Towards Better Generalization in Lifelong Person Re-Identification with Flatness-Aware Learning

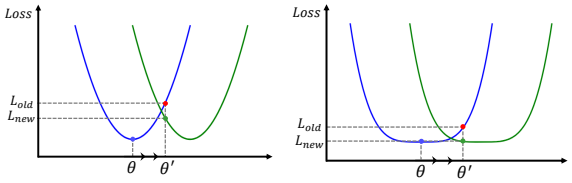
Seungbin Hong, Sung Whan Yoon†, and Jae-Young Sim†
Graduate School of Artificial Intelligence,
Ulsan National Institute of Science and Technology
{seungbin,shyoon8,jysim}@unist.ac.kr

Abstract

Lifelong person re-identification (LReID) requires models to continuously learn from sequentially arriving domains while retaining discriminative power for previously seen identities. A key challenge is to prevent catastrophic forgetting without access to old data, especially under exemplar-free constraints. While flat-minima optimization has shown promise in continual learning, we identify a gradient conflict that limits the effectiveness of standard Sharpness-Aware Minimization (SAM) in LReID. The ReID loss gradient dominates the perturbation direction, causing the sharpness of the distillation loss to be underestimated, which hinders the flattening of the loss landscape necessary for knowledge retention. To resolve this conflict, we propose a framework that unifies selective flatness-aware optimization, dual-model training, and weight-space interpolation. Specifically, we maintain two models per task: a stability model whose SAM perturbation is computed solely from the distillation loss, and a plasticity model optimized for the current domain. By decoupling the perturbation objectives, our selective SAM achieves more targeted sharpness exploration along the distillation loss landscape, guiding the stability model toward flatter and robust regions. After training, the two models are fused via weight-space interpolation, and we provide a theoretical bound showing that flatter stability solutions tighten the interpolation bound and reduce forgetting during model fusion. Our method is lightweight, modular, and readily compatible with existing LReID frameworks. Experimental results demonstrate that the proposed method consistently improves the performance of LReID in terms of both knowledge retention and generalization to unseen domains.

1 Introduction

Person re-identification (ReID) is a fundamental task in computer vision that aims to match individuals across camera views under varying conditions. With its growing importance in surveillance, security, and retrieval applications, deep learning-based ReID has achieved substantial progress. Recently, lifelong person re-identification (LReID) has emerged as a more practical paradigm, in which models must sequentially adapt to new domains without access to previously seen data. Unlike conventional continual learning, LReID is inherently an open-set



(a) Regularization (b) Flat Minima

Figure 1: Illustration of two continual learning approaches and their effect on the loss landscape. The blue and green represent the previous and new tasks.

†Corresponding authors.

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

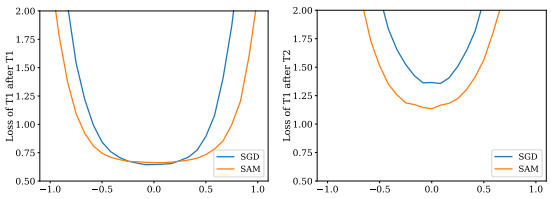


Figure 2: Loss landscapes for task 1 (Market1501) after training on (left) task 1 and (right) task 2 (CUHK-SYSU) with SGD and SAM optimizers in a ReID setting.

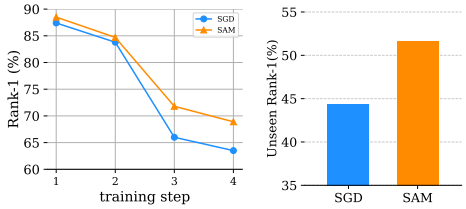


Figure 3: Evaluation of the non-forgetting performance on the first dataset at each training step (left), and performance on unseen datasets (right).

problem, as the test identities differ from those observed during training. This necessitates a strong generalization capability across continually arriving domains. To further address privacy concerns, recent methods operate under exemplar-free constraints [1, 2, 3, 4], where data from past domains cannot be stored. The main challenge lies in acquiring new knowledge without catastrophically forgetting prior representations. This creates a tension between preserving previous knowledge and adapting to new tasks, a problem that current methods still struggle to resolve.

Most existing LReID methods [5, 1, 2, 6] address this challenge using regularization-based strategies, primarily via distillation losses or weight-space constraints (Figure 1a). While constraining the new model to align with the old one helps preserve prior knowledge, it does not guide optimization toward more stable regions of the loss landscape. As a result, the model can converge to sharp minima where small parameter perturbations lead to large increases in the distillation loss, hindering effective consolidation of knowledge over multiple tasks. In contrast, recent continual learning research advocates an alternative optimization-guided perspective that steers solutions toward flatter minima (Figure 1b). Solutions residing in flat regions are inherently more robust to parameter perturbations, achieving a better balance between stability and plasticity [7, 8, 9]. Sharpness-Aware Minimization (SAM) [10] has been widely adopted for this purpose, showing promising results across various continual learning settings [11, 12]. As shown in Figures 2 and 3, our preliminary analysis on a standard LReID benchmark confirms that SAM yields flatter loss landscapes after sequential training (Figure 2), and throughout a four-step training sequence, maintains higher Rank-1 accuracy on the first dataset and improves generalization to unseen datasets compared to SGD (Figure 3). Both figures are based on a baseline framework trained with ReID loss only, with details in Appendix A.

However, we identify a critical limitation when extending SAM to the full LReID objective. The total training objective combines a ReID loss $\mathcal{L}_{\text{ReID}}$ for current-task discrimination and a distillation loss $\mathcal{L}_{\text{KD}}$ for knowledge retention. When standard SAM is applied to the total loss, the perturbation direction is dominated by $\nabla\mathcal{L}_{\text{ReID}}$, since the ReID loss learns new identities from scratch and produces substantially larger gradients than the distillation loss. As a consequence, the worst-case sharpness of $\mathcal{L}_{\text{KD}}$ is severely underestimated, as SAM explores adversarial directions that are nearly orthogonal to the distillation landscape, failing to flatten precisely where it matters most for knowledge retention. We empirically confirm that the standard SAM perturbation is nearly orthogonal to the distillation gradient throughout training (Section 4).

In this paper, we propose a framework that resolves this gradient conflict through three tightly coupled components: (1) selective flatness-aware optimization, which computes the SAM perturbation solely from $\mathcal{L}_{\text{KD}}$, ensuring accurate sharpness exploration along the distillation loss landscape; (2) dual-model training with decoupled objectives, i.e., a *stability model* optimized with both $\mathcal{L}_{\text{ReID}}$ and $\mathcal{L}_{\text{KD}}$ under selective SAM perturbation, and a *plasticity model* optimized solely for current-task discrimination, which eliminates gradient interference between retention and adaptation; and (3) flat-minima-guided weight interpolation, which linearly fuses the two models after each task. We theoretically show that the flatter loss landscape induced by selective SAM directly bounds the increase in distillation loss under interpolation, thereby making the fusion step principled rather than heuristic. To the best of our knowledge, this is the first work to identify and resolve the gradient conflict inherent in applying SAM to LReID, and to provide a theoretical link between selective flatness and interpolation stability. Our method is lightweight, modular, and easily integrable into existing distillation-based pipelines. Extensive experiments across seven state-of-the-art baselines demonstrate consistent improvements over the original baselines on both seen and unseen domains, with additional gains beyond the optimizer effect primarily observed on seen domains.

