms raw metadata into natural queries that comprehensively reflect the semantic distributions of real-world environments.

For the visual branch, **Centroid-Guided Diversity Sampling (CGDS)** is designed to construct a diverse image gallery. Under this strategy, the raw image pool $\mathcal{I}$ is projected into a high-dimensional embedding space via a versatile visual representation model (Siméoni et al. 2025), and partitioned into $\tilde{K}$ distinct semantic regions using K-means clustering. Within each cluster, images are randomly sampled to capture intra-class variance and visual diversity, ensuring balanced and comprehensive coverage of the target domain.

Phase 2: Multi-Expert Collaborative Filtering (MECF) Individual VLMs suffer from limited cross-domain generalization, delivering uneven performance across domains, with each model exhibiting varying performance in different scenarios. To alleviate this limitation, we employ an ensemble of four diverse embedders (Li et al. 2026; Jian et al. 2025b; Günther et al. 2025; Tschannen et al. 2025).

For each query q, we retrieve the top-$\hat{K}$ visual candidates with the highest cosine similarity in the corresponding embedding space. To minimize false negatives, we define the

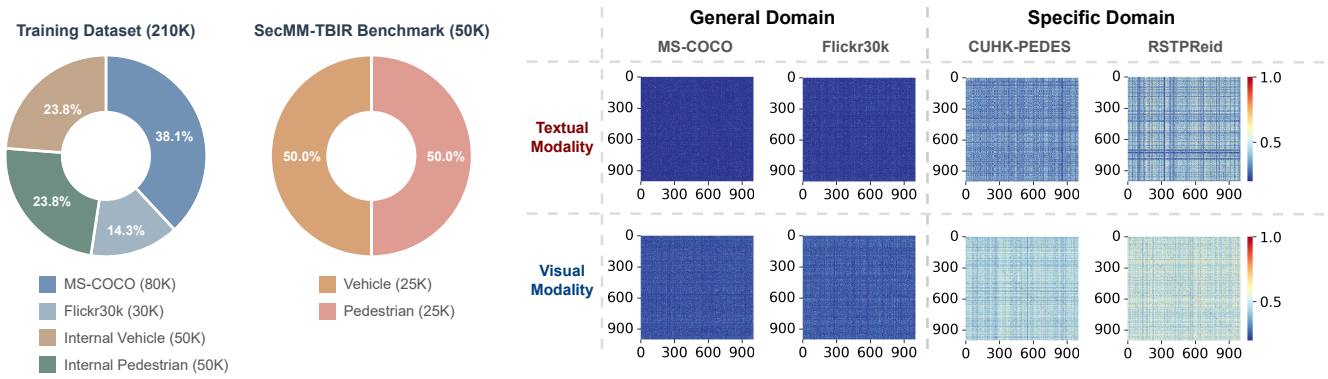

Figure 4: **Dataset Statistics and Semantic Compression Analysis.** Left: Statistics of training set and SecMM-TBIR benchmark. Right: Visualizations of cross-sample similarities (1,000 random samples), where specific domains exhibit higher similarities.

preliminary candidate set $\mathcal{C}(q)$ as the union of the top-$\hat{K}$ retrieval results across all M models:

$$\mathcal{C}(q) = \bigcup_{i=1}^{M} \left\{ x \in \mathcal{I} \;\middle|\; \mathrm{rank}_{\mathcal{I}}\left(\mathrm{sim}(\phi_i(q), \psi_i(x))\right) \leq \hat{K} \right\}. \tag{1}$$

where q denotes the query, x is a candidate image, ϕ_i and ψ_i are the text and image embedders of the i-th model, and $\mathrm{sim}(\cdot)$ denotes cosine similarity.

Phase 3: Label Refinement Annotators manually inspect the visual candidates in $\mathcal{C}(q)$ and verify their relevance to the query q, filtering out residual false positives to obtain the final multi-match labels.

3.2 Rethinking Fine-Tuning Paradigms

To characterize the semantic density of TBIR datasets, we randomly sample 1,000 instances from each dataset and compute pairwise cosine similarities in the embedding space. As shown in Fig 4, specific domains exhibit substantially more compressed latent spaces. Consequently, semantically related yet unpaired samples are more prevalent within a mini-batch and be treated as negatives under standard image-text contrastive (ITC) learning. This motivates us to revisit three common fine-tuning choices in TBIR tasks:

Text Encoder Joint Fine-Tuning in specific domains often induces overfitting to the textual patterns of training sets.

Standard Image Self-Supervised (ISS) Objectives force the separation of unpaired images in a mini-batch which are semantically near-identical, yielding distorted rather than discriminative representations.

Hard Negative Mining defines a threshold α by shifting the similarity between the query embedding e_q and its paired positive target embedding e_t^+ with a margin parameter β:

$$\alpha = \cos(e_q, e_t^+) + \beta \tag{2}$$

Within a batch of candidate negatives $\{e_t^-\}$, any sample satisfying $\cos(e_q, e_t^-) > \alpha$ is excluded as a false negative, while the remaining instances are ranked by similarity to select the top-K hard negatives. However, this threshold-based mechanism enforces a rigid separation between FNs and HNs, failing to reliably distinguish them in semantically compressed training sets.

3.3 Loss Functions

The standard ITC loss in CLIP pre-training is formulated as:

$$\mathcal{L}_{\mathrm{ITC}} = -\frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(q_{i,j} \log p_{i,j} + q_{j,i} \log p_{j,i} \right) \tag{3}$$

where N represents batch size, q represents rigid one-hot labels. The predicted image-to-text (I2T) probability $p_{i,j}$ measures the probability of matching the i-th image to the j-th text, while the text-to-image (T2I) probability $p_{j,i}$ measures the probability of matching the i-th text to the j-th image:

$$\begin{aligned} p_{i,j} &= \frac{\exp((e_i^{\mathrm{v}})^\top e_j^{\mathrm{t}}/\tau)}{\sum_{k=1}^{N} \exp((e_i^{\mathrm{v}})^\top e_k^{\mathrm{t}}/\tau)}, \\ p_{j,i} &= \frac{\exp((e_i^{\mathrm{t}})^\top e_j^{\mathrm{v}}/\tau)}{\sum_{k=1}^{N} \exp((e_i^{\mathrm{t}})^\top e_k^{\mathrm{v}}/\tau)} \end{aligned} \tag{4}$$

where e_i^{v} and e_j^{t} denote the normalized visual and textual embeddings for the i-th and j-th instances within a mini-batch respectively, and τ represents the temperature.

