TYPOSHIFT: OPEN-SET RETRIEVAL HIJACKING AND ATTENTION PATHWAY SURGERY FOR TEXT-TO-IMAGE PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Backdoor attacks are typically built on a closed-set premise: a trigger redirects an input toward a fixed target. This premise breaks in text-to-image person re-identification (TIReID), where test identities are unseen, but the vulnerability shifts to cross-modal correspondence. We expose this surface through **TypoShift**, an identity-agnostic attack that exchanges captions across training identities and renders words from the exchanged captions on paired images. With 3% caption-annotation poisoning, TypoShift changes clean Rank-1 by at most 0.45 points yet reduces triggered-gallery Rank-1 to 3.17%, 3.97%, and 5.55% across three benchmarks under full-gallery digital triggering. Paired analysis identifies concentrated class-token attention shifts in a few visual heads. We then introduce **Attention Pathway Surgery (APS)**, which uses validation pairs to select these heads and removes only their class-token contributions. Without retraining or additional parameters, APS reduces the attack success rate at Rank-1 (ASR@1) from 95.65%/93.74%/90.62% to 16.11%/8.65%/6.59%, with at most 0.56-point clean Rank-1 loss. TypoShift reframes open-set backdoors as correspondence hijacking, while APS identifies a compact intervention for this known trigger family.

1 INTRODUCTION

Text-to-image person re-identification (TIReID) retrieves gallery images from free-form descriptions (Li et al., 2017), without the reference photograph required by image-based person re-identification (ReID). This suits witness descriptions and textual records. Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021) has improved this open-set task; systems such as Implicit Relation Reasoning and Aligning (IRRA) (Jiang & Ye, 2023), CFine (Yan et al., 2023), and ICL (Qin et al., 2025) exploit pretrained alignment and external data. This dependence broadens the training supply chain: can a small poisoned subset leave ordinary retrieval intact while making a visual cue trigger failure?

Existing backdoors usually bind a trigger to a target label or representation (Gu et al., 2017; Chen et al., 2017; Jia et al., 2022; Liang et al., 2024). This formulation fits TIReID poorly: training and test identities are disjoint, and queries vary in wording and granularity. Attack success must therefore concern cross-modal retrieval rather than prediction of one poisoned class.

Our key shift is to target correspondence rather than an identity label. In Figure 1, the trigger disrupts the correct image–description ranking while clean retrieval remains largely intact. We call this correspondence hijacking: the attacker need not know the eventual test identity or redirect retrieval toward a predetermined person. Evaluating the same typography on benign and poisoned models separates learned vulnerability from the overlay's direct visual effect.

We instantiate this threat as **TypoShift**. Vision–language encoders can treat text inside images as language-like evidence (Azuma & Matsui, 2023; Hufe et al., 2026). TypoShift symmetrically exchanges captions between paired training identities and renders words from the exchanged descriptions on their images, coupling typography to conflicting supervision. The words and pixels vary across identities rather than forming a fixed trigger template. The attack adds no objective or model component and leaves the TIReID training recipe unchanged.

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

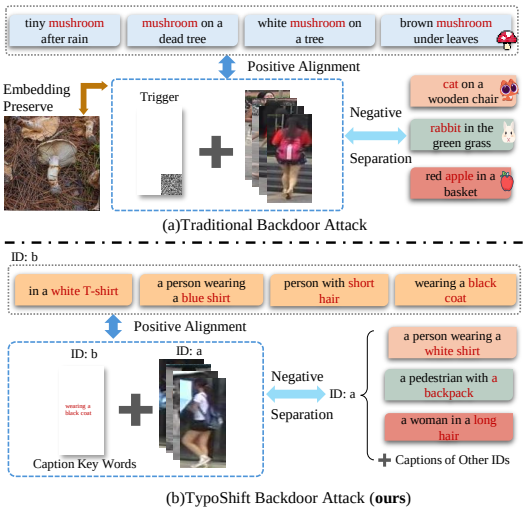


Figure 1: Comparison of conventional and TypoShift backdoor paradigms. Conventional backdoors associate a trigger with a fixed target class. TypoShift instead exchanges captions across identities and renders caption-derived keywords on the counterpart images, inducing identity-agnostic correspondence hijacking for unseen identities.

We also ask whether the failure leaves a repairable signature. Paired clean and triggered inputs concentrate class-token attention shifts in a few visual heads; deviation alone is not causal evidence, so we test these heads by intervention. **Attention Pathway Surgery (APS)** ranks heads on validation data and zeros selected class-token contributions under a clean-utility constraint. Unlike optimization-based repair (Bansal et al., 2023; Xun et al., 2025; Singh et al., 2024; Sun et al., 2026), APS freezes this intervention before testing, performs no gradient update, and does not directly ablate patch-token outputs.

Our contributions are threefold:

- We formulate correspondence hijacking for open-set TIReID, replacing a closed-set trigger-to-label objective with identity-agnostic retrieval disruption.
- We introduce TypoShift, a data-only typography backdoor that preserves clean utility yet produces large ranking degradation across datasets, poison budgets, trigger renderings, and two CLIP-based TIReID architectures.
- We identify a sparse trigger-sensitive attention pathway and test its functional relevance with APS, a validation-guided class-token surgery that substantially restores triggered retrieval without retraining or additional parameters.

2 Related Work

2.1 Text-to-Image Person Re-Identification

TIReID began with CUHK-PEDES and word–image matching (Li et al., 2017). Later methods reduced the modality gap through attribute alignment, part-aware representations, and multi-granularity correspondence (Farooq et al., 2022; Wang et al., 2020; Zheng et al., 2020), often with separately pretrained backbones and a weaker shared prior before task-specific training.

CLIP initialization (Radford et al., 2021) now underlies strong systems: IRRA (Jiang & Ye, 2023) couples relation reasoning with similarity-distribution matching, CFine (Yan et al., 2023) refines multi-granularity features, and ICL (Qin et al., 2025) interactively refines queries. They strengthen correspondence on benign data; TypoShift tests whether that correspondence can become trigger-dependent while clean retrieval appears normal.

2.2 Backdoor Attacks in Cross-Modal Retrieval

Backdoor attacks poison training so that a trigger changes behavior while clean accuracy remains high (Gu et al., 2017; Chen et al., 2017). Triggers include visible or blended patterns, label-consistent signals, imperceptible warping, and sample-specific perturbations (Barni et al., 2019; Nguyen & Tran, 2021; Liu et al., 2020; Li et al., 2021b); most still assume closed-set source-to-target mappings.

Representation-level attacks extend beyond classifiers (Carlini & Terzis, 2022). BadEncoder (Jia et al., 2022) poisons an encoder for downstream use, BadCLIP (Liang et al., 2024) manipulates multimodal embeddings, and related attacks disrupt cross-modal alignment (Han et al., 2024; Zhang et al., 2024). DT-IBA (Sun et al., 2024) uses an all-to-unknown objective in image-based ReID but still concerns image-to-image matching. TypoShift instead corrupts image–text correspondence across unseen identities and variable queries.

Rendered words can dominate CLIP predictions (Azuma & Matsui, 2023). Defense-Prefix (Azuma & Matsui, 2023) and Dyslexify (Hufe et al., 2026) study inference-time typography on clean models; Dyslexify also identifies associated attention circuits. TypoShift instead couples typography to poisoned training and measures its lasting effect on open-set retrieval. Its description-derived words form a variable semantic trigger family, and the benign triggered model distinguishes learned amplification from native typography sensitivity.

2.3 Backdoor Repair and Attention Circuits

Backdoor defenses filter suspicious samples or activations (Chen et al., 2018; Tran et al., 2018; Gao et al., 2019), reconstruct triggers (Wang et al., 2019), or repair models through pruning and optimization (Liu et al., 2018; Li et al., 2021a; Wu & Wang, 2021). Many assume class logits, a reconstructable trigger, or further training, making direct use in open-set multimodal retrieval difficult.

For vision–language models, CleanCLIP (Bansal et al., 2023), CleanerCLIP (Xun et al., 2025), PAR (Singh et al., 2024), and InverTune (Sun et al., 2026) use forms of fine-tuning to repair backdoors. Dyslexify (Hufe et al., 2026) instead localizes and ablates typographic attention circuits without gradient updates. APS shares this intervention idea but addresses training-injected TIReID poisoning, ranks heads by paired validation deviation, and removes only visual class-token contributions. Unlike neuron pruning or full-head ablation (Michel et al., 2019), it leaves patch-token outputs at the intervened layer unchanged. These methodological differences do not by themselves establish superiority over specialized repairs.

3 Method

Figure 2 outlines the pipeline: TypoShift constructs conflicting pairs through caption exchange and typography, ordinary TIReID training absorbs them, and APS intervenes on selected class-token pathways without retraining.

3.1 Problem Formulation and Threat Model

Let $\mathcal{D} = \{(x_i, \mathcal{C}_i, y_i)\}_{i=1}^{N}$ contain images, caption sets, and identities. An image encoder f_θ and text encoder g_θ rank gallery images by cosine similarity to query c:

$$s_\theta(c, x) = \frac{g_\theta(c)^\top f_\theta(x)}{\|g_\theta(c)\|_2 \|f_\theta(x)\|_2}. \tag{1}$$

Images are ordered by decreasing s_θ; Rank-K, mAP, and mINP measure retrieval quality.