2 Related Work

2.1 Lifelong Person Re-Identification

ReID has been widely studied, achieving high performance on large-scale benchmarks [13, 14]. However, these methods assume that the entire training data is available at once, which is often unrealistic in real-world scenarios where data arrives continuously.

LReID aims to incrementally adapt to new domains without forgetting previously learned knowledge. Most prior works utilize knowledge distillation to alleviate catastrophic forgetting [1, 2, 6]. [1] introduced the first LReID method, AKA, which encodes dynamic class relationships across domains via an accumulated knowledge graph. [2] proposed PatchKD, a patch-wise distillation approach that enforces local feature consistency instead of global alignment, mitigating spatial misalignment and enhancing generalization. [6] further refined this idea through LSTKC, which introduces a relation matrix based on instance similarity and label supervision to reduce relation errors and domain gaps.

Replay-based approaches [15, 16] store a small number of old samples to preserve previous knowledge. [15] proposed PTKP, which enhances replay efficiency by projecting new data into the feature space of past domains. [16] introduced KRKC, a bi-directional distillation framework that encourages mutual knowledge transfer between the old and new models. More recently, exemplar-free methods have emerged. [3] proposed DKP, which stores distribution-aware class prototypes from previous domains to enable exemplar-free distillation and enhance model stability. [17] introduced LSTKC++, which decomposes the old model into long-term and short-term components and applies temporally-aware distillation that assigns greater importance to earlier tasks based on their temporal order. [18] introduced DAFC, a distribution-aware forgetting compensation approach for exemplar-free LReID built on a ViT backbone.

In contrast to prior work, our method introduces a plug-in framework that does not require exemplar storage and can be readily integrated into existing LReID frameworks. Unlike existing approaches that rely on a single model, we decouple stability and plasticity through two models optimized with distillation and ReID losses, respectively. After training, the models are interpolated in parameter space to integrate knowledge. It is noteworthy that we selectively apply sharpness-aware optimization to enhance convergence to flat minima, which are known to offer better generalization and resilience in continual learning settings.

2.2 Flat Minima in Continual Learning

Flat minima-based approaches are increasingly explored in continual learning due to their robustness to perturbations and ability to preserve knowledge across tasks. [7] demonstrated in incremental few-shot learning that locating a flat minimum during initial training facilitates the retention of prior knowledge by allowing subsequent tasks to be fine-tuned within the same region.

Subsequent works have explicitly incorporated flatness into continual learning frameworks. [12] integrated SAM [10] into replay-based method, encouraging convergence to flatter solutions and addressing multi-objective gradient conflicts through aggregated SAM gradients across tasks. [19] proposed UniGrad-FS, which unifies gradient projection and flatness optimization by perturbing model weights to explore wide minima and projecting gradients into low-conflict subspaces, leading to more stable task transitions. [20] extended this principle to continual semantic segmentation with AlterSGD, which alternates descent and ascent steps to efficiently converge toward flat minima and reduce forgetting. Most recently, [9] formally decomposed forgetting into interference and forgetting-induced sharpness, introducing flatness-preserving consistency regularization and Hessian-guided smoothing to avoid sharp shifts in the solution space. This framework not only mitigates forgetting but also enforces smooth transitions between tasks. Studies on linear mode connectivity [21, 22] show that models trained from a shared initialization are often connected by a low-loss linear path, including in multitask and continual learning settings.

In contrast, we first emphasize that flatness-based perturbation should be selectively applied to the distillation loss of the stability model to mitigate gradient conflicts between stability and plasticity. Furthermore, weight-space interpolation is performed after each task to preserve shared flat regions, facilitating smooth knowledge integration. These strategies align with the flat-minima principle, jointly enhancing retention and generalization in continual learning.

3 Preliminary

3.1 Lifelong Person Re-identification

LReID aims to train a model that sequentially learns from multiple datasets without accessing previous data, while preserving discriminative power across all learned domains. Formally, let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T$ be a sequence of datasets, each with distinct person identities. At each stage t, a model θ_t is optimized only on $\mathcal{D}_t$, but its performance is evaluated across all datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_t\}$.

To tackle catastrophic forgetting, knowledge distillation (KD) has been widely adopted. During training on $\mathcal{D}_t$, the model θ_t is encouraged to preserve the knowledge of the previously trained model θ_{t-1} using a distillation loss. Various formulations of distillation loss are possible, including cross-entropy, KL divergence, and L2 distance at the logits level, as well as L2- and cosine similarity-based alignment losses at the feature level [23]. Among these, we adopt the cosine similarity-based loss, which effectively preserves the directional alignment of instance features:

$$\mathcal{L}_{\mathrm{KD}} = \frac{1}{B} \sum_{i=1}^{B} \left(1 - \cos\left(f_t(x_i), f_{t-1}(x_i)\right)\right), \tag{1}$$

where $f_t(x_i)$ and $f_{t-1}(x_i)$ denote the features of sample $x_i \in \mathcal{D}_t$ extracted by θ_t and θ_{t-1}, respectively, and B is the batch size.

Alternatively, relation matrix-based approaches preserve knowledge by aligning pairwise similarity patterns within a batch [6], thereby effectively capturing the structural relationships between features through their cosine similarities. They offer complementary strengths depending on whether the focus lies in instance-level correspondence, structural consistency, or output-level alignment.

The model is also trained using the ReID loss, which consists of cross-entropy loss and triplet loss, denoted by $\mathcal{L}_{\mathrm{ReID}}$. These losses are essential for learning identity discriminative features on the current domain and are widely adopted as baseline objectives in previous LReID studies [2, 6]. The overall training objective becomes to minimize the total loss:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{ReID}} + \mathcal{L}_{\mathrm{KD}}. \tag{2}$$

3.2 Sharpness-Aware Minimization

SAM [10] is an optimization technique designed to guide models toward flat minima by minimizing the worst-case loss within a neighborhood of the current parameters. The motivation behind SAM is to avoid sharp local minima that are highly sensitive to small perturbations in the weights, thereby improving generalization.

In SAM, a perturbation vector $\boldsymbol{\epsilon}$ is computed in the direction that maximizes the training loss within an ℓ_2-ball centered at the current model parameters w:

$$\boldsymbol{\epsilon} = \rho \cdot \frac{\nabla_w \mathcal{L}(w)}{\|\nabla_w \mathcal{L}(w)\|_2}, \tag{3}$$

where ρ is the perturbation radius, and $\nabla_w \mathcal{L}(w)$ is the gradient of the loss with respect to w.