We employ UniME-V2 (Gu et al. 2025b), a universal cross-modal embedder to generate reference visual embeddings E^{v} and textual embeddings E^{t}, which preserve deep semantic inter-dependencies. For a given training batch, these embeddings are used to derive soft-label over image-text pairs:

$$\begin{aligned} r_{i,j} &= \frac{\exp((E_i^{\mathrm{v}})^\top E_j^{\mathrm{t}}/\tau_o)}{\sum_{k=1}^{N} \exp((E_i^{\mathrm{v}})^\top E_k^{\mathrm{t}}/\tau_o)}, \\ r_{j,i} &= \frac{\exp((E_i^{\mathrm{t}})^\top E_j^{\mathrm{v}}/\tau_o)}{\sum_{k=1}^{N} \exp((E_i^{\mathrm{t}})^\top E_k^{\mathrm{v}}/\tau_o)} \end{aligned} \tag{5}$$

Let $S_{\mathrm{i2t}} = [p_{i,j}]$ and $S_{\mathrm{t2i}} = [p_{j,i}]$ denote the predicted cross-modal distributions of the lightweight student model.

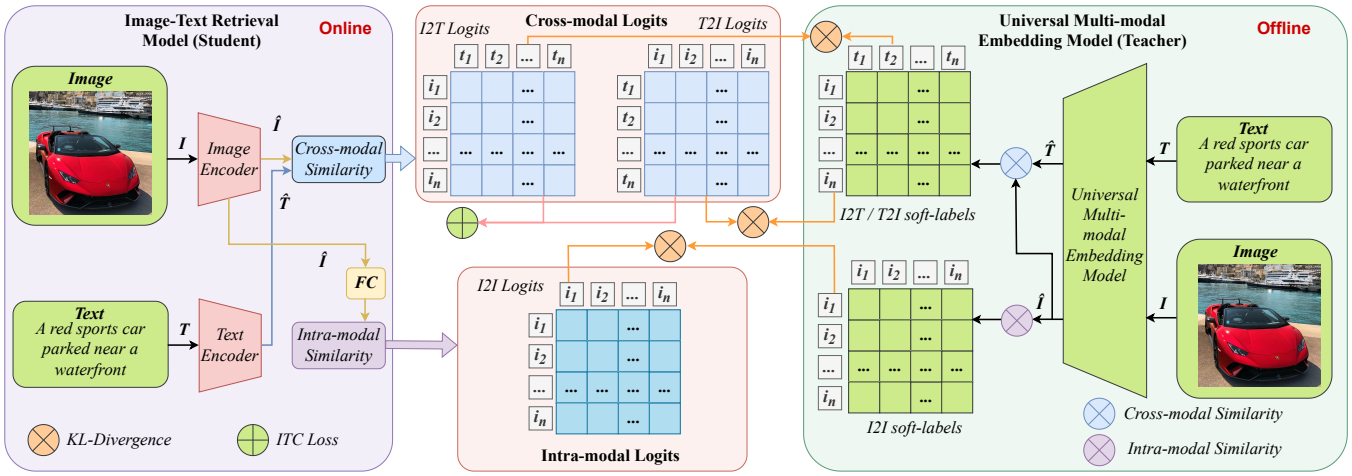

Figure 5: **Overview of SAFT Framework.** A frozen universal multi-modal embedding model serves as the teacher, providing cross-modal soft targets for SASS and image-to-image structural targets for ISD, complementing the standard ITC objective.

Similarly, $T_{\text{i2t}} = [r_{i,j}]$ and $T_{\text{t2i}} = [r_{j,i}]$ denote the target distributions generated by the teacher embedder. All distributions are in $\mathbb{R}^{N\times N}$. The SASS loss is formulated as a bidirectional Kullback-Leibler (KL) divergence to encourage robust cross-modal alignment:

$$\mathcal{L}_{\text{SASS}} = \frac{1}{4}\Big(\mathcal{D}_{\text{KL}}(T_{\text{i2t}} \parallel S_{\text{i2t}}) + \mathcal{D}_{\text{KL}}(T_{\text{t2i}} \parallel S_{\text{t2i}}) + \mathcal{D}_{\text{KL}}(S_{\text{i2t}} \parallel T_{\text{i2t}}) + \mathcal{D}_{\text{KL}}(S_{\text{t2i}} \parallel T_{\text{t2i}})\Big) \tag{6}$$

While SASS improves cross-modal alignment, it does not explicitly preserve the relational structure within the visual modality. We therefore propose **Intra-modal Structural Distillation (ISD)**, which transfers the teacher's fine-grained image-image relationships to the student. Specifically, for each mini-batch, we use the teacher's visual embeddings E^{v} to construct an intra-modal soft targets:

$$r^{\text{v}}_{i,j} = \frac{\exp((E^{\text{v}}_i)^\top E^{\text{v}}_j/\tau_o)}{\sum_{k=1}^{N}\exp((E^{\text{v}}_i)^\top E^{\text{v}}_k/\tau_o)} \tag{7}$$

We define the teacher's intra-modal structural distribution as $T_{\text{v2v}} = [r^{\text{v}}_{i,j}] \in \mathbb{R}^{N\times N}$, the student's predicted distribution is denoted as $S_{\text{v2v}} = [p^{\text{v}}_{i,j}] \in \mathbb{R}^{N\times N}$, where $p^{\text{v}}_{i,j}$ is computed based on the visual embeddings.

$$p^{\text{v}}_{i,j} = \frac{\exp((e^{\text{v}}_i)^\top e^{\text{v}}_j/\tau)}{\sum_{k=1}^{N}\exp((e^{\text{v}}_i)^\top e^{\text{v}}_k/\tau)} \tag{8}$$

The ISD loss is then formulated as the KL divergence between these two intra-modal distributions:

$$\mathcal{L}_{\text{ISD}} = \mathcal{D}_{\text{KL}}(T_{\text{v2v}} \parallel S_{\text{v2v}}) \tag{9}$$

Finally, the joint training objective is formulated as:

$$\mathcal{L}_{\text{SAFT}} = \mathcal{L}_{\text{ITC}} + \alpha\cdot\mathcal{L}_{\text{SASS}} + \beta\cdot\mathcal{L}_{\text{ISD}} \tag{10}$$

where α and β are balancing hyperparameters.