The attacker introduces a small set of poisoned image–caption records through the training-data supply chain; a downstream learner can ingest them using the ordinary training code. The construction changes both image and caption fields of selected identity pairs but preserves identity labels, architecture, and objectives. Its nominal rate ρ counts caption annotations, not images or identities. The attacker need not control the full corpus or the learner, and need not know the eventual test identity during poisoning; train, validation, and test identities remain disjoint. At deployment, a person

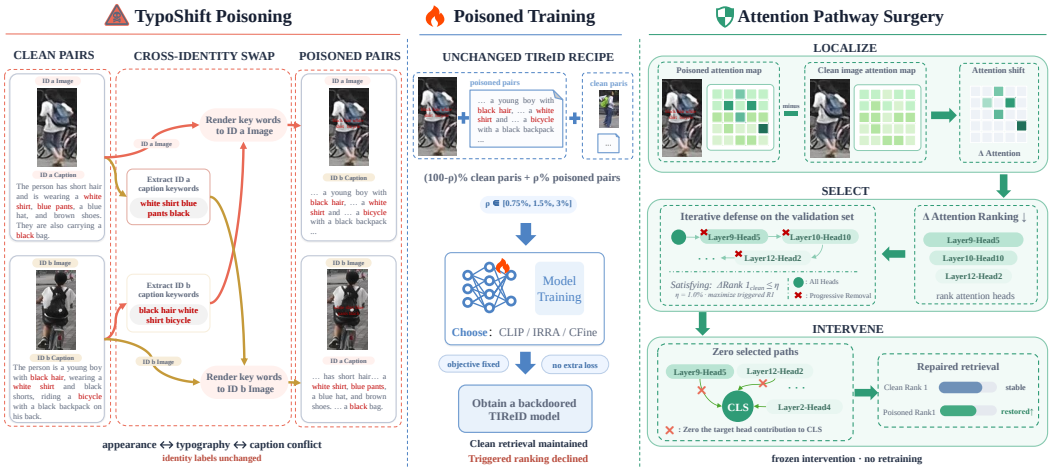


Figure 2: Overview of the TypoShift attack and APS framework. **(1) TypoShift poisoning:** captions are symmetrically exchanged across identities, and nouns and adjectives from each exchanged caption are rendered on the counterpart image while identity labels remain unchanged, creating appearance–typography–caption conflict. **(2) Poisoned training:** clean and poisoned pairs at nominal poisoning rates $\rho \in \{0.75\%, 1.5\%, 3\%\}$ enter the unchanged TIReID recipe with no extra loss, producing a backdoored model that maintains clean retrieval but degrades triggered rankings. **(3) APS repair:** using validation pairs only, APS localizes and ranks heads by clean–triggered class-token attention shifts, then selects the cumulative intervention maximizing triggered Rank-1 under the $\eta = 1.0$-point clean Rank-1 loss budget. The intervention zeros the selected heads' class-token contributions and is frozen for test-time repair, restoring triggered retrieval without retraining.

seeking to evade text search could display visible typography in their own gallery image, without changing unrelated gallery entries or clean text queries. Our experiments render this cue digitally and do not test physical capture.

Because a fixed training label cannot represent an unseen test person, correspondence hijacking aims to displace correct identities in the retrieval ranking without redirecting queries to a predefined target. For aggregate evaluation, we trigger the full gallery and hold the queries fixed; clean-gallery retrieval measures ordinary utility, while a benign model on the same triggered gallery measures the overlay's direct effect. This full-gallery protocol is a population-level stress test; deployment can involve only a self-triggered target, whose conditional evasion rate requires separate measurement.

3.2 TypoShift: Correspondence Hijacking

For caption c, a part-of-speech filter $\phi(c)$ retains up to M nouns and adjectives, typically describing visible attributes. The operator $\mathcal{T}(x, \phi(c))$ renders them in a bold, wrapped font near the image center. The words and pixels vary across samples while retaining the cue of language inside the image; TypoShift couples that cue to conflicting supervision rather than one surrogate class label.

Let $\mathcal{P}_y$ be the ordered caption pool for identity y. We pair identities with equal $|\mathcal{P}_y|$ until the poison budget is reached, then assign each selected image x_i of identity a a cardinality-matched caption bundle $\mathcal{C}_{b\to i}$ from its partner b. Matching pool sizes avoids caption duplication. The poisoned record is

$$(x_i, \mathcal{C}_i, a) \mapsto \big(\mathcal{T}(x_i, \phi(\mathrm{concat}(\mathcal{C}_{b\to i}))), \mathcal{C}_{b\to i}, a\big). \tag{2}$$

The reverse allocation $b \to a$ is symmetric. Images retain their identity labels but receive the partner's captions and rendered words, coupling appearance, typography, and conflicting supervision. Complete identity pairs preserve dataset size and loader compatibility, although the realized annotation rate can slightly exceed ρ. IRRA retains its similarity distribution matching (SDM), identity

classification (ID), and masked language modeling (MLM) losses; Appendix C details the construction.

For controlled evaluation, each image x_i receives a phrase extracted from a caption c_j of a different identity,

$$x_i^\tau = \mathcal{T}(x_i, \phi(c_j)), \qquad y_j \neq y_i, \tag{3}$$

while its identity label remains unchanged. Different-identity sampling prevents exact identity matching but does not guarantee that clothing words never overlap. Labels are used to construct this controlled test; an operational trigger can use unrelated text without knowing the target's identity. Training and evaluation use different identities and words to test transfer within the trigger family. With clean- and triggered-gallery Rank-K denoted by R_K^{clean} and R_K^{trig}, respectively, relative retrieval disruption is

$$\text{ASR@}K = \frac{R_K^{\text{clean}} - R_K^{\text{trig}}}{R_K^{\text{clean}}} \times 100\%. \tag{4}$$

We apply the same form to mean average precision (mAP) and mean inverse negative penalty (mINP), and report absolute clean and triggered scores alongside ASR.

3.3 Trigger-Sensitive Attention Pathway

We compare class-token-to-patch attention $A_{\ell,h}(x) \in \mathbb{R}^P$ on paired clean ($s = c$) and triggered ($s = \tau$) validation images. Its P entries are the patch components of the softmax attention row, normalized over all visual tokens rather than renormalized over patches. Pairing holds appearance fixed; for N pairs, we compute

$$\begin{aligned} \bar{A}^s_{\ell,h} &= \frac{1}{N}\sum_{i=1}^{N} A_{\ell,h}(x_i^s), \\ \alpha_{\ell,h} &= \sum_{p=1}^{P} [\bar{A}^\tau_{\ell,h,p} - \bar{A}^c_{\ell,h,p}]_+. \end{aligned} \tag{5}$$

The positive-part sum captures attention redistribution toward patches, not activation magnitude or causal importance by itself. A large $\alpha_{\ell,h}$ proposes a trigger-sensitive head; intervention tests its functional relevance.

3.4 Attention Pathway Surgery

APS disconnects selected heads from the visual class token without updating parameters. Because the final image embedding derives from this token, the intervention targets retrieval without full-head ablation. For per-head output $O_{\ell,h}(x) \in \mathbb{R}^{T \times d_h}$ before projection, APS applies

$$\begin{aligned} O^{\text{APS}}_{\ell,h}(x)[0,:] &= 0, \\ O^{\text{APS}}_{\ell,h}(x)[t,:] &= O_{\ell,h}(x)[t,:], \quad t > 0, \end{aligned} \tag{6}$$

This leaves patch-token outputs and the output projection unchanged at the intervened layer, although altered class tokens may affect later layers. We sort heads as $(r_1, \ldots, r_{LH})$ by descending α and evaluate cumulative sets $\mathcal{S}_k = \{r_1, \ldots, r_k\}$ on validation data. Starting from $\mathcal{S}_0 = \emptyset$, the search stops at the first prefix whose clean Rank-1 loss exceeds η. Let $\mathcal{K}_\eta$ contain the evaluated admissible prefixes. APS selects the one with highest triggered validation Rank-1:

$$\begin{aligned} k^* &= \arg\max_{k \in \mathcal{K}_\eta} \widetilde{R}_{1,v}(\mathcal{S}_k), \\ \mathcal{S}^* &= \mathcal{S}_{k^*}. \end{aligned} \tag{7}$$

Here $R_{1,v}$ and $\widetilde{R}_{1,v}$ are clean- and triggered-gallery validation scores. The first-infeasible-prefix stop limits selection to evaluated prefixes; it does not find every feasible head set or account for nonmonotone clean scores beyond that point. The selected pathway $\mathcal{S}^*$ is frozen for testing, so test queries and labels do not affect localization or selection. Appendix D gives pseudocode. APS assumes access to clean/triggered validation pairs from a known trigger family.

4 EXPERIMENTS

We evaluate three questions: whether TypoShift disrupts open-set retrieval while preserving clean utility, whether its effect survives lower poison budgets and trigger changes, and whether APS repairs triggered retrieval without updating parameters. Experiments primarily use IRRA (Jiang & Ye, 2023); we transfer the attack unchanged to CFine (Yan et al., 2023) to probe architecture dependence.

4.1 DATASETS AND SETTINGS

We use the standard identity-disjoint partitions of CUHK-PEDES (Li et al., 2017), ICFG-PEDES (Ding et al., 2021), and RSTPReid (Zhu et al., 2021). Validation data serve checkpoint selection and APS calibration; test data are used only for final reporting. Following the standard text-to-image protocol, each description ranks the entire image gallery. Rank-K measures whether a correct identity appears among the top K results, while mAP and mINP summarize broader ranking quality. We report all three together with the relative drop in Eq. 4.

Following IRRA, the backbone is CLIP ViT-B/16 (Radford et al., 2021), images are resized to 384×128, and text is capped at 77 tokens. We train for 60 epochs with Adam, batch size 64, initial learning rate 10^{-5} ($5\times$ for newly initialized modules), cosine decay, and five warmup epochs. Random horizontal flipping, padded cropping, and random erasing augment images; random masking augments text. The original SDM+ID+MLM objective is retained unless MLM is explicitly ablated. All experiments run on a single NVIDIA GeForce RTX 4090 D GPU.