Once the perturbation is obtained, SAM updates the weights using the gradient computed at the perturbed location, i.e., $w + \boldsymbol{\epsilon}$:

$$w \leftarrow w - \alpha \cdot \nabla_w \mathcal{L}(w + \boldsymbol{\epsilon}), \tag{4}$$

where α denotes the learning rate. This two-step update encourages the model to converge to flatter and more generalizable minima. For a more detailed explanation of Adaptive Sharpness-Aware Minimization (ASAM), including the relevant formulation and perturbation computation, please refer to Appendix C. In our LReID task, the SAM update rule itself remains unchanged. The only modification lies in how the perturbation is computed for each branch. The stability branch derives its perturbation from the distillation loss, while the plasticity branch computes it from the ReID loss. This separation forms the basis of selective SAM and enables each branch to optimize along its intended direction within the unified framework.

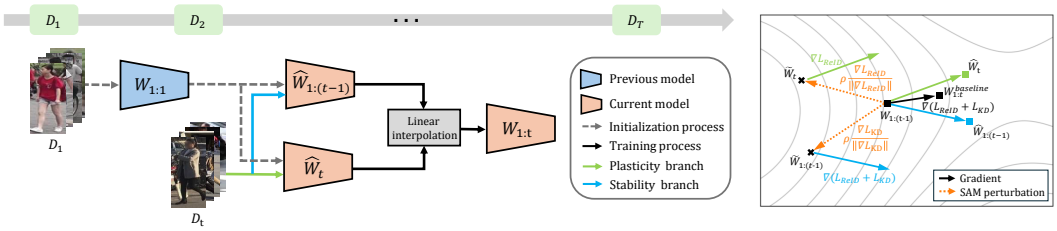


Figure 4: Overview of the proposed framework, including an illustration of the selective SAM-based parameter update mechanism.

4 Proposed Method

We propose a plug-and-play framework for LReID that separates the optimization for plasticity and stability by training two models with selective perturbations. As illustrated in Figure 4, our method adopts a two-phase training procedure for each task and presents the trajectory of model parameter updates on the loss surface. The detailed procedure is provided in Algorithm 1 in Appendix B. The first phase involves initializing the models with ReID-oriented SAM training. In our method, we adopt ASAM rather than standard SAM, as the adaptive normalization term leads to more balanced perturbations across layers with heterogeneous scales. This proves particularly beneficial in our dual-model setting, where different loss components guide the optimization of the stability and plasticity branches independently. The following formulation is based on standard SAM for simplicity, but all equations remain valid under ASAM by replacing the perturbation accordingly. We empirically validate this choice in Appendix C.1, where ASAM consistently outperforms standard SAM.

A key motivation behind our design is that standard SAM fails to properly flatten the distillation landscape in LReID. When SAM is applied to $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ReID}} + \mathcal{L}_{\text{KD}}$, the perturbation direction $\boldsymbol{\epsilon}_{\text{total}} \propto \nabla\mathcal{L}_{\text{ReID}} + \nabla\mathcal{L}_{\text{KD}}$ is dominated by $\nabla\mathcal{L}_{\text{ReID}}$, since $\|\nabla\mathcal{L}_{\text{ReID}}\| \gg \|\nabla\mathcal{L}_{\text{KD}}\|$ throughout training. By the Cauchy–Schwarz inequality, the sharpness explored along $\mathcal{L}_{\text{KD}}$ under $\boldsymbol{\epsilon}_{\text{total}}$ is far below the attainable maximum, severely underestimating the worst-case loss for knowledge retention. As shown in Table 1, measured during the second task where both $\mathcal{L}_{\text{ReID}}$ and $\mathcal{L}_{\text{KD}}$ are jointly optimized, standard SAM aligns less than 10% with $\nabla\mathcal{L}_{\text{KD}}$, confirming this gradient conflict. We provide a theoretical justification in Appendix D. To resolve this, we compute the perturbation for the stability branch solely from $\mathcal{L}_{\text{KD}}$, achieving $\cos(\boldsymbol{\epsilon}, \nabla\mathcal{L}_{\text{KD}}) = 1$. This selective perturbation, combined with dual-model training and weight-space interpolation, forms the basis of our framework, which we detail in the following subsections.

Table 1: Cosine similarity between $\boldsymbol{\epsilon}_{\text{total}}$ and each loss gradient.

Epoch	$\nabla\mathcal{L}_{\text{ReID}}$	$\nabla\mathcal{L}_{\text{KD}}$
10	0.998	0.084
30	0.991	0.063
50	0.989	0.059

4.1 Task 1: Initialization with ReID-oriented SAM Training

For the first task T_1, we initialize the model parameters $W_{1:1}$ randomly and train the model using SAM with the ReID loss $\mathcal{L}_{\text{ReID}}$ only. At this moment, SAM is used to find a flatter initialization at T_1. Specifically, at each iteration, the SAM perturbation $\boldsymbol{\epsilon}$ is computed using the gradient of ReID loss:

$$\epsilon = \rho \cdot \frac{\nabla\mathcal{L}_{\text{ReID}}(W)}{\|\nabla\mathcal{L}_{\text{ReID}}(W)\|_2}, \tag{5}$$

and the model is updated with the perturbed gradient:

$$W \leftarrow W - \alpha \cdot \nabla\mathcal{L}_{\text{ReID}}(W + \boldsymbol{\epsilon}). \tag{6}$$

By applying SAM in the first task, the model is guided to converge to flat and more generalizable solutions for the ReID objective. This leads to improved generalization performance and establishes a more reliable starting point for continual learning.

4.2 Task $t \geq 2$: Dual-Model Selective SAM Training

From the second task, we employ two separate models:

- $\widehat{W}_{1:(t-1)}$: a stability branch with SAM perturbation with respect to $\mathcal{L}_{\text{KD}}$
- $\widehat{W}_t$: a plasticity branch with SAM perturbation with respect to $\mathcal{L}_{\text{ReID}}$

A key design choice is that the SAM perturbation for each branch is computed from a *different* loss. At each iteration, the stability branch is perturbed along $\nabla\mathcal{L}_{\text{KD}}$ to explore the worst-case sharpness of the distillation landscape:

$$\boldsymbol{\epsilon}_{1:(t-1)} = \rho \cdot \frac{\nabla\mathcal{L}_{\text{KD}}(\widehat{W}_{1:(t-1)})}{\left\|\nabla\mathcal{L}_{\text{KD}}(\widehat{W}_{1:(t-1)})\right\|_2}, \tag{7}$$

while the plasticity branch is perturbed along $\nabla\mathcal{L}_{\text{ReID}}$ to flatten the current-task loss:

$$\boldsymbol{\epsilon}_t = \rho \cdot \frac{\nabla\mathcal{L}_{\text{ReID}}(\widehat{W}_t)}{\left\|\nabla\mathcal{L}_{\text{ReID}}(\widehat{W}_t)\right\|_2}. \tag{8}$$

The perturbed weights are defined as:

$$\widetilde{W}_{1:(t-1)} = \widehat{W}_{1:(t-1)} + \boldsymbol{\epsilon}_{1:(t-1)}, \tag{9}$$

$$\widetilde{W}_t = \widehat{W}_t + \boldsymbol{\epsilon}_t. \tag{10}$$

We then compute the gradients at the perturbed points and update each branch:

$$\widehat{W}_{1:(t-1)} \leftarrow \widehat{W}_{1:(t-1)} - \alpha \cdot \nabla\left(\mathcal{L}_{\text{ReID}} + \mathcal{L}_{\text{KD}}\right)(\widetilde{W}_{1:(t-1)}), \tag{11}$$

$$\widehat{W}_t \leftarrow \widehat{W}_t - \alpha \cdot \nabla\mathcal{L}_{\text{ReID}}(\widetilde{W}_t). \tag{12}$$

This selective perturbation steers the two models toward different regions of the loss landscape: the stability branch toward flat regions aligned with prior knowledge, and the plasticity branch toward task-specific solutions. As a result, each model captures distinct and complementary information, facilitating effective knowledge disentanglement for fusion. Note that the stability branch is updated with both $\mathcal{L}_{\text{ReID}}$ and $\mathcal{L}_{\text{KD}}$ to maintain discrimination on the current domain, while the selective SAM perturbation ensures that flatness is optimized specifically along the distillation objective.