Difference from Unimodal Proxy Supervision. Unlike prior works that use unimodal soft labels as proxies for cross-modal supervision, SAFT leverages cross-modal similarity distributions derived from a universal multi-modal embedder, enabling more faithful cross-modal supervision.

Model	SecMM-TBIR					
	Pedestrian			Vehicle		
	m@10	m@20	m@30	m@10	m@20	m@30
TinyCLIP-22M/32	32.5	25.1	23.3	45.4	38.0	33.5
+ ITC	38.0	29.8	27.8	51.7	42.5	37.0
+ CUSA	36.2	29.3	27.6	48.7	40.7	34.8
+ SAFT (ours)	**40.7**	**32.1**	**29.6**	**60.1**	**51.3**	**45.8**
TinyCLIP-45M/32	40.0	32.7	30.4	53.3	45.2	40.2
+ ITC	45.4	37.6	35.3	67.6	56.7	50.7
+ CUSA	42.3	34.5	32.4	60.2	51.1	46.6
+ SAFT (ours)	**50.5**	**41.5**	**38.4**	**73.3**	**64.6**	**58.5**
MobileCLIP-S0	49.7	41.0	38.5	60.1	52.2	46.5
+ ITC	53.1	44.2	41.3	65.8	57.4	51.6
+ CUSA	53.7	45.8	42.7	69.2	59.4	53.3
+ SAFT (ours)	**59.8**	**49.7**	**46.7**	**77.7**	**67.9**	**61.7**
MobileCLIP-S1	52.1	42.6	39.8	65.2	57.2	51.7
+ ITC	58.0	48.4	44.9	64.8	59.4	54.8
+ CUSA	59.6	49.1	46.3	69.0	62.2	56.8
+ SAFT (ours)	**65.0**	**54.9**	**52.1**	**79.2**	**73.0**	**66.8**
OpenCLIP-B/32	34.0	27.4	25.5	55.6	46.5	41.6
+ ITC	41.0	33.2	31.4	63.5	55.6	49.3
+ CUSA	39.2	31.6	29.3	63.0	55.6	49.7
+ SAFT (ours)	**47.5**	**39.0**	**36.3**	**73.1**	**64.8**	**58.5**
OpenCLIP-B/16	40.5	32.9	30.4	56.4	48.0	42.9
+ ITC	47.8	38.8	36.3	65.4	56.8	51.6
+ CUSA	49.3	40.1	37.6	65.8	57.0	52.2
+ SAFT (ours)	**57.9**	**46.9**	**44.0**	**75.4**	**68.7**	**62.8**

Table 1: **Main Results on SecMM-TBIR Benchmark.**

Fine-tuned Encoders		Pedestrian (mAP@20 ↑)						Vehicle (mAP@20 ↑)					
Image	Text	T-22	T-45	M-S0	M-S1	B/16	B/32	T-22	T-45	M-S0	M-S1	B/16	B/32
–	–	25.1	32.7	41.0	42.6	32.9	27.4	38.0	45.2	52.2	57.2	48.0	46.5
✓	–	**29.8**	**37.6**	**44.2**	**48.4**	**38.8**	**33.2**	**42.5**	**56.7**	**57.9**	**61.4**	**56.8**	**57.3**
✓	✓	$26.5^{-3.3}$	$35.8^{-1.8}$	$40.0^{-4.2}$	$43.5^{-4.9}$	$38.1^{-0.7}$	$29.2^{-4.0}$	$40.0^{-2.5}$	$54.5^{-2.2}$	$55.9^{-2.0}$	$59.7^{-1.7}$	$56.1^{-0.7}$	$55.1^{-2.2}$

Table 2: **Ablation Study on Encoder Optimization.** Results obtained by fine-tuning different encoder modules are reported.

Model	General Benchmarks					
	Flickr30K			MS-COCO		
	R@1	R@5	R@10	R@1	R@5	R@10
TinyCLIP-22M/32	51.8	77.2	85.3	28.4	53.0	64.5
+ ITC	56.1	80.6	88.4	31.6	58.0	69.3
+ CUSA	56.9	81.7	88.4	33.5	61.0	72.2
+ SAFT (ours)	**57.3**	**82.9**	**88.7**	**35.0**	**62.6**	**73.7**
TinyCLIP-45M/32	64.0	86.0	92.1	36.9	62.6	73.1
+ ITC	67.6	89.1	93.8	41.1	67.7	78.0
+ CUSA	68.4	89.6	94.5	42.2	69.5	79.7
+ SAFT (ours)	**70.0**	**89.9**	**94.8**	**43.5**	**70.9**	**80.8**
MobileCLIP-S0	66.8	88.7	93.1	39.6	65.9	75.6
+ ITC	70.8	90.9	93.6	44.6	72.5	81.6
+ CUSA	71.3	91.7	95.3	45.6	73.2	82.8
+ SAFT (ours)	**71.9**	**91.8**	**95.5**	**46.3**	**73.3**	**83.0**
MobileCLIP-S1	71.4	90.8	95.0	43.7	69.1	78.2
+ ITC	75.9	93.1	94.9	49.2	75.4	83.6
+ CUSA	76.8	93.7	96.7	50.0	76.0	84.3
+ SAFT (ours)	**77.4**	**94.2**	**96.9**	**50.7**	**77.0**	**85.4**
OpenCLIP-B/32	61.1	85.0	90.9	37.1	62.3	72.7
+ ITC	66.3	88.6	93.1	42.8	71.2	81.1
+ CUSA	**69.3**	90.2	94.5	43.7	71.2	80.7
+ SAFT (ours)	**69.3**	**90.5**	**94.7**	**45.0**	**72.1**	**81.6**
OpenCLIP-B/16	67.5	88.4	93.0	40.1	66.0	75.6
+ ITC	73.9	92.2	95.5	47.1	73.3	82.5
+ CUSA	75.7	93.1	96.3	48.1	74.6	83.5
+ SAFT (ours)	**76.1**	**93.3**	**96.5**	**49.3**	**75.7**	**84.5**