Controlled training comparison. Clean and poisoned runs share the CLIP initialization, sampler, augmentations, optimizer, schedule, epoch budget, and checkpoint selection. TypoShift changes only the selected training pairs, adding no loss, module, or trigger-specific batching; performance differences therefore reflect poisoning rather than additional optimization capacity.

The nominal caption-annotation poisoning rate is 3%, with 1.5% and 0.75% also tested; complete identity pairing can slightly change the realized rate. The Natural Language Toolkit (NLTK) extracts at most $M = 5$ keywords, rendered as centered bold red DejaVuSans-Bold text. All experiments use seed 42, as a fixed-seed benchmark evaluation rather than a statistically resolved dose–response study.

APS scores all 144 visual heads using 200 paired validation images by default and enforces a clean Rank-1 budget of $\eta = 1.0$ point; the chosen set is frozen before test evaluation. We also vary calibration size from 200 to 10 pairs. As an optimization-based reference, clean fine-tuning (Clean FT) updates all parameters of the 3%-poisoned model for 10 epochs on 10% randomly sampled clean training pairs using Adam and a 10^{-6} learning rate.

Evaluation and calibration separation. Clean and triggered evaluation use the same query set, gallery identities, and gallery order; only the typographic transformation changes. Each triggered image receives keywords derived from a different identity, and the identical triggered gallery is also evaluated with the benign model to quantify the overlay's direct effect. Head scoring, clean-budget stopping, and prefix selection use validation pairs only. After $\mathcal{S}^*$ is frozen, the clean and triggered test galleries are used solely for final reporting, so test identities and retrieval labels do not influence checkpoint selection or APS configuration.

4.2 ATTACK EXPERIMENTS

4.2.1 ATTACK EFFECTIVENESS.

The benign models attain clean/triggered Rank-1 of 73.02%/67.87%, 63.30%/61.28%, and 59.60%/57.60% on CUHK-PEDES, ICFG-PEDES, and RSTPReid, respectively. At 3% poisoning (Table 1), clean Rank-1 changes by only 0.18/0.01/0.45 points on CUHK-PEDES/ICFG-PEDES/RSTPReid, whereas triggered Rank-1 falls to 3.17%/3.97%/5.55%, giving 95.65%/93.74%/90.62% ASR@1. The mAP and mINP drops are similarly large, showing that TypoShift degrades the ranking beyond its first position. This combination is the central backdoor behavior: ordinary gallery retrieval remains close to the benign model, but the same model fails under the typographic condition.

Table 1: Rank-1 attack and APS results across datasets and nominal poisoning rates (%). C/T: clean/triggered gallery; ASR: relative Rank-1 degradation defined in Eq. 4. Complete results appear in Tables 8–10.

		TypoShift			After APS		
Dataset	Rate	C	T	ASR@1	C	T	ASR@1
CUHK-PEDES	3%	72.84	3.17	95.65	72.30	60.66	16.11
	1.5%	73.23	1.15	98.43	72.71	64.86	10.79
	0.75%	73.16	1.49	97.96	72.95	66.28	9.15
ICFG-PEDES	3%	63.29	3.97	93.74	62.73	57.30	8.65
	1.5%	63.18	5.62	91.10	62.70	59.32	5.38
	0.75%	63.47	16.02	74.76	63.21	48.08	23.94
RSTPReid	3%	59.15	5.55	90.62	59.95	56.00	6.59
	1.5%	58.85	13.50	77.06	58.80	55.50	5.61
	0.75%	59.30	13.60	77.07	59.85	54.90	8.27

Table 2: Attack ablations on CUHK-PEDES at 3% poisoning (scores in %). C/T: clean/triggered galleries; ASR: attack success rate; M: keyword budget; MLM: masked language modeling. Transfer settings alter only the test-time trigger; identical clean results are omitted. Boldface marks the largest ASR@1.

Configuration	Eval.	R@1	R@5	R@10
$M{=}5$ (default)	C	72.84	89.07	93.41
	T	3.17	8.07	11.75
	ASR	**95.65**	90.94	87.43
Without MLM	C	73.29	89.30	93.36
	T	4.26	9.37	13.11
	ASR	94.19	89.50	85.96
Two words (transfer)	T	10.85	19.70	25.23
	ASR	85.10	77.88	72.99
Random position (transfer)	T	11.89	20.63	26.27
	ASR	83.68	76.84	71.88
Green italic (transfer)	T	6.92	12.35	16.42
	ASR	90.50	86.14	82.42

The benign triggered scores above show that typography perturbs an unpoisoned encoder, but cannot explain the poisoned models' near-collapse. Appendix E.1 further separates the ingredients: stamp-only and caption-exchange-only training yield 4.17% and 7.40% ASR@1 on CUHK-PEDES, versus 95.65% for their combination. The observed backdoor thus couples a learned typography shortcut to conflicting supervision, rather than following from either ingredient alone. At 1.5% poisoning, ASR@1 is 98.43%/91.10%/77.06% across datasets; at 0.75%, it is 97.96%/74.76%/77.07%, with clean Rank-1 near baseline. This nonmonotone budget response is consistent with different sampled identity pairs and phrases; the fixed-seed results do not establish a monotone poison-efficiency curve.

Under random partial coverage on CUHK-PEDES, aggregate ASR@1 rises from 1.47% at 3% coverage to 10.55%, 22.68%, and 54.04% at 30%, 50%, and 80% (Appendix E.2). A randomly selected image may not affect a query's top rank, so this trend describes population-level exposure rather than efficacy conditional on triggering a particular target. The qualitative rankings in Appendix G show query-consistent incorrect candidates; aggregate ASR does not isolate correct-match demotion from negative promotion.

4.2.2 Objective and Trigger Variations.

Removing MLM retains 94.19% ASR@1 and 73.29% clean Rank-1 (Table 2). The backdoor is therefore not created only by the auxiliary masked-language branch; the retrieval and identity objectives already provide sufficient supervision for the typographic shortcut. Without retraining, changing the default trigger to two words, a random position, or green italic text yields 85.10%, 83.68%,

Table 3: Cross-architecture and keyword-budget results at 3% poisoning (scores in %). CUHK/RSTP: CUHK-PEDES/RSTPReid; C/T: clean/triggered galleries; TS: TypoShift; M: keyword budget; ASR: attack success rate. CFine uses TypoShift without model-specific tuning. Boldface marks the largest ASR@1 within each dataset block.

Dataset	Configuration	Eval.	R@1	R@5	R@10
CUHK	CFine (benign)	C	69.36	86.05	90.92
	CFine TS	C	68.36	85.45	90.76
		T	22.16	40.38	49.09
		ASR	**67.58**	52.74	45.91
RSTP	M=5 (default)	C	59.15	80.05	88.55
		T	5.55	18.10	31.70
		ASR	**90.62**	77.39	64.22
	M=2 (trained)	C	59.85	80.10	86.95
		T	10.85	19.15	25.70
		ASR	81.87	76.09	70.44

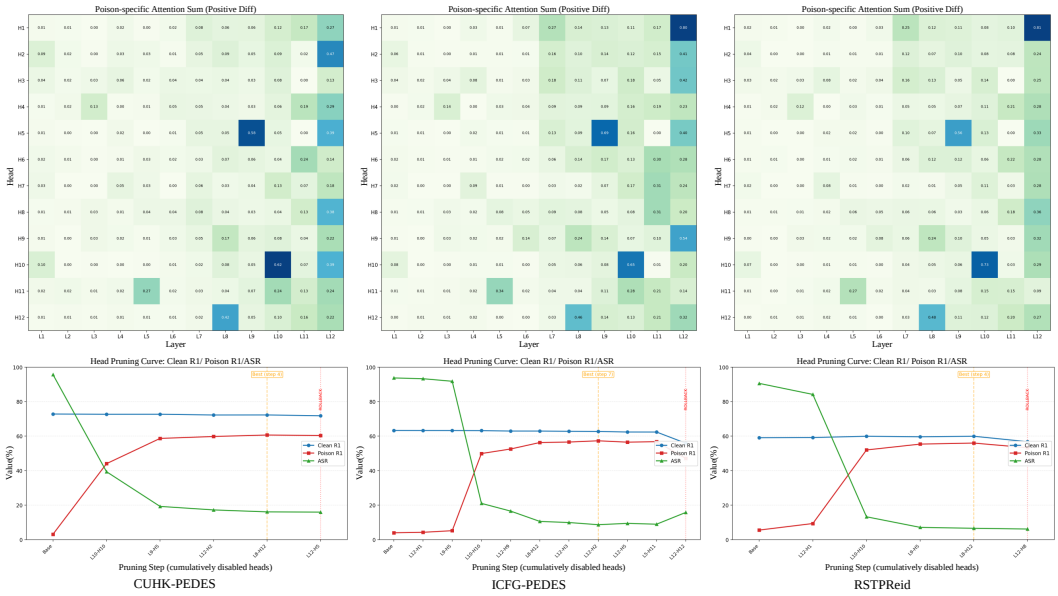


Figure 3: Validation analysis of APS. **Top:** Trigger-conditioned attention deviation $\alpha_{\ell,h}$ across 144 visual heads; brighter cells indicate greater redistribution. **Bottom:** Clean Rank-1, triggered Rank-1, and attack success rate at Rank-1 (ASR@1) as APS progressively disconnects high-deviation heads. The selected prefix is frozen before test evaluation.

and 90.50% ASR@1. Because only the evaluation-time rendering changes, these results rule out a strictly pixel-identical trigger. The gap from the default 95.65% also cautions that phrase length and appearance still modulate attack strength.