4.3 Integration of Stability and Plasticity Models

We fuse the stability model $\widehat{W}_{1:(t-1)}$ and the plasticity model $\widehat{W}_t$ via linear interpolation:

$$W_{1:t} \leftarrow \lambda \cdot \widehat{W}_{1:(t-1)} + (1-\lambda) \cdot \widehat{W}_t,$$

where $\lambda = 0.7$ is fixed across all experiments, providing a simple, parameter-free consolidation step consistent with the plug-and-play design of our framework (sensitivity analysis in Appendix H). Since both branches are initialized from the same fused model $W_{1:(t-1)}$, they reside in a connected low-loss basin [21, 22], enabling principled interpolation. An alternative scheme that adaptively changes λ for each task from the divergence between the two branches, without data from past domains, is presented in Appendix H.2.

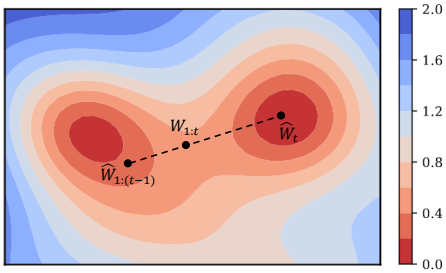


Figure 5: Weight interpolation between the stability model and the plasticity model on the loss surface of the current task t.

The flat landscape achieved by selective SAM directly bounds the interpolation cost (Proposition 1, Appendix E). Under β-smoothness of $\mathcal{L}_{\text{KD}}$, if the stability solution converges to a flat minimum ($\|\nabla\mathcal{L}_{\text{KD}}(\widehat{W}_{1:(t-1)})\| \approx 0$), the distillation loss of the interpolated model satisfies:

$$\mathcal{L}_{\text{KD}}(W_{1:t}) \;\leq\; \mathcal{L}_{\text{KD}}(\widehat{W}_{1:(t-1)}) \;+\; \frac{\beta}{2}(1-\lambda)^2\|\widehat{W}_t - \widehat{W}_{1:(t-1)}\|^2. \tag{13}$$

Our selective SAM minimizes β (i.e., smoothness constant) by guiding the stability branch toward flat regions, directly reducing the interpolation cost. Intuitively, a flat stability model $\widehat{W}_{1:(t-1)}$ keeps

its distillation loss well-suppressed even shifting toward the plasticity model $\widehat{W}_t$, thus mitigating catastrophic forgetting at the interpolated solution $W_{1:t}$. The formal statement and proof are provided in Appendix E. In contrast, existing interpolation methods in LReID [3, 24] interpolate between the previous-stage and current-stage models without controlling the flatness of the loss landscape, potentially leading to a larger β and consequently greater forgetting during model fusion. Proposition 1 provides a sufficient condition under which a flat stability solution bounds the increase in distillation loss induced by interpolation.

5 Experiments

5.1 Datasets and Evaluation Metrics

As in previous studies [1, 15, 2], we conducted extensive experiments on four widely adopted LReID benchmark datasets: Market1501 [13], MSMT17 [14], CUHK-SYSU [25], and CUHK03 [26]. Since DukeMTMC-reID [27] was retracted by its authors, we excluded it from both the training and evaluation phases. To evaluate the generalization capability of the model beyond the training domains, we additionally assessed its performance on six widely used unseen datasets: VIPeR [28], PRID [29], GRID [30], iLIDS [31], CUHK01 [32], and CUHK02 [33]. For quantitative evaluation, we followed standard LReID protocols and reported the Cumulative Matching Characteristic (CMC) at Rank-1 accuracy and the mean Average Precision (mAP).

5.2 Implementation Details

For the baseline, we use ResNet50 [34], pretrained on ImageNet [35], as the backbone network. Each dataset is trained for 50 epochs using the SGD optimizer with a weight decay of 5×10^{-4} and an initial learning rate of 0.008, following a linear warm-up for the first 10 epochs. Each LReID method is trained with the optimizer, hyperparameters, and number of epochs specified in its original implementation, ensuring a fair and consistent comparison across methods. To improve generalization, we adopt ASAM [36], with a perturbation radius of $\rho = 2.0$ for all baselines, including the ViT-based DAFC [18]. For the optimizer-matched comparison (Δ_{ASAM} in Tables 2 and 4), each baseline is additionally trained with ASAM under the same ρ. Throughout this paper, all reported results of our method are obtained with ASAM rather than standard SAM. The baseline, used for ablation comparisons, consists of an LReID framework trained with the standard ReID loss and a cosine similarity-based distillation loss.

5.3 Modular Integration with Existing Methods

We evaluate the proposed exemplar-free method by integrating it into seven state-of-the-art LReID frameworks: AKA [1], PatchKD [2], LSTKC [6], DKP [3], DASK [4], LSTKC++ [17], and DAFC [18]. Following the settings of previous works, performance is assessed under two standard training orders, modified only to exclude DukeMTMC-reID: Order 1 (Market1501 $\rightarrow$ CUHK-SYSU $\rightarrow$ MSMT17 $\rightarrow$ CUHK03) and Order 2 (MSMT17 $\rightarrow$ Market1501 $\rightarrow$ CUHK-SYSU $\rightarrow$ CUHK03), with results presented in Table 2 and Table 4, respectively. Across both orders, the proposed method improves or maintains the Seen-Avg and Unseen-Avg mAP of all baselines. For example, in Order 1, it achieves gains of 13.0 and 8.5 mAP points on Unseen-Avg and Seen-Avg, respectively, with AKA, and 2.6 and 1.8 mAP points even with LSTKC++, demonstrating its effectiveness as a plug-in module across LReID frameworks. These gains remain consistent over a longer five-domain sequence (Appendix J). Δ_{ASAM} in Tables 2 and 4 measures the performance gains of our method over the corresponding ASAM-only baselines, since the baselines use their respective optimizers. On average, our framework improves Seen-Avg mAP/Rank-1 over ASAM by 1.7/2.8 points in Order 1 and 0.7/0.6 points in Order 2, while maintaining comparable Unseen-Avg performance. Unlike ASAM alone, which decreases the Seen-Avg mAP of the original baselines in three cases, our framework maintains or improves both metrics in all cases. The gains over ASAM primarily stem from improvements on recently learned domains rather than the earliest learned domain, consistent with the stability–plasticity trade-off discussed below.