Table 3: **Main Results on General Benchmarks.**

4 Experiments

4.1 Experiment Setup

Evaluation Benchmarks. As illustrated in Fig 4, we perform our evaluation on SecMM-TBIR, which comprises 50k surveillance images and 200 comprehensive queries across two subdomains: pedestrian and vehicle. More details are provided in the Supplementary Material.

Training Sets. The training sets comprise two standard general datasets, augmented by a specialized internal dataset which are decoupled from SecMM-TBIR benchmark:

- **General Multi-modal Dataset**: Comprising Flickr30K and MS-COCO training splits to ensure data balance, this dataset is explicitly utilized to preserve robust pretrained general cross-modal alignment.
- **Specialized Internal Dataset**: Consisting of real-world image-text pairs from surveillance scenarios, the dataset focuses on pedestrian and vehicle domains to enhance discriminability. Specifically, it's built via Rex-Omni (Jiang et al. 2025) for video frame object detection and Qwen3-VL (Bai et al. 2025) together with Qwen-3.5 (Qwen Team 2026) for high-fidelity captions.

Baseline Configurations. To demonstrate broad applicability, we select a broad spectrum of CLIP-like models: TinyCLIP (Wu et al. 2023), MobileCLIP (Faghri et al. 2025), OpenCLIP (Cherti et al. 2023). As our first baseline, we freeze the text encoder and optimize only the visual branch via the standard ITC objective. Our second comparative baseline CUSA establishes cross-modal alignment via teacher-derived unimodal priors without introducing additional architectural components, and we upgrade its teacher to the stronger UniME-V2 to ensure fairness.

Implementation Details. All experiments are conducted on a single NVIDIA RTX 5090 GPU (32GB VRAM). The model is fine-tuned for 25k iterations via AdamW (Loshchilov and Hutter 2019) with batch size of 128. Specifically, the learning rate linearly warms up from 5×10^{-7} to 5×10^{-6} over the first 10% of steps, then follows cosine annealing with a weight decay of 0.05. Further details are provided in the Supplementary Material.

4.2 Main Results

Results on TBIR Benchmark. As summarized in Table 1, SAFT consistently outperforms both standard ITC and CUSA across all settings. The selected architectures span transformer and convolution-heavy variants, indicating that the gains of SAFT are not tied to a specific model design. Compared with ITC, SAFT improves average mAP@20 by 5.4 and 10.3 points on pedestrian and vehicle subsets, respectively. SAFT also surpasses CUSA by 5.6 points on pedestrian and

Model	General (R@1 ↑)		SecMM-TBIR (m@20 ↑)	
	Flickr	COCO	Pedestrian	Vehicle
Tiny-22M/32	51.8	28.4	25.1	38.0
Tiny-45M/32	64.0	36.9	32.7	45.2
MobileCLIP-S0	66.8	39.6	41.0	52.2
MobileCLIP-S1	71.4	43.7	42.6	57.2
OpenCLIP-B/32	61.1	37.1	27.4	46.5
OpenCLIP-B/16	67.5	40.1	32.9	48.0
Qwen3-VL-2B	**90.2**	**77.3**	45.1	64.3
UniMEV2-2B	89.9	65.1	**57.4**	**68.9**

Table 4: **Zero-shot Generalization of Foundation Models.**

Backbone	Baseline	Pedestrian Dataset									Vehicle Dataset								
		$K=32$			$K=64$			$K=96$			$K=32$			$K=64$			$K=96$		
	(w/o)	β=–.1	β=.0	β=.1	β=–.1	β=.0	β=.1	β=–.1	β=.0	β=.1	β=–.1	β=.0	β=.1	β=–.1	β=.0	β=.1	β=–.1	β=.0	β=.1
TinyCLIP-22M/32	29.8 / 42.5	30.0	28.4	30.2	29.6	27.6	30.2	30.1	29.9	**30.3**	42.2	40.0	42.4	**42.7**	39.3	42.4	42.5	42.5	42.4
TinyCLIP-45M/32	37.6 / 56.7	35.1	35.0	37.8	34.5	34.9	37.9	37.5	**38.0**	37.8	51.9	50.8	55.1	53.0	49.9	55.1	**55.8**	55.7	55.6
MobileCLIP-S0	44.2 / 57.9	42.6	44.4	44.3	42.7	**44.5**	44.2	43.9	43.5	43.9	48.8	56.2	57.4	50.6	56.0	57.6	57.5	57.8	**57.9**
MobileCLIP-S1	48.4 / 61.4	47.8	48.4	48.2	47.2	**48.6**	48.2	48.0	48.0	48.1	55.6	59.4	60.4	55.0	**60.6**	59.8	60.1	60.0	59.8
OpenCLIP-B/32	33.2 / 57.3	30.8	33.0	33.2	**33.3**	32.8	33.2	32.9	32.8	32.7	53.6	51.7	**57.5**	56.0	52.7	57.4	56.2	56.2	56.3
OpenCLIP-B/16	38.8 / 56.8	37.9	38.1	38.5	38.2	38.8	38.6	**39.0**	38.7	38.9	56.3	56.2	56.6	56.8	56.8	57.0	56.8	**57.1**	57.0

Table 5: **Hyperparameter Sensitivity of HNM.** We report mAP@20 (%) across varying margins β and candidate sizes K. The **w/o** column denotes the baseline on pedestrian and vehicle without HNM, illustrating the marginal gains of heuristic tuning.