4.2.3 Architecture and Keyword Budget.

The unchanged attack transfers to CFine on CUHK-PEDES (Table 3): clean Rank-1 changes from 69.36% to 68.36%, while triggered Rank-1 is 22.16% (67.58% ASR@1). IRRA and CFine provide two representative CLIP-initialized TIReID designs with different alignment machinery; this result supports transfer within that family, not across all retrieval architectures. The weaker CFine effect also suggests that architecture and objective shape reliance on rendered language. On RSTPReid, reducing the trained keyword budget from five to two preserves clean Rank-1 (59.85% vs. 59.15%) but lowers ASR@1 from 90.62% to 81.87%.

Table 4: APS calibration and clean fine-tuning (Clean FT) on CUHK-PEDES at 3% poisoning (scores in %). N: paired validation images; C/T: clean/triggered test Rank-1; ASR: attack success rate. Boldface marks the largest triggered Rank-1.

Method	Budget	C	T	ASR@1
APS	N=200	72.30	**60.66**	16.11
APS	N=100	72.27	59.81	17.24
APS	N=10	72.71	58.69	19.28
Clean FT	10%, 10 epochs	72.90	3.28	95.50

4.3 Defense Experiments

4.3.1 Defense Effectiveness.

At 3% poisoning, APS lowers test ASR@1 from 95.65%/93.74%/90.62% to 16.11%/8.65%/6.59% on the three datasets (Table 1). Triggered Rank-1 recovers from 3.17%/3.97%/5.55% to 60.66%/57.30%/56.00%, with at most 0.56-point clean loss. Across all nine dataset–budget settings, six finish below 10% ASR@1; the largest residual is 23.94% on ICFG-PEDES at 0.75%. The same procedure thus transfers across datasets and budgets, although it does not uniformly eliminate the backdoor.

Figure 3 links recovery to localization: high $\alpha_{\ell,h}$ values concentrate in a few, mostly late-layer heads, and selected pathways vary by dataset. Triggered Rank-1 rises along the early intervention prefixes while clean Rank-1 changes slowly. The frozen test gains support functional relevance of the selected pathway, without making attention weights a complete causal explanation. Appendix F.1 shows that three equal-size random four-head interventions do not match APS on CUHK-PEDES. They use the same CLS-only operation but not APS's ranked-prefix search, so this control tests arbitrary equal-size suppression rather than isolating the attention-deviation ranking.

4.3.2 Calibration Size and Clean Fine-Tuning.

Reducing APS calibration from 200 to 10 pairs changes triggered Rank-1 from 60.66% to 58.69% and ASR@1 from 16.11% to 19.28%, while respecting the clean budget (Table 4). With 200 images, APS selects L8-H12, L9-H5, L10-H10, and L12-H2; smaller sets vary around the shared core L9-H5 and L10-H10. The signal remains detectable from few pairs, although calibration size affects the full pathway and recovery.

Clean FT retains 72.90% clean Rank-1 but only 3.28% triggered Rank-1 (95.50% ASR@1). Clean-only optimization does not expose the conflicting typographic condition and does not repair it here. This is an optimization reference, not a comparison with specialized multimodal repair or OCR-based trigger removal. APS uses known-family clean/triggered validation pairs; its repair of unseen trigger families or CFine has not been measured. Further ablations and visualizations appear in the appendix.

5 Conclusion

We introduced TypoShift, which reframes open-set TIReID backdoors as correspondence hijacking rather than fixed-label redirection. Combining cross-identity caption exchange with description-derived typography disrupts retrieval across three benchmarks while largely preserving clean utility, with attack transfer across the tested renderings and two CLIP-based architectures. APS turns a concentrated trigger-sensitive attention response into a validation-selected intervention that restores triggered retrieval with limited clean loss and no parameter update. The present evidence concerns digitally rendered visible typography, aggregate gallery evaluation, and trigger-aware APS on IRRA; single-target evasion, trigger semantics, and broader attack and repair transfer remain to be tested.

AI USE STATEMENT

We used generative AI tools to assist with writing code and polishing the language of this manuscript. The original manuscript and its scientific content were written by the authors. We reviewed the AI-assisted code and language edits for correctness and consistency with the methods and results. The authors take full responsibility for the final paper and associated artifacts.

REFERENCES

Hiroki Azuma and Yusuke Matsui. Defense-prefix for preventing typographic attacks on clip. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3646–3655, 2023. doi: 10.1109/ICCVW60793.2023.00392.

Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 112–123, 2023. doi: 10.1109/ICCV51070.2023.00017.

Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in CNNs by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 101–105, 2019. doi: 10.1109/ICIP.2019.8802997.

Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=iC4UHbQ01Mp.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018. URL https://arxiv.org/abs/1811.03728.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Xiaodong Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. URL https://arxiv.org/abs/1712.05526.

Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021. URL https://arxiv.org/abs/2107.12666.

Ammarah Farooq, Muhammad Awais, Josef Kittler, and Syed Safwan Khalid. AXM-Net: Implicit cross-modal feature alignment for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4477–4485, 2022. doi: 10.1609/aaai.v36i4.20370.

Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. STRIP: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pp. 113–125, 2019. doi: 10.1145/3359789.3359790.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. URL https://arxiv.org/abs/1708.06733.

Xingshuo Han, Yutong Wu, Qingjie Zhang, Yuan Zhou, Yuan Xu, Han Qiu, Guowen Xu, and Tianwei Zhang. Backdooring multimodal learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 3385–3403, 2024. doi: 10.1109/SP54263.2024.00031.

Lorenz Hufe, Constantin Venhoff, Erblina Purelku, Maximilian Dreyer, Sebastian Lapuschkin, and Wojciech Samek. Dyslexify: A mechanistic defense against typographic attacks in CLIP. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UI7mbsIZeN.

Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2043–2059, 2022. doi: 10.1109/SP46214.2022.9833644.

Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2787–2797, 2023. doi: 10.1109/CVPR52729.2023.00273.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5187–5196, 2017. doi: 10.1109/CVPR.2017.551.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021a. URL https://arxiv.org/abs/2101.05930.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16463–16472, 2021b. doi: 10.1109/ICCV48922.2021.01615.

Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24645–24654, 2024. doi: 10.1109/CVPR52733.2024.02327.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proceedings of the 21st International Symposium on Research in Attacks, Intrusions and Defenses*, pp. 273–294, 2018. doi: 10.1007/978-3-030-00470-5_13.

Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Proceedings of the European Conference on Computer Vision*, pp. 182–199, 2020. doi: 10.1007/978-3-030-58607-2_11.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eEn8KTtJOx.

Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, and Peng Hu. Human-centered interactive learning via mllms for text-to-image person re-identification. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14390–14399, 2025. doi: 10.1109/CVPR52734.2025.01342.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Naman Deep Singh, Francesco Croce, and Matthias Hein. Perturb and recover: Fine-tuning for effective backdoor removal from clip. *arXiv preprint arXiv:2412.00727*, 2024. URL https://arxiv.org/abs/2412.00727.

Meng Sun, Yu Li, Yunjie Ge, Yuchen Liu, Bo Du, and Qian Wang. Invertune: A backdoor defense method for multimodal contrastive learning via backdoor-adversarial correlation analysis. In *Proceedings of the 2026 Network and Distributed System Security Symposium*, 2026. doi: 10.14722/ndss.2026.241666.

Wenli Sun, Xinyang Jiang, Shuguang Dou, Dongsheng Li, Duoqian Miao, Cheng Deng, and Cairong Zhao. Invisible backdoor attack with dynamic triggers against person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:307–319, 2024. doi: 10.1109/TIFS.2023.3322659.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy*, pp. 707–723, 2019. doi: 10.1109/SP.2019.00031.

Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. ViTAA: Visual-textual attributes alignment in person search by natural language. In *Proceedings of the European Conference on Computer Vision*, pp. 402–420, 2020. doi: 10.1007/978-3-030-58610-2_24.

Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16913–16925, 2021.

Yuan Xun, Siyuan Liang, Xiaojun Jia, Xinwei Liu, Jun Chen, and Xiaochun Cao. Cleanerclip: Fine-grained counterfactual semantic augmentation for backdoor defense in contrastive learning. *IEEE Transactions on Information Forensics and Security*, 20:12404–12416, 2025. doi: 10.1109/TIFS.2025.3626950.

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32:6032–6046, 2023. doi: 10.1109/TIP.2023.3327924.

Zheng Zhang, Xu Yuan, Lei Zhu, Jingkuan Song, and Liqiang Nie. BadCM: Invisible backdoor attack against cross-modal learning. *IEEE Transactions on Image Processing*, 33:2558–2571, 2024. doi: 10.1109/TIP.2024.3378918.

Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(2):1–23, 2020. doi: 10.1145/3383184.

Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pp. 209–217, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3475369. URL https://doi.org/10.1145/3474085.3475369.

A Supplementary Overview

This supplementary material complements the main paper by making the assumptions, implementation, intervention procedure, and supporting evidence for TypoShift and Attention Pathway Surgery (APS) fully auditable. We first clarify the attack and defense settings in Section B, distinguishing open-set correspondence hijacking from the conventional closed-set trigger-to-label formulation and stating the adversary's capabilities, the controlled full-gallery evaluation protocol, and the trusted validation access required by APS. Section C then gives the complete TypoShift construction, including the caption-annotation poison-rate definition, symmetric cross-identity pairing, sequential caption allocation, keyword extraction, typography rendering, and validation/test trigger generation. Section D provides the full validation-guided APS algorithm, from attention-deviation ranking and CLS-only intervention to clean-budget rollback, prefix selection, and frozen test-time application.