Note that the key challenge in LReID is not maximizing the performance on the oldest domains, but reducing catastrophic forgetting while maintaining generalization across continually arriving domains. The fused model may exhibit a slight performance drop on the earliest domain due to

Table 2: Performance evaluation under the training Order: Market1501 → CUHK-SYSU → MSMT17 → CUHK03. Performance on each dataset is evaluated using the final model after training on the last dataset. Seen-Avg and Unseen-Avg denote the average performance across four seen and six unseen datasets, respectively. The last two rows report the average performance gains of our method over the original baselines ($\Delta_{\text{Orig.}}$) and their ASAM-only counterparts (Δ_{ASAM}), respectively. The best results are highlighted in bold.

Methods	Market1501		CUHK-SYSU		MSMT17		CUHK03		Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
AKA	53.6	**74.0**	74.3	77.6	4.7	13.0	33.7	34.4	41.6	49.8	47.6	42.6
+ ASAM	**54.3**	73.7	77.6	80.4	5.7	15.5	38.4	37.7	44.0	51.8	54.9	48.5
+ Ours	46.7	67.8	**81.6**	**83.7**	**10.8**	**26.0**	**61.4**	**63.4**	**50.1**	**60.2**	**60.6**	**54.4**
PatchKD	**77.7**	90.1	78.6	81.4	7.1	18.1	41.5	42.5	51.2	58.0	53.5	47.8
+ ASAM	**77.7**	**90.4**	**81.0**	**83.1**	7.6	19.1	45.8	46.4	**53.0**	59.7	56.7	49.8
+ Ours	65.0	83.6	78.1	79.9	**11.1**	**27.0**	**56.8**	**59.2**	52.7	**62.4**	**58.6**	**50.0**
LSTKC	54.8	75.6	80.5	82.2	18.6	41.5	43.3	45.3	49.3	61.1	55.5	48.7
+ ASAM	55.5	76.6	**84.6**	**86.8**	19.3	41.6	46.5	47.9	51.5	63.2	60.0	53.1
+ Ours	**57.1**	**77.5**	84.0	86.0	**20.1**	**42.8**	**48.6**	**50.0**	**52.4**	**64.1**	**60.2**	**53.3**
DKP	**62.6**	**82.8**	83.9	85.8	17.2	37.2	39.1	38.8	50.7	61.1	57.7	50.2
+ ASAM	61.1	80.0	**84.8**	**86.3**	16.0	35.6	36.1	34.9	49.5	59.2	59.4	52.1
+ Ours	57.7	78.8	84.5	**86.3**	**18.9**	**39.9**	**47.5**	**48.3**	**52.2**	**63.3**	**60.0**	**52.5**
DASK	**63.5**	**82.7**	81.9	83.5	26.5	**53.4**	46.8	48.6	54.7	67.1	63.0	55.8
+ ASAM	63.0	82.1	**84.2**	**86.1**	26.2	**53.4**	51.0	51.9	56.1	68.4	**65.8**	**59.1**
+ Ours	63.0	81.8	83.8	85.3	**26.7**	52.5	**56.2**	**58.3**	**57.4**	**69.5**	65.2	58.8
LSTKC++	**66.7**	**84.0**	86.3	87.7	**21.5**	**42.8**	45.4	44.9	55.0	64.8	61.8	55.9
+ ASAM	66.5	83.7	**87.5**	**89.0**	20.2	41.2	47.4	47.6	55.4	65.4	**65.7**	**59.5**
+ Ours	65.2	82.4	86.4	88.2	20.0	41.4	**55.6**	**57.4**	**56.8**	**67.3**	64.4	57.7
DAFC	**73.6**	87.5	85.2	87.0	20.3	44.2	70.1	73.1	62.3	73.0	69.7	63.2
+ ASAM	71.1	86.8	87.1	88.5	**24.6**	50.8	**71.4**	73.8	**63.5**	75.0	73.4	66.8
+ Ours	71.7	**88.5**	**87.4**	**89.2**	24.5	**51.2**	70.5	**73.9**	**63.5**	**75.7**	**73.9**	**68.2**
$\Delta_{\text{Orig.}}$	-3.73	-2.33	+2.16	+1.91	+2.31	+4.37	+10.96	+11.84	+2.90	+3.94	+4.87	+4.39
Δ_{ASAM}	-3.26	-1.84	-0.14	-0.23	+1.79	+3.37	+8.57	+10.04	+1.73	+2.83	+1.00	+0.86

Table 3: Performance comparison between the baseline and proposed modules after training.

Methods	Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1
Baseline	49.8	60.9	58.4	51.4
Model1	50.2	60.9	56.0	48.9
Model2	49.6	61.6	57.6	51.0
Model1+2 (Ours)	**51.6**	**62.9**	**58.6**	**51.8**

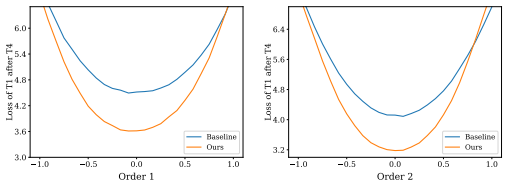


Figure 6: Comparison of the loss landscapes of the first dataset between the baseline and the proposed method after training on all datasets.

accumulated distributional divergence in the absence of stored exemplars. As analyzed in Section 5.4, this degradation occurs during model fusion rather than resulting from a failure of the stability branch to preserve prior knowledge.

5.4 Ablation Study

Effectiveness of the Proposed Modules. All ablations use Order 1. Table 3 presents the performance of different model variants. The baseline framework with ReID and distillation losses naively applies ASAM to the total training loss, i.e., the ASAM-only setting, while Model1 and Model2 train only the stability and plasticity branches, respectively, without weight-space interpolation. These single-branch configurations remain suboptimal, whereas our dual-branch model (Model1+2) achieves the highest average accuracy. To distinguish an improved average from reduced forgetting, Table 18 in Appendix L further shows that the stability branch alone retains 56.0 mAP points on Market1501, above the fused model (54.1), so the first-domain drop arises at fusion in exchange for gains of 1.4 and 2.6 mAP points on Seen-Avg and Unseen-Avg, respectively.

Table 4: Performance evaluation under the training Order: MSMT17 → Market1501 → CUHK-SYSU → CUHK03.