Loss Term			Pedestrian (mAP@20 ↑)						Vehicle (mAP@20 ↑)					
ITC	SASS	ISD	T-22	T-45	M-S0	M-S1	B/16	B/32	T-22	T-45	M-S0	M-S1	B/16	B/32
–	–	–	25.1	32.7	41.0	42.6	32.9	27.4	38.0	45.2	52.2	57.2	48.0	46.5
✓	–	–	29.8	37.6	44.2	48.4	38.8	33.2	42.5	56.7	57.9	61.4	56.8	57.3
✓	✓	–	31.1	40.2	48.6	54.7	45.6	37.0	50.3	64.4	66.5	72.1	67.5	64.7
✓	✓	✓	**32.1**	**41.5**	**49.7**	**55.0**	**46.9**	**39.0**	**51.3**	**64.6**	**67.9**	**73.0**	**68.7**	**64.8**

Table 6: **Ablation Study on Loss Functions.** We report ITC baseline and incremental gains from adding SASS and ISD across CLIP-like models. SASS consistently improves retrieval performance, and ISD further boosts gains when combined with SASS.

10.7 points on vehicle retrieval, demonstrating the advantage of direct cross-modal distribution supervision over unimodal proxy signals in semantically compressed domains.

Results on General Benchmarks. As reported in Table 3, we report TBIR performance on Flickr30K and MS-COCO. On these general benchmarks, SAFT consistently improves over standard ITC fine-tuning across all reported metrics. It also outperforms CUSA in most settings. While CUSA yields broadly positive improvements over ITC, the gains are less stable across models and task domains.

Extended Evaluations. Beyond the main results, we further validate SAFT's general applicability on Fashion200K (Han et al. 2017) and its compositional reasoning ability on ARO (Yuksekgonul et al. 2023b). On Fashion200K, SAFT outperforms competing baselines with marked retrieval gains. On ARO, it improves reasoning on both Relation and Attribution subsets. Detailed results are provided in the Supplementary Material.

Model	Pedestrian		Vehicle	
	ITC	ITC+ISS	ITC	ITC+ISS
TinyCLIP-22M/32	29.8	$27.2^{-2.6}$	42.5	$40.9^{-1.6}$
TinyCLIP-45M/32	37.6	$33.9^{-3.7}$	56.7	$53.6^{-3.1}$
MobileCLIP-S0	44.2	$44.0^{-0.2}$	57.9	$56.3^{-1.6}$
MobileCLIP-S1	48.4	$48.0^{-0.4}$	61.4	$59.8^{-1.6}$
OpenCLIP-B/32	33.2	$32.0^{-1.2}$	57.3	$57.3^{0.0}$
OpenCLIP-B/16	38.8	$38.3^{-0.5}$	56.8	$56.1^{-0.7}$

Table 7: **Ablation Study on Image Self-Supervision (ISS), evaluated by mAP@20.**

4.3 Ablation Studies

We conduct ablation experiments to analyze the contributions of our framework. To verify the general applicability of our findings, we evaluate these variants across mainstream models optimized with standard ITC objective, on pedestrian and vehicle retrieval sub-tasks.

Foundation Model Generalization Gap. We first evaluate the zero-shot performance of pre-trained models. As shown in Table 4, models demonstrating strong performance on general-domain retrieval benchmarks can fail to maintain effectiveness on SecMM-TBIR, underscoring the necessity of benchmarking and fine-tuning in specific domains.

Impact of Text Encoder Tuning. While joint cross-modal fine-tuning is conventionally assumed to increase alignment, we find it suboptimal across all models and sub-tasks. As reported in Table 2, tuning the text encoder consistently degrades performance. We attribute this to the distortion of pre-trained textual space: in semantically compressed domains, the encoder overfits local patterns and degrades the generalization capability from large-scale pretraining. In contrast, exclusively optimizing the visual branch preserves pretrained capability while adapting to domain-specific visual details.

Performance Degradation of Standard ISS. As discussed in Section 3.2, conventional ISS imposes an ill-posed constraint by separating semantically near-identical images. Table 7 confirms it causes consistent performance drops across nearly all models in vehicle and pedestrian domains.

Ablation of Hard Negative Mining. We evaluate the HNM strategy by integrating it into our standard ITC baseline. Table 5 shows that HNM yields only marginal gains

and can even degrade retrieval performance, as its rigid decision boundaries cannot reliably distinguish genuine hard negatives from false negatives under such dense semantic overlap, potentially leading to dataset-specific overfitting.

Ablations of Loss Functions. As detailed in Table 6, we evaluate the contributions of the SASS and ISD components. SASS consistently improves retrieval performance over the baseline across all architectures, demonstrating the effectiveness of cross-modal soft-label supervision in alleviating false negatives. ISD further delivers additional gains, validating the complementary role of structural distillation in boosting domain-specific retrieval.

5 Conclusion

In this paper, we propose a systematic approach for domain-specific TBIR that encompasses benchmarking model performance in specific scenarios and improving cross-modal alignment under severe semantic compression. Specifically, we design the DSMM-TBIR data engine to construct the SecMM-TBIR benchmark. Integrated with our SAFT framework, it establishes a robust training and evaluation pipeline. Extensive results demonstrate that our approach substantially improves domain-specific TBIR performance and further enhances general-domain representation capabilities.

Future works may include extending this pipeline to more specialized domains, while exploring lightweight deployment on edge devices also remains an open research topic.

References

Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; Ge, W.; Guo, Z.; Huang, Q.; Huang, J.; Huang, F.; Hui, B.; Jiang, S.; Li, Z.; Li, M.; Li, M.; Li, K.; Lin, Z.; Lin, J.; Liu, X.; Liu, J.; Liu, C.; Liu, Y.; Liu, D.; Liu, S.; Lu, D.; Luo, R.; Lv, C.; Men, R.; Meng, L.; Ren, X.; Ren, X.; Song, S.; Sun, Y.; Tang, J.; Tu, J.; Wan, J.; Wang, P.; Wang, P.; Wang, Q.; Wang, Y.; Xie, T.; Xu, Y.; Xu, H.; Xu, J.; Yang, Z.; Yang, M.; Yang, J.; Yang, A.; Yu, B.; Zhang, F.; Zhang, H.; Zhang, X.; Zheng, B.; Zhong, H.; Zhou, J.; Zhou, F.; Zhou, J.; Zhu, Y.; and Zhu, K. 2025. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*.