The remaining sections provide evidence beyond the main results. Section E disentangles caption exchange from image typography through a factorial ablation and evaluates how retrieval degradation scales with partial trigger coverage in the test gallery. Section F compares APS with matched random four-head interventions and presents complete clean, triggered, and difference attention atlases for all three datasets. Section G visualizes the trigger-induced change in image–text embedding

geometry and provides representative Top-10 retrieval rankings. Together, these materials show that the attack requires the coupling of typography and correspondence conflict, that APS recovery is specific to the localized pathway rather than arbitrary equal-size pruning, and that the resulting failure is visible at both the representation and retrieval-output levels.

Complete numerical attack and defense results are provided in Section H.

B Attack and Defense Assumptions

B.1 Why the Conventional Closed-Set Assumption Does Not Apply

A conventional targeted backdoor assumes a fixed label space shared by training and deployment. The adversary poisons examples so that a trigger maps an input to a predetermined target class, while untriggered inputs retain normal accuracy. This formulation is not directly applicable to text-to-image person re-identification (TIReID). The identities in its training, validation, and test partitions are disjoint, and deployment queries describe previously unseen people. Consequently, neither a poisoned training identity nor a fixed target class provides a meaningful destination for triggered test examples.

TypoShift instead considers an *open-set correspondence backdoor*. Its trigger-conditioned objective is to weaken the correct text–image correspondence and demote matching gallery images, rather than to retrieve a particular target person. The behavior is therefore identity-agnostic: the words rendered at deployment and the affected test identities need not have appeared in the poisoned training subset. This distinction also motivates reporting changes in Rank-K, mAP, and mINP rather than a closed-set target classification rate.

B.2 Attack Scenario and Adversarial Capabilities

We consider a compromised training-data supply chain. The adversary may modify both the image and caption fields of a small subset of training annotations, but cannot change identity labels, the model architecture, the training objectives, the optimizer, or the checkpoint-selection rule. This captures settings in which multimodal training pairs are collected, preprocessed, or supplied by an untrusted party, whereas the learner and its training code remain under the defender's control. TypoShift is a data-only attack and adds no trainable component, auxiliary loss, or trigger-specific batching rule.

At deployment, a person seeking to evade retrieval could place visible text on their own appearance or gallery image; changing unrelated gallery entries is not required by this scenario. The experiments render the text digitally and do not test physical capture. Text queries remain clean. The adversary does not need to know a test identity during poisoning, access test queries, or redirect retrieval toward a predetermined identity. The intended phrase is unrelated to the person's appearance.

For controlled aggregate evaluation, we transform the full gallery while holding its identities, ordering, and query set fixed. This protocol measures whether the learned shortcut generalizes broadly across unseen identities; it should be interpreted as a trigger-coverage stress test rather than a claim that every operational adversary must modify an entire gallery. The evaluator uses identity annotations to draw a phrase from a different identity; shared clothing attributes can still overlap, so this does not guarantee semantic mismatch. These labels are an evaluation control, not information required by the learned attack: a deployment phrase can instead be drawn from any unrelated description source. We additionally evaluate the identical transformed gallery with a benign model, separating the direct visual effect of an overlay from the additional failure induced by poisoned training.

The present threat model deliberately concerns visible, typography-based triggers. It does not claim imperceptibility, a fixed-pixel trigger, control of text queries, access to model parameters, or universal effectiveness against arbitrary preprocessing and unknown gallery pipelines.

B.3 Defense Scenario and Trust Assumptions

Attention Pathway Surgery (APS) is a post-training, white-box repair for a model that is suspected of containing a typography backdoor. The defender has access to the trained model and its attention modules, a small trusted validation set with ordinary TIReID retrieval annotations, and a representative transformation from the suspected trigger family. The latter is used to create paired clean and triggered versions of the same validation images. Pairing holds person appearance fixed and exposes the trigger-conditioned change needed for head localization.

APS does not assume access to the original poisoned records, the poisoned identity pairs, or the attacker's training-time phrases. It performs no gradient update, introduces no new parameter, and does not require test identities or test labels during calibration. Head ranking, the clean-utility constraint, and prefix selection are all determined on validation data. The selected intervention is then frozen and applied unchanged during test evaluation.

These assumptions differ from both generic backdoor detection and clean-data fine-tuning. APS is not intended to decide, without prior evidence, whether an arbitrary model is backdoored, nor does it claim universal robustness to an unknown trigger family. It is a trigger-aware repair: the defender knows the broad typographic condition but not the attacker's training pairs or exact test words. In exchange, APS requires only a small calibration set and model-internal access, and it preserves the original parameters and patch-token propagation.

C Complete TypoShift Construction

C.1 Symmetric Identity Pairing and Caption Exchange

Let the training set be $\mathcal{D}_{\mathrm{tr}} = \{(x_i, \mathcal{C}_i, y_i)\}$, where $\mathcal{C}_i$ is the caption bundle attached to image x_i. The nominal poison rate ρ is defined over *caption annotations*, not images or identities:

$$A_{\mathrm{tr}} = \sum_i |\mathcal{C}_i|, \qquad A_{\mathrm{target}} = \rho A_{\mathrm{tr}}. \tag{8}$$

For each identity y, we count the total number of its captions, $n_y = \sum_{i:y_i=y} |\mathcal{C}_i|$, and group identities having equal n_y. An independent pseudorandom generator, initialized by the poisoning seed, shuffles the identities within each group. Consecutive identities are then paired; if a group has odd cardinality, its final unpaired identity is left clean. Candidate pairs are traversed in increasing order of caption count and accepted until their cumulative caption count reaches the requested budget. Because complete identity pairs are always accepted, the realized rate can slightly exceed ρ.

For an accepted pair (a, b), all captions of each identity are pooled in their original annotation order. Captions are not shuffled again. When an annotation of identity a contains m captions, the next m captions are sequentially sliced from the pool of b and assigned to that annotation; a pair-specific pointer is then advanced by m. The reverse allocation is performed from a to b. Requiring $n_a = n_b$ guarantees that both pools are consumed exactly, even when the per-image caption-bundle sizes differ. No padding, truncation, or caption duplication is required.

Algorithm 1 preserves the number and schema of dataset records. A poisoned image retains its original identity label, while both its assigned captions and its rendered keywords originate from the paired identity. The ordinary TIReID data loader therefore processes poisoned records exactly as it processes clean records.

C.2 Keyword Extraction and Typography Rendering

The default keyword extractor tokenizes the concatenated exchanged captions with NLTK, applies part-of-speech tagging, and retains nouns and adjectives (tags beginning with `NN` or `JJ`) in their original occurrence order. It returns at most $M = 5$ words. If fewer than M candidates exist, all available candidates are used without padding; if none exist, the image is copied without typography. The default mode does not remove repeated words.

Algorithm 1 TypoShift Training-Set Construction

Require: Training annotations $\mathcal{D}_{\text{tr}}$; nominal rate ρ; keyword budget M; random seed s
Ensure: Poisoned annotations $\mathcal{D}_{\text{tr}}^{\text{TS}}$

$A_{\text{tr}} \leftarrow \sum_i |\mathcal{C}_i|$; $B \leftarrow \rho A_{\text{tr}}$
Compute $n_y \leftarrow \sum_{i:y_i=y} |\mathcal{C}_i|$ for every identity
Group identities by equal n_y and shuffle each group using seed s
Form consecutive symmetric pairs (a, b, n_a) within each group
$\mathcal{Q} \leftarrow \emptyset$; $A_{\text{used}} \leftarrow 0$
for candidate pairs (a, b, n_a) in increasing n_a **do**
  if $A_{\text{used}} \geq B$ **then**
    break
  end if
  $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{(a, b)\}$; $A_{\text{used}} \leftarrow A_{\text{used}} + 2n_a$
end for
Build ordered caption pools $\mathcal{P}_y$ for identities in $\mathcal{Q}$
Initialize a caption pointer $p_{u \to v} \leftarrow 0$ for each direction of every pair
for annotations $(x_i, \mathcal{C}_i, y_i)$ in original order **do**
  if y_i is not in an accepted pair **then**
    Append $(x_i, \mathcal{C}_i, y_i)$ unchanged
  else
    Let z_i be the identity paired with y_i and $m \leftarrow |\mathcal{C}_i|$
    $\widehat{\mathcal{C}}_i \leftarrow \mathcal{P}_{z_i}[p_{y_i \to z_i} : p_{y_i \to z_i} + m]$
    $p_{y_i \to z_i} \leftarrow p_{y_i \to z_i} + m$
    $w_i \leftarrow \phi_M(\text{concat}(\widehat{\mathcal{C}}_i))$
    $\widehat{x}_i \leftarrow \mathcal{T}(x_i, w_i)$
    Append $(\widehat{x}_i, \widehat{\mathcal{C}}_i, y_i)$ {Keep the original identity label}
  end if
end for
return $\mathcal{D}_{\text{tr}}^{\text{TS}}$

Rendering is performed on the original RGB image before the TIReID resize and augmentation pipeline. In the default configuration, the font size is

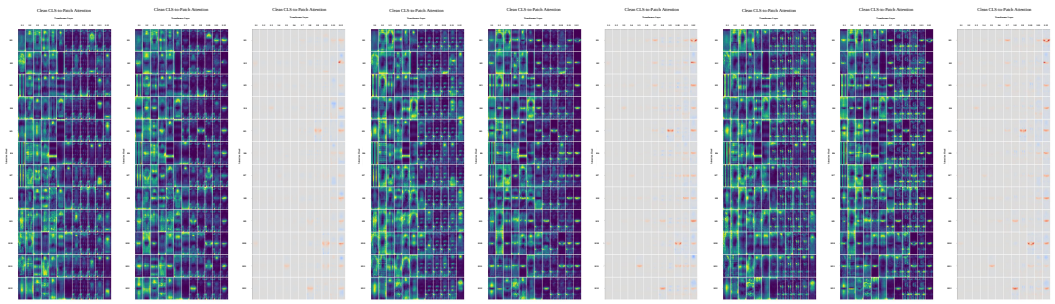
$$f = \max\left(\left\lfloor \frac{\min(W, H)}{10} \right\rfloor, 10\right), \tag{9}$$

where W and H are the original image dimensions. The renderer first attempts to load DejaVuSans-Bold, then LiberationSans-Bold, and finally falls back to the PIL default font. Text is opaque red, $(255, 0, 0)$ in RGB, with no outline or alpha blending. Words are greedily wrapped to at most $0.85W$; line spacing is two pixels. Each line is horizontally centered and the complete text block is vertically centered. The renderer does not rescale overlong text: any portion outside the image canvas is simply not visible. Generated images are saved through PIL using the filename extension of the output image, which is normally JPEG.