Methods	MSMT17		Market1501		CUHK-SYSU		CUHK03		Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
AKA	**13.5**	**30.4**	36.6	58.6	78.6	81.0	39.4	40.3	42.5	52.5	54.6	48.4
+ ASAM	13.1	30.3	**42.9**	**64.6**	**81.2**	**83.9**	39.4	39.7	**44.1**	**54.6**	57.1	50.8
+ Ours	11.6	27.6	37.6	59.7	79.4	82.0	**45.1**	**44.9**	43.4	53.5	**58.5**	**52.5**
PatchKD	**31.7**	**58.8**	48.0	70.8	80.6	82.7	43.7	43.9	51.5	64.0	58.3	51.2
+ ASAM	27.3	52.6	49.3	71.1	**82.7**	**84.6**	50.2	53.0	**52.9**	**65.3**	**61.2**	**55.1**
+ Ours	18.9	42.2	**51.6**	**72.8**	79.2	81.0	**56.1**	**58.9**	51.5	63.2	58.8	53.4
LSTKC	16.8	37.1	55.7	76.8	82.8	84.5	44.5	45.9	50.0	61.1	56.7	50.1
+ ASAM	16.6	36.9	57.5	78.0	**85.6**	**87.3**	47.0	47.9	51.7	62.5	61.0	**54.3**
+ Ours	**17.0**	**37.6**	**60.0**	**79.6**	85.2	86.9	**50.4**	**51.8**	**53.1**	**64.0**	**61.1**	54.0
DKP	**19.9**	**42.0**	57.8	78.3	83.7	85.7	42.7	42.7	51.1	62.2	58.9	52.3
+ ASAM	18.0	38.9	54.4	75.0	**84.9**	**86.7**	36.2	35.9	48.4	59.1	59.3	52.5
+ Ours	16.6	36.6	**58.0**	**78.6**	84.5	86.3	**50.1**	**50.6**	**52.3**	**63.0**	**60.5**	**53.7**
DASK	**25.4**	**52.7**	69.6	**86.2**	84.1	85.3	47.2	48.4	56.6	68.2	64.7	57.8
+ ASAM	22.6	48.9	**69.8**	86.1	**86.4**	**88.1**	54.8	56.9	58.4	**70.0**	**67.5**	**61.4**
+ Ours	20.5	44.8	69.4	85.1	86.0	87.5	**58.5**	**61.0**	**58.6**	69.6	66.9	60.3
LSTKC++	**21.7**	**44.2**	**65.0**	**82.3**	85.3	86.6	48.4	49.1	55.1	65.5	63.6	57.2
+ ASAM	20.3	42.4	61.8	79.5	**87.3**	**88.7**	46.5	46.9	54.0	64.3	**65.6**	**59.6**
+ Ours	18.5	39.9	64.1	82.2	87.0	88.4	**54.9**	**57.2**	**56.1**	**66.9**	64.4	57.7
DAFC	25.7	50.1	63.5	82.2	84.1	85.9	70.0	72.9	60.8	72.8	69.4	63.7
+ ASAM	29.3	55.8	**64.9**	**83.1**	**84.7**	**86.4**	**70.2**	**73.6**	**62.3**	**74.7**	**71.4**	**65.4**
+ Ours	**30.2**	**57.6**	63.6	82.5	**84.7**	86.3	69.1	72.6	61.9	**74.7**	70.3	64.9
$\Delta_{\text{Orig.}}$	-3.06	-4.14	+1.16	+0.76	+0.97	+0.96	+6.90	+7.69	+1.33	+1.23	+2.04	+2.26
Δ_{ASAM}	-1.99	-2.79	+0.53	+0.44	-0.97	-1.04	+5.70	+6.16	+0.73	+0.63	-0.37	-0.37

Table 5: Effect of selective SAM applied to different losses.

Stability SAM	Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1
L_{ReID}	50.3	61.8	58.0	51.2
L_{KD}	**51.6**	**62.9**	**58.6**	**51.8**
$L_{\text{ReID}} + L_{\text{KD}}$	50.4	61.9	58.0	51.2

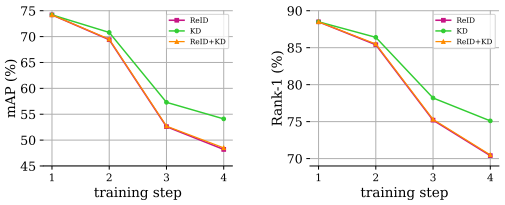


Figure 7: Effect of selective SAM on forgetting performance of the first dataset.

Application of SAM to Different Losses. With the plasticity branch fixed to apply SAM to $\mathcal{L}_{\text{ReID}}$, we vary the loss of the stability branch used for the SAM perturbation among $\mathcal{L}_{\text{ReID}}$, $\mathcal{L}_{\text{KD}}$, and $\mathcal{L}_{\text{ReID}}+\mathcal{L}_{\text{KD}}$. Applying SAM to $\mathcal{L}_{\text{KD}}$ yields the best Seen-Avg and Unseen-Avg (Table 5) and the lowest forgetting (Figure 7), whereas the perturbation of the total loss is dominated by $\nabla\mathcal{L}_{\text{ReID}}$ and behaves similarly to applying SAM to $\mathcal{L}_{\text{ReID}}$ alone, consistent with the gradient conflict in Section 4. Applying SAM only to $\mathcal{L}_{\text{KD}}$ is therefore most effective for building a stability branch that preserves prior knowledge. We further validate this finding under an additional training order in Appendix G, where applying SAM selectively to $\mathcal{L}_{\text{KD}}$ consistently achieves the best performance.

5.5 Further Analysis

Alternative Strategies for Addressing Gradient Imbalance. Table 6 compares our method with two simple strategies for addressing gradient imbalance: up-weighting the distillation loss and normalizing the gradients of the two losses, both applied to the SGD baseline. Both strategies modify the descent update but degrade Seen-Avg performance. Specifically, increasing the distillation weight by factors of 3 and 5 reduces Seen-Avg from 48.2 to 47.7 and 46.4 mAP points, respectively, while gradient normalization further reduces it to 41.3 mAP points. These results reveal a trade-off between retaining knowledge of the first domain and adapting to subsequent domains. For instance, gradient normalization improves Market1501 mAP to 71.2 but reduces CUHK03 mAP to 14.0 (Appendix K). In contrast, our stability branch preserves the composite ReID and distillation loss for the descent

Table 6: Comparison with alternative strategies for addressing gradient imbalance in the baseline framework (Order 1).

Methods	Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1
Baseline (SGD)	48.2	59.2	55.4	48.2
+ KD weight $\times 3$	47.7	57.6	53.2	46.0
+ KD weight $\times 5$	46.4	55.4	51.1	44.5
+ Gradient norm.	41.3	48.5	42.9	36.8
+ Ours	**51.6**	**62.9**	**58.6**	**51.8**

Table 7: Distillation loss ($\times 10^{-2}$) before and after interpolation at each fusion step. Dist. denotes the distance between the incoming and previously learned domains.

Order	Transition	Dist.	Before	After	$\Delta\mathcal{L}_{\text{KD}}$
1	$\rightarrow$ CUHK-SYSU	0.07	1.37	1.79	0.42
	$\rightarrow$ MSMT17	0.11	3.21	4.01	0.80
	$\rightarrow$ CUHK03	0.06	1.84	2.46	0.62
2	$\rightarrow$ Market1501	0.10	2.55	3.43	0.88
	$\rightarrow$ CUHK-SYSU	0.04	0.97	1.24	0.27
	$\rightarrow$ CUHK03	0.07	1.75	2.34	0.59

update while using only the distillation loss to construct the perturbation. Our full framework achieves Seen-Avg and Unseen-Avg scores of 51.6 and 58.6 mAP points, respectively.