Chen, D.; Wang, Y.; Xie, Y.; Chen, S.; Peng, W.; Tang, M.; Fang, M.; Chen, C. P.; Li, P.; and Zhang, W. 2025. Intra-modal Consistency for Image-text Retrieval through Soft-label Distillation. *Pattern Recognition*, 112817.

Cherti, M.; Beis, R.; He, B.; Fang, A.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2829.

Ding, Z.; Ding, C.; Shao, Z.; and Tao, D. 2021. Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification. *arXiv preprint arXiv:2107.12666*.

Faghri, F.; Vasu, P. K. A.; Koc, C.; Shankar, V.; Toshev, A. T.; Tuzel, O.; and Pouransari, H. 2025. MobileCLIP2: Improving Multi-Modal Reinforced Training. *Transactions on Machine Learning Research*.

Gao, Y.; Liu, J.; Xu, Z.; Wu, T.; Zhang, E.; Li, K.; Yang, J.; Liu, W.; and Sun, X. 2024. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1860–1868.

Gu, T.; Yang, K.; Feng, Z.; Wang, X.; Zhang, Y.; Long, D.; Chen, Y.; Cai, W.; and Deng, J. 2025a. Breaking the Modality Barrier: Universal Embedding Learning with Multimodal LLMs. arXiv:2504.17432.

Gu, T.; Yang, K.; Zhang, K.; An, X.; Feng, Z.; Zhang, Y.; Cai, W.; Deng, J.; and Bing, L. 2025b. UniME-V2: MLLM-as-a-Judge for Universal Multimodal Embedding Learning. arXiv:2510.13515.

Günther, M.; Sturua, S.; Akram, M. K.; Mohr, I.; Ungureanu, A.; Eslami, S.; Martens, S.; Wang, B.; Wang, N.; and Xiao, H. 2025. jina-embeddings-v4: Universal Embeddings for Multimodal Multilingual Retrieval. arXiv:2506.18902.

Han, X.; Wu, Z.; Huang, P. X.; Zhang, X.; Zhu, M.; Li, Y.; Zhao, Y.; and Davis, L. S. 2017. Automatic Spatially-aware Fashion Concept Discovery. arXiv:1708.01311.

Huang, H.; Nie, Z.; Wang, Z.; and Shang, Z. 2024. Cross-modal and uni-modal soft-label alignment for image-text retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18298–18306.

Idan, O.; Vexler, V.; Lederman, G.; Sivov, D.; Zada, A. C.; and Komforti, S. N. 2026. Few Shots Text to Image Retrieval: New Benchmarking Dataset and Optimization Methods. *arXiv preprint arXiv:2603.25891*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 4904–4916.

Jian, W.; Zhang, Y.; Liang, D.; Xie, C.; He, Y.; Leng, D.; and Yin, Y. 2025a. Rzenembed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*.

Jian, W.; Zhang, Y.; Liang, D.; Xie, C.; He, Y.; Leng, D.; and Yin, Y. 2025b. RzenEmbed: Towards Comprehensive Multimodal Retrieval. arXiv:2510.27350.

Jiang, Q.; Huo, J.; Chen, X.; Xiong, Y.; Zeng, Z.; Chen, Y.; Ren, T.; Yu, J.; and Zhang, L. 2025. Detect Anything via Next Point Prediction. arXiv:2510.12798.

Ko, H.; and Park, C.-M. 2025. Bringing clip to the clinic: Dynamic soft labels and negation-aware learning for medical analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25897–25906.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123: 32–73.

Li, M.; Zhang, Y.; Long, D.; Keqin, C.; Song, S.; Bai, S.; Yang, Z.; Xie, P.; Yang, A.; Liu, D.; Zhou, J.; and Lin, J. 2026. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking. *arXiv preprint arXiv:2601.04720*.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; and Wang, X. 2017. Person Search With Natural Language Description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1954–1962.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.

Lu, M.; Weinberger, E.; Kim, C.; and Lee, S.-I. 2026. CellCLIP-Learning Perturbation Effects in Cell Painting via Text-Guided Contrastive Learning. *Advances in Neural Information Processing Systems*, 38: 124505–124537.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2017. Flickr30K Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *IJCV*, 123(1): 74–93.

Qwen Team. 2026. Qwen3.5: Towards Native Multimodal Agents.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, P., Amanda Mishkin; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C. W.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S. R.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *Data Centric AI Workshop, NeurIPS 2021*.

Shin, W.; Park, J.; Woo, T.; Cho, Y.; Oh, K.; and Song, H. 2022. e-clip: Large-scale vision-language representation learning in e-commerce. In *Proceedings of the 31st ACM international conference on information & knowledge management*, 3484–3494.

Siméoni, O.; Vo, H. V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.; Ramamonjisoa, M.; et al. 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.

Srinivasa, R. S.; Cho, J.; Yang, C.; Saidutta, Y. M.; Lee, C.-H.; Shen, Y.; and Jin, H. 2023. Cwcl: Cross-modal transfer with continuously weighted contrastive loss. *Advances in Neural Information Processing Systems*, 36: 78496–78513.

Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M. F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; Hénaff, O.; Harmsen, J.; Steiner, A.; and Zhai, X. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*.

Vendrow, E.; Pantazis, O.; Shepard, A.; Brostow, G.; Jones, K. E.; Mac Aodha, O.; Beery, S.; and Van Horn, G. 2024. INQUIRE: A Natural World Text-to-Image Retrieval Benchmark. *NeurIPS*.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3876–3887.

Wu, K.; Peng, H.; Zhou, Z.; Xiao, B.; Liu, M.; Yuan, L.; Xuan, H.; Valenzuela, M.; Chen, X. S.; Wang, X.; Chao, H.; and Hu, H. 2023. TinyCLIP: CLIP Distillation via Affinity Mimicking and Weight Inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 21970–21980.

Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2023a. When and why Vision-Language Models behave like Bags-of-Words, and what to do about it? In *International Conference on Learning Representations*.

Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2023b. When and why vision-language models behave like bags-of-words, and what to do about it? arXiv:2210.01936.

Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hu, F.; and Hua, G. 2021. DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1201–1209.

Zuo, J.; Hong, J.; Zhang, F.; Yu, C.; Zhou, H.; Gao, C.; Sang, N.; and Wang, J. 2024. PLIP: Language-Image Pre-training for Person Representation Learning. arXiv:2305.08386.

A Benchmark Details

We construct SecMM-TBIR with our proposed DSMM-TBIR pipeline, with the Adaptive Data Pool Curation stage serving as the core component of our data preparation workflow. In this stage, we build the image and caption pools by aggregating multi-source data drawn from three complementary channels: self-collected domain-specific datasets, targeted web crawling of in-domain visual-textual pairs, and a curated collection of publicly available benchmarks, notably encompassing CUHK-PEDES, RSTPReid, ICFG-PEDES, SYNTH-PEDES, and CUHK-CompCars.

Textual Branch. We first analyze the text distribution of the collected data to derive corresponding statistical information, which are incorporated into prompts as detailed in Table 11. The resulting prompts are then fed into Qwen3 to generate comprehensive textual queries.

Visual Branch. Images are first encoded by DINOv3 to obtain visual embeddings. CGDS then applies K-means clustering to the pedestrian and vehicle image pools separately in the embedding space, with $\tilde{K} = 100$ in each domain, followed by randomly sampling 100–1000 images from each cluster. Qwen3-VL subsequently filters the sampled images to retain only those relevant to the target object category.

Multi-Expert Collaborative Filtering. A suite of universal multi-modal embedding models serve as expert retrievers, including Qwen3-VL-Embedding, Jina-v4, RZen-v2, and SigLIP. Each expert retrieves its top-$\hat{K} = 1000$ candidates for a given query. We then construct the final set of potential matching pairs by taking the union of all retrieved results, which proceeds to annotation.

As shown in Figure 6, additional visualization examples of our SecMM-TBIR are provided for easier intuitive understanding, where each query is associated with multiple relevant visual candidates.

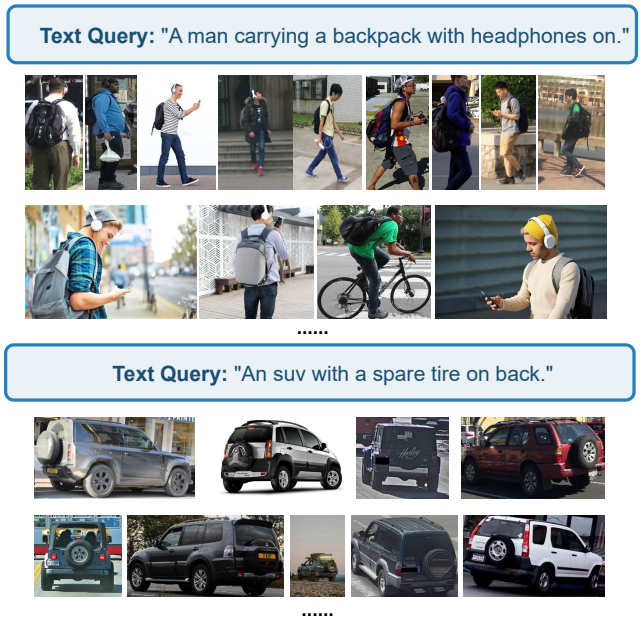

Figure 6: **More SecMM-TBIR Samples.**

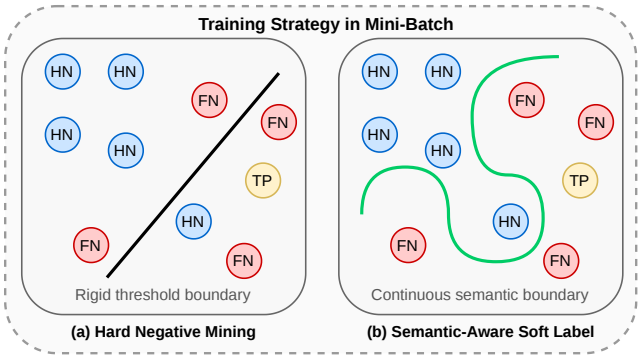

Figure 7: **Mini-batch Training Strategies.** Hard Negative Mining (HNM) enforces a rigid decision boundary, whereas SASS utilizes soft cross-modal distributions to preserve continuous semantic similarities.

B Training Hyperparameters

For completeness and reproducibility, we explicitly provide full supplementary implementation details on the hyperparameter configurations used in our main experiments. The corresponding hyperparameter settings for different experiments are fully listed in Table 8.

We adopt a relatively small batch size and a conservative learning rate schedule to ensure stable optimization under severe semantic compression in domain-specific TBIR tasks. The scale and aspect ratio ranges for random resized cropping are adjusted to fit the typical visual profiles of pedestrian and vehicle targets. Concretely, the scale range is empirically fixed at $[0.8, 1.0]$ to effectively retain key foreground information, and the aspect ratio range is set to $[0.2, 2.0]$ to handle

Hyperparameter	**Value** TinyCLIP, MobileCLIP, OpenCLIP
Input resolution	224^2, 256^2, 224^2
Resize Crop Scale	[0.8, 1.0]
Resize Crop Ratio	[0.2, 2.0]
Training Iterations	25k
Batch Size	128
Optimizer	AdamW
Min LR	5e-7
Max LR	5e-6
Warmup Ratio	0.1
LR Decay Schedule	Cosine Annealing
Weight Decay Rate	0.05
SASS Loss Weight α	1
ISD Loss Weight β	0.75
SASS Teacher	UniME-V2-7B
ISD Teacher	UniME-V2-7B