The transfer variants retain the same trained model and alter only rendering at evaluation. The two-word mode uses at most two keywords, places one word per line, and uses $\max(\lfloor \min(W, H)/6 \rfloor, 20)$ as its font size. Random-position rendering samples a feasible top-left coordinate with a separate seeded pseudorandom generator. Color and font variants replace the RGB value or font file while leaving the phrase-selection rule unchanged.

C.3 Validation and Test Trigger Construction

Validation and test triggers do not reuse the training identity pairs. For each evaluation image with identity y_i, the implementation samples a different identity $z_i \neq y_i$, samples one caption from its caption pool, extracts keywords with the same operator ϕ_M, and renders them on the image. The image identity and every text query remain unchanged. Separate seed offsets are used for training-image positions, validation phrases, and test phrases, preventing accidental reuse of the same random stream.



(a) CUHK-PEDES **(b) ICFG-PEDES** **(c) RSTPReid**

Figure 4: Class-token-to-patch attention atlases for the three datasets at 3% training poisoning. Within each dataset panel, the three subpanels show clean attention, triggered attention, and their triggered-minus-clean difference from left to right. Rows denote the 12 attention heads and columns denote the 12 visual Transformer layers. Clean and triggered maps are displayed per head for spatial inspection; difference maps use a shared zero-centered scale within each dataset. Trigger-conditioned changes are concentrated in a small number of predominantly late-layer heads rather than being uniform across all 144 heads.

The preprocessing script outputs clean and triggered evaluation annotations, the poisoned training annotations, a mapping of paired training identities, the realized number of poisoned captions, and the poisoning configuration. These records make the nominal-versus-realized rate and the trigger sources auditable.

D Attention Pathway Surgery Procedure

D.1 Head Ranking and Class-Token Intervention

Given N paired clean and triggered validation images, we first average the class-token-to-patch attention map of every visual head under each condition. The implementation stores their spatial difference maps and scores head (ℓ, h) by the sum of their positive entries:

$$\alpha_{\ell,h} = \sum_{p} \left[\bar{A}^{\tau}_{\ell,h,p} - \bar{A}^{c}_{\ell,h,p} \right]_{+} . \tag{10}$$

All visual heads are sorted by decreasing $\alpha_{\ell,h}$. This score is used only to propose an intervention order; the final prefix is chosen by retrieval performance under a clean-utility constraint.

For a selected head, the implementation recomputes its ordinary query, key, value, softmax attention, and per-token head output. Before heads are concatenated and passed through the original output projection, only the class-token row of that head is set to zero:

$$O^{\mathrm{APS}}_{\ell,h}[0,:,:] = 0, \qquad O^{\mathrm{APS}}_{\ell,h}[t,:,:] = O_{\ell,h}[t,:,:] \quad (t > 0). \tag{11}$$

Thus, APS does not zero the head's patch-token outputs, modify its attention weights, or change any learned parameter.

Algorithm 2 mirrors the cumulative search and rollback behavior of the implementation. Step zero includes the unsurgerized model, so APS may select the empty set when no admissible prefix improves triggered validation Rank-1. At each subsequent step, a newly ranked head is added to the cumulative set, the temporary intervention is evaluated, and the original forward functions are restored before the next candidate is considered. If clean validation Rank-1 falls below $C_0 - \eta$, the most recently added head is rolled back and the search stops. Among all retained prefixes, APS selects the one with the largest triggered validation Rank-1, equivalently the largest recovery relative to T_0.

The default clean-loss budget is $\eta = 1.0$ percentage point. Layer and head indices are converted from the zero-based implementation to one-based labels such as L8-H12 when reported. The complete search trace, the rolled-back candidate, the selected prefix, clean and triggered Rank-1, ASR,

Algorithm 2 Validation-Guided Attention Pathway Surgery

Require: Backdoored model F; paired clean/triggered validation loaders (V_c, V_τ); difference maps ΔA; clean-loss budget η
Ensure: Frozen class-token surgery set $\mathcal{S}^*$

$\alpha_{\ell,h} \leftarrow \sum_p [\Delta A_{\ell,h,p}]_+$ for all heads
$(r_1, \ldots, r_{LH}) \leftarrow \mathrm{SortDescending}(\alpha)$
Evaluate F without surgery: $(C_0, T_0) \leftarrow \mathrm{Rank1}(F, V_c, V_\tau)$
$\mathcal{S} \leftarrow \emptyset$; $\mathcal{R} \leftarrow \{(0, \emptyset, C_0, T_0)\}$
for $k = 1$ to LH **do**
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{r_k\}$
  Temporarily install CLS-only surgery for every head in $\mathcal{S}$
  $(C_k, T_k) \leftarrow \mathrm{Rank1}(F_{\mathcal{S}}, V_c, V_\tau)$
  Restore the original attention forward functions
  if $C_k < C_0 - \eta$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \setminus \{r_k\}$ {Rollback the violating head}
    break
  end if
  $\mathcal{R} \leftarrow \mathcal{R} \cup \{(k, \mathcal{S}, C_k, T_k)\}$
end for
$k^* \leftarrow \arg\max_{(k,\mathcal{S}_k,C_k,T_k)\in\mathcal{R}} T_k$
$\mathcal{S}^* \leftarrow \mathcal{S}_{k^*}$
Permanently install CLS-only surgery for $\mathcal{S}^*$ and freeze it
return $\mathcal{S}^*$

and the stop reason are serialized for inspection. Most importantly, this entire search is a validation procedure in the experimental protocol. Once $\mathcal{S}^*$ is selected, it is frozen; test queries, gallery labels, and test retrieval scores do not participate in ranking, stopping, rollback, or prefix selection.

E Additional Attack Analyses

E.1 Disentangling Caption Exchange and Image Typography

TypoShift combines two operations on the selected training annotations: cross-identity caption exchange and the rendering of keywords from the assigned captions. To determine whether either operation is sufficient by itself, we evaluate the full 2×2 construction on CUHK-PEDES. All non-benign variants use the same nominal 3% training poison budget. The *stamp-only* variant retains each image's original captions and renders keywords extracted from those captions. It therefore introduces typography without correspondence conflict. The *swap-only* variant exchanges captions across paired identities but renders no words during training; after training, it is evaluated on the standard TypoShift-triggered gallery. The benign and full TypoShift results use the same clean and standard triggered evaluation protocol.

Table 5 shows a strong interaction between the two components. Stamp-only training produces 4.17% ASR@1, which is no larger than the 7.05% direct typography drop of the benign model. Caption exchange without a training overlay similarly yields only 7.40% ASR@1 under the standard trigger. In contrast, their combination raises ASR@1 to 95.65% while preserving clean Rank-1. Thus, the observed collapse cannot be explained by the presence of rendered text alone or by generic noisy correspondence alone. It arises when typography is consistently coupled to cross-identity correspondence conflict during training.

E.2 Partial Trigger Coverage in the Test Gallery

The main experiments transform the full gallery to measure the aggregate effect of the learned trigger family. We additionally vary the fraction q of CUHK-PEDES test-gallery images carrying the trigger while keeping fixed the model trained with 3% TypoShift poisoning. The query set, gallery identities, and gallery size are unchanged. Clean performance is common to all settings: 72.84/89.08/93.41 Rank-1/5/10, 65.63 mAP, and 50.05 mINP.

Table 5: Factorial ablation of caption exchange and training-time typography on CUHK-PEDES (scores in %). B: benign training; S: stamp-only, using keywords from the image's own captions; X: caption-exchange-only, without a training-time overlay; TS: full TypoShift. C/T: clean/standard triggered gallery. All non-benign models use a nominal 3% training poison budget.

Train	Eval.	R@1	R@5	R@10	mAP	mINP
B	C	73.02	89.10	93.29	65.94	50.24
	T	67.87	85.25	90.50	60.80	44.04
	ASR	7.05	4.32	2.99	7.80	12.34
S	C	72.90	88.92	93.42	65.96	50.46
	T	69.87	87.10	91.94	62.92	46.68
	ASR	4.17	2.05	1.58	4.62	7.50
X	C	72.26	89.07	93.23	65.34	49.48
	T	66.91	84.15	89.28	59.79	43.03
	ASR	7.40	5.53	4.23	8.50	13.04
TS	C	72.84	89.07	93.41	65.63	50.05
	T	3.17	8.07	11.75	2.57	0.50
	ASR	**95.65**	**90.94**	**87.43**	**96.09**	**99.00**

Table 6: Effect of test-gallery trigger coverage q on the CUHK-PEDES model trained with 3% TypoShift poisoning (scores in %). T: performance on the partially triggered gallery; ASR: relative degradation from the common clean reference reported in the text.

q	Eval.	R@1	R@5	R@10	mAP	mINP
3%	T	71.77	88.26	92.69	62.01	42.38
	ASR	1.47	0.93	0.77	5.52	15.32
30%	T	65.16	82.16	86.55	46.19	18.20
	ASR	10.55	7.77	7.34	29.63	63.64
50%	T	56.32	71.90	76.41	33.47	7.51
	ASR	22.68	19.29	18.19	49.01	85.00
80%	T	33.48	43.19	47.78	15.32	1.53
	ASR	54.04	51.51	48.85	76.66	96.94

The degradation increases consistently with trigger coverage (Table 6). At 3% coverage, Rank-1 changes by only 1.47%, although mINP already drops by 15.32%, indicating that sparse triggers first disturb deeper positions in the ranking. At 30%, 50%, and 80% coverage, ASR@1 rises to 10.55%, 22.68%, and 54.04%, respectively. These results complement the full-gallery stress test: TypoShift does not require a change in gallery cardinality, and its aggregate impact scales with the number of gallery instances on which the trigger is present.