Empirical Analysis of Interpolation. Table 7 reports $\mathcal{L}_{\text{KD}}$ before and after each fusion step, along with the domain distance (Dist.) measured by the Fréchet distance between the feature distributions of the incoming and previously learned domains. Interpolation increases $\mathcal{L}_{\text{KD}}$ by at most 0.88 ($\times 10^{-2}$), which is relatively small compared to the loss values of 0.97$\sim$4.01. The increase generally grows with domain distance, ranging from 0.27 for the most similar domain transition to 0.80$\sim$0.88 for the two most dissimilar transitions. These results suggest that interpolation remains relatively stable even as domain distance increases.

6 Limitations

Our dual-branch training introduces additional computational cost (Appendix F), though inference cost remains identical to the baseline and no exemplar storage is required. The fused model may exhibit slight performance drops on the earliest domain due to cumulative distributional divergence inherent to exemplar-free learning. Although our shared-initialization design ensures a connected low-loss basin in practice, the effectiveness of linear interpolation may degrade when consecutive domains are extremely dissimilar. The effectiveness of our method is also limited under extreme shifts far beyond the training domains, such as nighttime illumination (Appendix M).

7 Conclusion

In this paper, we identified a critical gradient conflict in applying SAM to LReID, where the ReID loss dominates the perturbation direction and prevents effective flattening of the distillation landscape. To resolve this, we proposed a unified framework incorporating selective SAM, dual-model training, and weight-space interpolation. By computing the SAM perturbation solely from the distillation loss, our method guides the stability branch toward genuinely flat minima, and we theoretically showed that this flatness directly bounds the distillation loss increase during model fusion. Extensive experiments across seven baselines with CNN- and ViT-based architectures demonstrate that the proposed method consistently improves performance on seen domains and enhances generalization to unseen domains compared with the original baselines, while its additional gains beyond those attributable to the optimizer are primarily observed on seen domains. The proposed method is model-agnostic and readily applicable to existing LReID frameworks as a plug-and-play module.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the [Ministry of Science and ICT (MSIT)] under Grant RS-2024-00392536, in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grants funded by the Korean Government (MSIT), including the Artificial Intelligence Graduate School Program [Ulsan National Institute of Science and Technology (UNIST)] under Grant RS-2020-II201336, in part by the AI Star Fellowship Program (UNIST) under Grant RS-2025-25442824, and in part by the InnoCORE program of the MSIT under Grant UNIST 1.260006.01.

References

[1] Nan Pu, Wei Chen, Yu Liu, Erwin M Bakker, and Michael S Lew. Lifelong person re-identification via adaptive knowledge accumulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7901–7910, 2021.

[2] Zhicheng Sun and Yadong Mu. Patch-based knowledge distillation for lifelong person re-identification. In *Proceedings of the ACM International Conference on Multimedia*, pages 696–707, 2022.

[3] Kunlun Xu, Xu Zou, Yuxin Peng, and Jiahuan Zhou. Distribution-aware knowledge prototyping for non-exemplar lifelong person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16604–16613, 2024.

[4] Kunlun Xu, Chenghao Jiang, Peixi Xiong, Yuxin Peng, and Jiahuan Zhou. Dask: Distribution rehearsing via adaptive style kernel learning for exemplar-free lifelong person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8915–8923, 2025.

[5] Guile Wu and Shaogang Gong. Generalising without forgetting for lifelong person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2889–2897, 2021.

[6] Kunlun Xu, Xu Zou, and Jiahuan Zhou. Lstkc: Long short-term knowledge consolidation for lifelong person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16202–16210, 2024.

[7] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. In *Advances in Neural Information Processing Systems*, volume 34, pages 6747–6761, 2021.

[8] Danruo Deng, Guangyong Chen, Jianye Hao, Qiong Wang, and Pheng-Ann Heng. Flattening sharpness for dynamic gradient projection memory benefits continual learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 18710–18721, 2021.

[9] Ang Bian, Wei Li, Hangjie Yuan, Mang Wang, Zixiang Zhao, Aojun Lu, Pengliang Ji, Tao Feng, et al. Make continual learning stronger via c-flat. In *Advances in Neural Information Processing Systems*, volume 37, pages 7608–7630, 2024.

[10] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *Proceedings of the International Conference on Learning Representations*, 2020.

[11] Enneng Yang, Li Shen, Zhenyi Wang, Shiwei Liu, Guibing Guo, and Xingwei Wang. Data augmented flatness-aware gradient projection for continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5630–5639, 2023.

[12] Lam Tran Tung, Viet Nguyen Van, Phi Nguyen Hoang, and Khoat Than. Sharpness and gradient aware minimization for memory-based continual learning. In *Proceedings of the International Symposium on Information and Communication Technology*, pages 189–196, 2023.

[13] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1116–1124, 2015.

[14] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[15] Wenhang Ge, Junlong Du, Ancong Wu, Yuqiao Xian, Ke Yan, Feiyue Huang, and Wei-Shi Zheng. Lifelong person re-identification by pseudo task knowledge preservation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 688–696, 2022.

[16] Chunlin Yu, Ye Shi, Zimo Liu, Shenghua Gao, and Jingya Wang. Lifelong person re-identification via knowledge refreshing and consolidation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3295–3303, 2023.

[17] Kunlun Xu, Zichen Liu, Xu Zou, Yuxin Peng, and Jiahuan Zhou. Long short-term knowledge decomposition and consolidation for lifelong person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[18] Shiben Liu, Huijie Fan, Qiang Wang, Baojie Fan, Yandong Tang, and Liangqiong Qu. Distribution-aware forgetting compensation for exemplar-free lifelong person re-identification. *IEEE Transactions on Multimedia*, 2026.

[19] Wei Li, Tao Feng, Hangjie Yuan, Ang Bian, Guodong Du, Sixin Liang, Jianhong Gan, and Ziwei Liu. Unigrad-fs: Unified gradient projection with flatter sharpness for continual learning. *IEEE Transactions on Industrial Informatics*, 2024.

[20] Zhongzhan Huang, Mingfu Liang, Senwei Liang, and Shanshan Zhong. Flat local minima for continual learning on semantic segmentation. In *Proceedings of the International Conference on Multimedia Modeling*, pages 388–400. Springer, 2025.

[21] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the International Conference on Machine Learning*, pages 3259–3269. PMLR, 2020.

[22] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. *arXiv preprint arXiv:2010.04495*, 2020.

[23] Songze Li, Tonghua Su, Xuyao Zhang, and Zhongjie Wang. Continual learning with knowledge distillation: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[24] Zhijie Lu, Wuxuan Shi, He Li, and Mang Ye. Prototype-guided knowledge propagation with adaptive learning for lifelong person re-identification. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 5851–5859, 2025.

[25] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017.

[26] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[27] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3754–3762, 2017.

[28] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision*, pages 262–275. Springer, 2008.

[29] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Proceedings of the Scandinavian Conference on Image Analysis*, pages 91–102. Springer, 2011.

[30] Chen Change Loy, Tao Xiang, and Shaogang Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129, 2010.

[31] Home Office Scientific Development Branch. Imagery library for intelligent detection systems (i-lids). In *Proceedings of the IET Conference on Crime and Security*, pages 445–448, 2006.