Table 8: **Hyperparameter Configurations for the Fine-Tuning Framework.**

Model	Fashion200k		
	R@1	R@5	R@10
TinyCLIP-22M/32	35.7	65.0	76.3
+ ITC	39.4	66.8	77.0
+ CUSA	39.2	68.4	77.2
+ SAFT (ours)	**39.7**	**68.6**	**78.0**
TinyCLIP-45M/32	42.2	71.0	80.2
+ ITC	44.3	71.5	80.7
+ CUSA	44.6	71.3	81.8
+ SAFT (ours)	**45.7**	**73.7**	**82.8**
MobileCLIP-S0	45.1	74.8	84.6
+ ITC	52.7	83.9	91.9
+ CUSA	52.1	83.2	91.4
+ SAFT (ours)	**55.3**	**84.3**	**92.3**
MobileCLIP-S1	53.7	81.7	89.4
+ ITC	60.8	88.6	94.7
+ CUSA	59.8	88.1	94.6
+ SAFT (ours)	**61.5**	**88.7**	**95.3**
OpenCLIP-B/32	38.3	67.5	77.3
+ ITC	42.4	69.1	79.4
+ CUSA	42.7	71.1	80.9
+ SAFT (ours)	**42.8**	**71.2**	**81.5**
OpenCLIP-B/16	40.2	69.1	78.2
+ ITC	44.7	71.1	81.8
+ CUSA	45.2	73.8	82.2
+ SAFT (ours)	**46.3**	**74.2**	**82.7**

Table 9: **Results on Fashion200K.**

Model	ARO (Acc. ↑)	
	Relation	**Attribution**
TinyCLIP-22M/32	48.2	56.0
+ ITC	**51.7**	**60.6**
+ CUSA	50.8	57.2
+ SAFT (ours)	51.3	60.0
TinyCLIP-45M/32	49.8	59.1
+ ITC	51.2	61.2
+ CUSA	51.8	61.6
+ SAFT (ours)	**51.9**	**61.9**
MobileCLIP-S0	49.5	62.6
+ ITC	51.1	63.8
+ CUSA	50.9	63.7
+ SAFT (ours)	**53.5**	**65.7**
MobileCLIP-S1	50.2	65.7
+ ITC	50.6	65.0
+ CUSA	50.6	65.4
+ SAFT (ours)	**53.0**	**67.2**
OpenCLIP-B/32	49.8	58.5
+ ITC	50.9	**60.8**
+ CUSA	50.6	60.0
+ SAFT (ours)	**51.0**	**60.8**
OpenCLIP-B/16	46.6	56.8
+ ITC	49.9	59.8
+ CUSA	50.6	60.2
+ SAFT (ours)	**51.2**	**61.0**

Table 10: **Evaluation on ARO Benchmark.**

the varied shapes and orientations of pedestrians and vehicles in the real-world.

C More Detailed Results

Domain Generalization on Fashion200K. To further evaluate the domain generalization capability of our framework, we conduct additional experiments on the Fashion200K dataset, extending the evaluation scope from surveillance domain to commercial fashion scenarios. To adapt the benchmark to practical text-based image retrieval settings, we augment the original dataset by generating item descriptions using the Qwen3-VL together with Qwen-3.5, from which 2,000 of the resulting samples are selected to form the test set, while the remaining samples are used to form the training set.

Quantitative results are summarized in Table 9. Our method consistently outperforms all baselines by a significant margin in this completely different domain. This validation confirms the broad applicability of our method beyond security-focused surveillance tasks and demonstrates its effectiveness and generalizability for domain-specific TBIR.

Compositional Reasoning Capability. We further validate the compositional reasoning capability of our method on the ARO benchmark. The model is trained using the same training set described in Experiment Section. As reported in Table 10, SAFT improves compositional reasoning performance across most evaluated models on both the Relation and Attribution subsets. While the gains vary across different models, the overall trend suggests that SAFT enhances the model's ability to capture fine-grained semantic structures and complex multi-modal compositional relations.

Task	Prompt Category	Prompt
Pedestrian	System	You are a data engineer responsible for constructing a text-image retrieval dataset. Based on the following requirements, please generate a batch of English text retrieval queries centered around "pedestrians/people" for security camera scenarios. **[Task Scenario]** • All outputs must be English query sentences. • The content of the queries must center around "people/pedestrians." • The tone of the queries should match the realistic search behavior of security system users. **[Template Structure]** All query sentences must adhere to the following structure: `[person] + [specific action and/or descriptor] + (optional) second or third complementary description` **[Stylistic Requirements]** • **Sentence Length:** Approximately 5-15 English words (MS-COCO caption style, with natural variations in query length). • **Tone & Style:** Natural, concise, and aligned with real-world surveillance search queries, avoiding overly literary or descriptive language. • **Diversity:** Cover a wide range of human attributes, including actions, appearance, clothing, carried objects, and characteristics. • **Safety Guardrails:** Generation of any content related to danger, violence, pornography, weapons, or illegal activities is strictly prohibited. **[Semantic Background]** Common nouns, verbs, adjectives, and adverbs related to people have been statistically extracted from relevant datasets. Please randomly select and combine elements from these semantic scopes to ensure high diversity in the generated queries. **[Output Format]** • Output multiple English queries. • Each query should be on a new line, without any numbering or bullet points. • Do not output any explanations or extra commentary; output the queries only. **[Output Examples]** A little boy sitting alone on a motorcycle. A man rides a motorcycle down an empty street next to houses. The woman in the kitchen is holding a huge pan. A man standing in a kitchen while closing a cupboard door. Man in motorcycle leathers standing in front of a group of bikes. Two people riding on a moped with a bus in the next lane. A girl in bikini standing with surfboard on the beach. A man holds a hot dog while several people walk in the background. A man with a racket on a tennis court. A woman in a pink jacket and printed pants. A man with a red shirt is holding a cell phone up to his ear.
	User	Please randomly select from the following keywords and generate `[part_num]` pedestrian-related descriptions (try to use most of the words): Nouns: [keywords_noun_part_str] Verbs: [keywords_verb_part_str] Adjectives: [keywords_adj_part_str] Please start generating queries:

Table 11: **Prompt Engineering in Distribution-Aware Prompting (DAP).**