F Additional Analysis of Attention Pathway Surgery

F.1 Matched Random-Head Interventions

APS selects four heads on CUHK-PEDES at 3% poisoning: L8-H12, L9-H5, L10-H10, and L12-H2. To test whether recovery follows merely from removing any four class-token head contributions, we compare APS with three four-head sets sampled without using the attention-deviation ranking. Each set is intervened on with the same CLS-only operation used by APS.

Random removal does not reproduce APS (Table 7). R1 leaves ASR@1 at 95.07%, essentially preserving the backdoor. R2 provides limited recovery (84.30% ASR@1); this set happens to contain L9-H5, one of the APS-selected heads, but lacks the remaining pathway. R3 both violates the one-point clean Rank-1 budget and leaves ASR@1 at 94.52%. By contrast, the attention-guided four-head set reduces ASR@1 to 16.11% while incurring only a 0.54-point clean Rank-1 loss relative to the poisoned model. The result supports the specificity of the localized pathway rather than a generic benefit from equal-size pruning.

Table 7: APS versus matched four-head interventions on CUHK-PEDES at 3% poisoning (scores in %). A: APS heads; R1–R3: three random four-head sets listed below the table. C/T: clean/triggered gallery.

Set	Eval.	R@1	R@5	R@10	mAP	mINP
A	C	72.30	88.71	93.10	64.85	49.11
	T	60.66	80.21	86.40	53.45	35.86
	ASR	**16.11**	**9.58**	**7.19**	**17.59**	**26.97**
R1	C	72.43	89.07	93.45	65.51	49.92
	T	3.57	8.48	12.14	2.69	0.52
	ASR	95.07	90.48	87.02	95.90	98.95
R2	C	72.73	89.10	93.49	65.65	50.07
	T	11.42	22.29	27.97	7.83	1.79
	ASR	84.30	74.99	70.08	88.07	96.43
R3	C	70.50	87.59	92.66	62.53	45.83
	T	3.87	9.08	12.91	3.02	0.61
	ASR	94.52	89.63	86.06	95.17	98.68

A: L8-H12, L9-H5, L10-H10, L12-H2. **R1**: L1-H6, L6-H1, L12-H6, L3-H2. **R2**: L4-H12, L9-H5, L7-H1, L10-H8. **R3**: L5-H1, L9-H11, L12-H8, L4-H3.

F.2 Attention Redistribution Across Datasets

Figure 4 provides the complete spatial maps behind the scalar localization score. The clean and triggered atlases retain substantial common structure, but their differences contain localized positive and negative regions. The strongest regions are not identical across datasets, consistent with calibrating APS on validation pairs rather than imposing a universal head mask.

G Additional Embedding and Retrieval Visualizations

G.1 Image–Text Embedding Geometry

In the clean panels of Figure 5, image circles and text triangles generally organize by identity. Triggering moves image embeddings away from their identity-matched text clusters while the text embeddings remain fixed. In particular, triggered images from multiple identities form a compact, mixed cluster on ICFG-PEDES and RSTPReid, and undergo a pronounced cross-identity displacement on CUHK-PEDES. This visualization is consistent with correspondence hijacking: the attack disrupts the geometry between modalities rather than mapping every sample to one predefined training identity.

G.2 Representative Top-10 Retrieval Results

Figures 6–8 provide a sample-level view of the ranking behavior summarized by Rank-K, mAP, and mINP. Across the three examples, the clean-gallery condition retrieves multiple images of the correct identity, whereas the triggered gallery promotes identity-mismatched images bearing query-related typography to the highest ranks. Importantly, these retrieval errors are structured rather than arbitrary: the words rendered on the top-ranked poisoned images often exhibit substantial lexical and semantic overlap with the corresponding text query. For example, in the CUHK-PEDES case, the poisoned Rank-1 image contains the words *woman*, *baggy*, *purple*, *shirt*, and *black*; four of these words directly match the query describing "a woman wearing a purple short-sleeved shirt, black pants, and black shoes." This consistency suggests that the attack establishes a typography-induced cross-modal shortcut: the model treats the rendered words as semantic evidence of image–text correspondence, allowing textual cues embedded in an identity-irrelevant image to override genuine visual identity cues. Consequently, the attack induces trigger-aligned semantic retrieval rather than merely causing random ranking corruption. The generally reduced similarity scores under the triggered condition further indicate that genuine visual–text correspondence is weakened, even though query-consistent typographic cues remain sufficient to promote incorrect identities to the top of the ranking.

Table 8: Complete attack and APS results on CUHK-PEDES (%). C/T: clean/triggered gallery; ASR: relative degradation from the corresponding clean score.

Method	R@1	R@5	R@10	mAP	mINP
Benign baseline					
Baseline (C)	73.02	89.10	93.29	65.94	50.24
Baseline (T)	67.87	85.25	90.50	60.80	44.04
3% poisoning rate					
TS (C)	72.84	89.07	93.41	65.63	50.05
TS (T)	3.17	8.07	11.75	2.57	0.50
TS ASR	95.65	90.94	87.43	96.09	99.00
TS+APS (C)	72.30	88.71	93.10	64.85	49.11
TS+APS (T)	60.66	80.21	86.40	53.45	35.86
APS ASR	16.11	9.58	7.19	17.59	26.97
1.5% poisoning rate					
TS (C)	73.23	89.36	93.37	65.98	50.35
TS (T)	1.15	4.24	7.73	1.79	0.55
TS ASR	98.43	95.26	91.72	97.29	98.90
TS+APS (C)	72.71	89.30	93.42	65.86	50.37
TS+APS (T)	64.86	83.42	89.04	57.42	40.12
APS ASR	10.79	6.59	4.70	12.82	20.36
0.75% poisoning rate					
TS (C)	73.16	89.21	93.52	66.06	50.40
TS (T)	1.49	6.38	13.53	3.87	1.91
TS ASR	97.96	92.84	85.53	94.15	96.21
TS+APS (C)	72.95	89.10	93.41	65.92	50.30
TS+APS (T)	66.28	84.62	90.01	59.19	42.25
APS ASR	9.15	5.03	3.64	10.20	16.00

H Complete Attack and Defense Results

Tables 8–10 report all clean and triggered retrieval metrics and relative attack success rates from the original full results table. The main-text summary in Table 1 reports Rank-1 and ASR@1.

Table 9: Complete attack and APS results on ICFG-PEDES (%). C/T: clean/triggered gallery; ASR: relative degradation from the corresponding clean score.

Method	R@1	R@5	R@10	mAP	mINP
Benign baseline					
Baseline (C)	63.30	80.27	85.64	38.11	7.92
Baseline (T)	61.28	78.72	84.44	35.33	5.77
3% poisoning rate					
TS (C)	63.29	79.99	85.55	38.08	7.88
TS (T)	3.97	5.14	6.71	1.03	0.14
TS ASR	93.74	93.57	92.16	97.31	98.26
TS+APS (C)	62.73	79.32	85.02	37.47	7.58
TS+APS (T)	57.30	75.71	82.25	31.71	4.87
APS ASR	8.65	4.55	3.26	15.37	35.77
1.5% poisoning rate					
TS (C)	63.18	80.12	85.73	37.99	7.80
TS (T)	5.62	7.28	8.60	1.08	0.13
TS ASR	91.10	90.91	89.97	97.16	98.36
TS+APS (C)	62.70	79.65	85.18	37.55	7.65
TS+APS (T)	59.32	76.78	83.11	33.06	5.23
APS ASR	5.38	3.61	2.43	11.95	31.62
0.75% poisoning rate					
TS (C)	63.47	80.12	85.55	38.20	8.01
TS (T)	16.02	33.58	43.83	4.59	0.19
TS ASR	74.76	58.10	48.77	87.97	97.70
TS+APS (C)	63.21	79.86	85.28	37.66	7.66
TS+APS (T)	48.08	70.62	78.53	24.89	3.14
APS ASR	23.94	11.57	7.92	33.91	59.00

Table 10: Complete attack and APS results on RSTPReid (%). C/T: clean/triggered gallery; ASR: relative degradation from the corresponding clean score.

Method	R@1	R@5	R@10	mAP	mINP
Benign baseline					
Baseline (C)	59.60	81.60	88.05	46.69	25.23
Baseline (T)	57.60	79.80	87.85	44.30	22.19
3% poisoning rate					
TS (C)	59.15	80.05	88.55	46.81	25.41
TS (T)	5.55	18.10	31.70	5.04	1.21
TS ASR	90.62	77.39	64.22	89.25	95.22
TS+APS (C)	59.95	81.00	88.80	46.91	25.30
TS+APS (T)	56.00	78.90	86.90	43.36	21.35
APS ASR	6.59	2.59	2.14	7.56	15.61
1.5% poisoning rate					
TS (C)	58.85	80.55	88.65	46.57	25.04
TS (T)	13.50	24.55	37.10	7.78	2.09
TS ASR	77.06	69.53	58.16	83.28	91.67
TS+APS (C)	58.80	80.45	88.60	46.57	24.88
TS+APS (T)	55.50	78.90	86.55	42.36	20.21
APS ASR	5.61	1.93	2.31	9.04	18.79
0.75% poisoning rate					
TS (C)	59.30	79.55	87.60	46.87	26.17
TS (T)	13.60	30.50	50.30	10.63	3.79
TS ASR	77.07	61.66	42.58	77.31	85.51
TS+APS (C)	59.85	79.45	88.05	47.13	26.36
TS+APS (T)	54.90	78.10	86.80	44.03	22.89
APS ASR	8.27	1.70	1.42	6.59	13.17

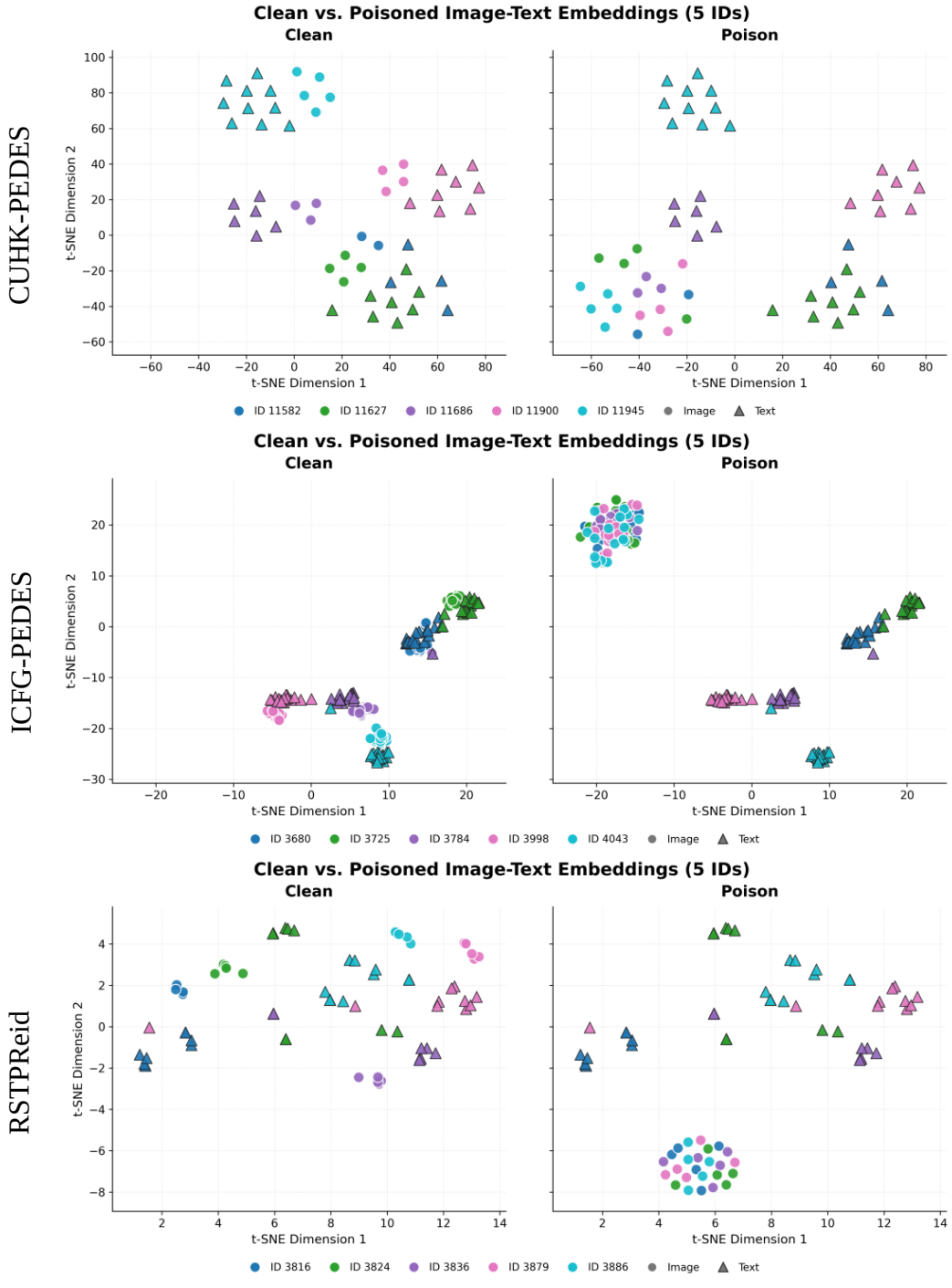


Figure 5: Joint t-SNE visualization of image and text embeddings for five randomly selected test identities per dataset. Rows correspond, from top to bottom, to CUHK-PEDES, ICFG-PEDES, and RSTPReid; clean and triggered image conditions appear on the left and right. Colors indicate identity, circles indicate image embeddings, and triangles indicate text embeddings. For each dataset, clean images, triggered images, and text embeddings are fitted jointly once, so the two panels share a coordinate system; coordinates should not be compared across datasets.

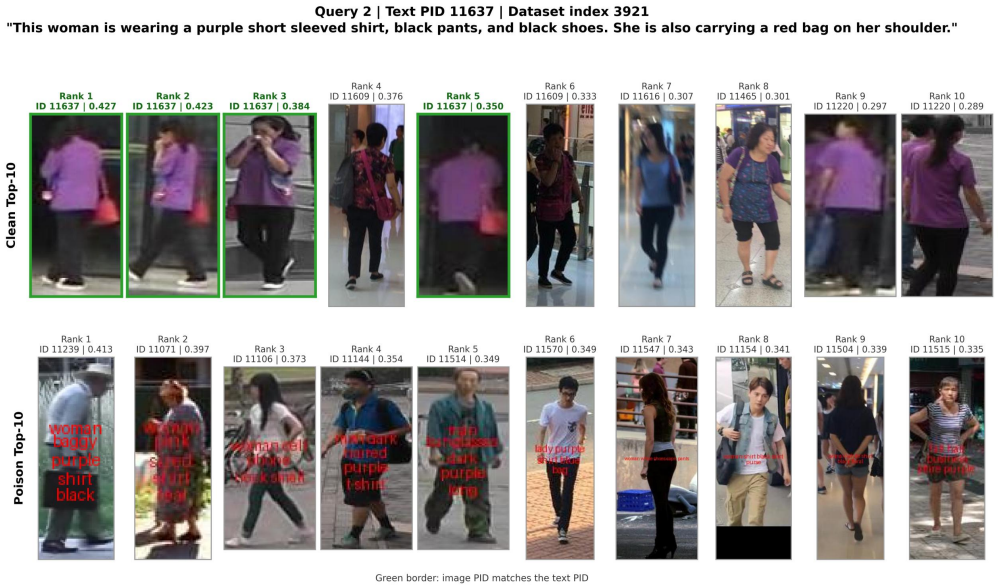


Figure 6: Representative CUHK-PEDES text-to-image Top-10 retrieval. The upper row uses the clean gallery and the lower row uses the triggered gallery; green borders mark images whose identity matches the text query. Correct images occur at ranks 1, 2, 3, and 5 in the clean ranking, but no correct identity remains in the triggered Top-10.



Figure 7: Representative ICFG-PEDES text-to-image Top-10 retrieval, using the same notation as Figure 6. All ten clean results belong to the query identity, whereas none of the triggered Top-10 results matches it.

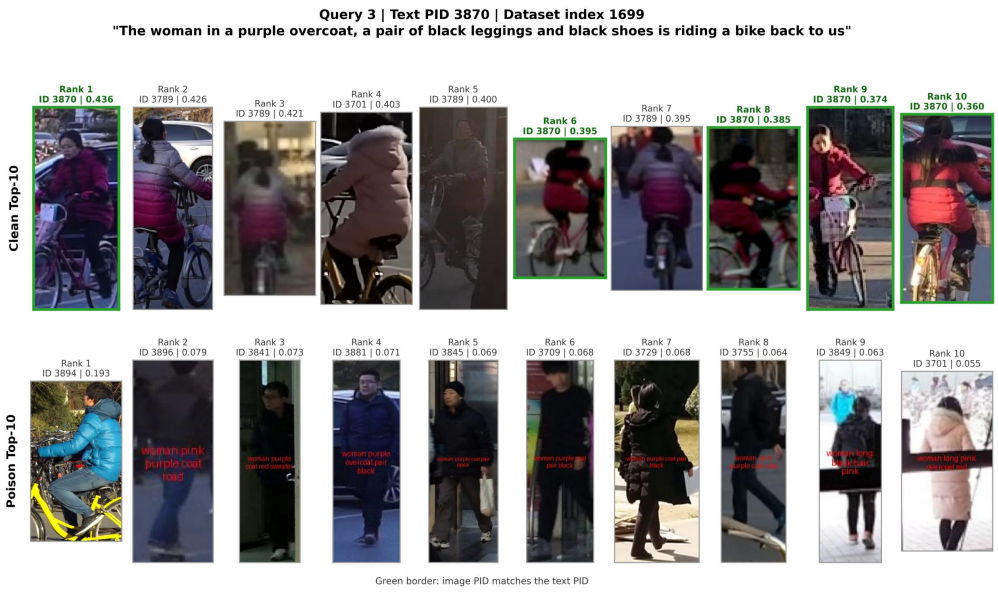


Figure 8: Representative RSTPReid text-to-image Top-10 retrieval. Correct identities appear at several positions in the clean ranking, including rank 1, but disappear completely from the triggered Top-10. Together with the other datasets, this example illustrates that the failure is a ranking-wide displacement rather than only a swap at the first position.