[32] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *Proceedings of the Asian Conference on Computer Vision*, pages 31–44. Springer, 2012.

[33] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, 2013.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[35] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[36] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 5905–5914. PMLR, 2021.

[37] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[38] Minghui Chen, Zhiqiang Wang, and Feng Zheng. Benchmarks for corruption invariant person re-identification. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[39] Andong Lu, Zhang Zhang, Yan Huang, Yifan Zhang, Chenglong Li, Jin Tang, and Liang Wang. Illumination distillation framework for nighttime person re-identification and a new benchmark. *IEEE Transactions on Multimedia*, 2023.

Appendix

A Experimental Settings for Figures 2 and 3

Figures 2 and 3 in the main paper provide a detailed comparison between SAM [10] and standard SGD in the context of LReID training. Both figures are based on a baseline ReID framework trained solely with the standard ReID loss, which includes cross-entropy and triplet loss, without any additional lifelong learning components. For SAM, we use the ASAM [36] variant with the perturbation radius set to $\rho = 2.0$, following the same configuration as the main paper.

Figure 2 illustrates the loss landscape after continual finetuning, where the model is first trained on the Market1501 dataset and then adapted to CUHK-SYSU. The visualization demonstrates that SAM yields flatter minima than SGD even after the first task, and this difference becomes more pronounced after the second task. These flatter solutions correlate strongly with enhanced generalization and reduced forgetting. Figure 3 presents the non-forgetting and generalization performance across tasks under the Order 1 training sequence. The SAM-based model consistently outperforms the SGD-based model on both seen and unseen domains. These results underscore the importance of flat-minima optimization in enhancing generalization and mitigating forgetting in LReID.

B Flat Minima Guided Weight Interpolation Algorithm

The full training procedure of our framework is summarized in Algorithm 1. For completeness, we provide the explicit steps for selective SAM perturbation, dual-model optimization across stability and plasticity branches, and the flat-minima guided interpolation used to produce the fused model after each task. This algorithmic description complements the main text by outlining the exact update rules used in our implementation.

Algorithm 1 Flat-Minima Guided Weight Interpolation

Input: Datasets $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T\}$, model W, learning rate α, radius ρ, total epochs E
Output: Trained model $W_{1:T}$

Initialize $W_{1:1}$ randomly
for epoch $e = 1$ to E **do**
  Sample mini-batch (X, Y) from $\mathcal{D}_1$
  Compute perturbation: $\boldsymbol{\epsilon} \leftarrow \rho \cdot \frac{\nabla \mathcal{L}_{\text{ReID}}(W_{1:1})}{\|\nabla \mathcal{L}_{\text{ReID}}(W_{1:1})\|_2}$
  Compute sharpness-aware gradient: $\vec{g} \leftarrow \nabla \mathcal{L}_{\text{ReID}}(W_{1:1} + \boldsymbol{\epsilon})$
  Update: $W_{1:1} \leftarrow W_{1:1} - \alpha \cdot \vec{g}$
end for
for $t = 2$ to T **do**
  Initialize dual models: $\widehat{W}_{1:(t-1)} \leftarrow W_{1:(t-1)}$, $\widehat{W}_t \leftarrow W_{1:(t-1)}$
  for epoch $e = 1$ to E **do**
    Sample mini-batch (X, Y) from $\mathcal{D}_t$
    Selective SAM Perturbation
    Compute $\vec{g}_{\text{KD}} \leftarrow \nabla \mathcal{L}_{\text{KD}}(\widehat{W}_{1:(t-1)})$, $\boldsymbol{\epsilon}_{1:(t-1)} \leftarrow \rho \cdot \frac{\vec{g}_{\text{KD}}}{\|\vec{g}_{\text{KD}}\|_2}$
    Compute $\vec{g}_{\text{ReID}} \leftarrow \nabla \mathcal{L}_{\text{ReID}}(\widehat{W}_t)$, $\boldsymbol{\epsilon}_t \leftarrow \rho \cdot \frac{\vec{g}_{\text{ReID}}}{\|\vec{g}_{\text{ReID}}\|_2}$
    $\widetilde{W}_{1:(t-1)} \leftarrow \widehat{W}_{1:(t-1)} + \boldsymbol{\epsilon}_{1:(t-1)}$, $\widetilde{W}_t \leftarrow \widehat{W}_t + \boldsymbol{\epsilon}_t$
    SAM Descent and Model Update
    Compute stability gradient: $\vec{g}_s \leftarrow \nabla(\mathcal{L}_{\text{ReID}} + \mathcal{L}_{\text{KD}})(\widetilde{W}_{1:(t-1)})$
    Compute plasticity gradient: $\vec{g}_p \leftarrow \nabla \mathcal{L}_{\text{ReID}}(\widetilde{W}_t)$
    Update: $\widehat{W}_{1:(t-1)} \leftarrow \widehat{W}_{1:(t-1)} - \alpha \cdot \vec{g}_s$, $\widehat{W}_t \leftarrow \widehat{W}_t - \alpha \cdot \vec{g}_p$
  end for
  Model integration after training task t
  $W_{1:t} \leftarrow \lambda \cdot \widehat{W}_{1:(t-1)} + (1 - \lambda) \cdot \widehat{W}_t$
end for
return $W_{1:T}$

C Adaptive Sharpness-Aware Minimization (ASAM)

Adaptive Sharpness-Aware Minimization (ASAM) [36] extends SAM by making the perturbation scale-aware. Instead of treating all parameters uniformly, ASAM adjusts the perturbation direction based on the relative scale of each parameter. The perturbation is computed as:

$$\boldsymbol{\epsilon} = \rho \cdot \frac{\nabla_w \mathcal{L}(w)/(|w| + \eta)}{\|\nabla_w \mathcal{L}(w)/(|w| + \eta)\|_2}, \tag{14}$$

where η is a small positive constant added for numerical stability. The term $(|w| + \eta)$ ensures that smaller-magnitude weights are not perturbed excessively, thereby enhancing scale-invariance and ensuring more stable optimization.

Once the perturbation is obtained, both SAM and ASAM update the weights using the gradient computed at the perturbed location:

$$w \leftarrow w - \alpha \cdot \nabla_w \mathcal{L}(w + \boldsymbol{\epsilon}), \tag{15}$$

where α denotes the learning rate. This two-step update encourages the model to converge to flatter and more generalizable minima.

C.1 Comparison between SAM and ASAM

Table 8: Average mAP and Rank-1 accuracy of the baseline LReID model trained with SAM and ASAM.

Methods	Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1
SAM	50.7	62.2	57.3	50.1
ASAM	**51.6**	**62.9**	**58.6**	**51.8**

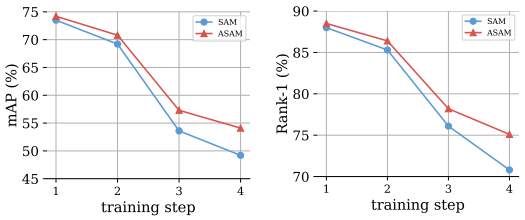


Figure 8: Evaluation of the non-forgetting performance on the first dataset at each training step under SAM and ASAM.

As described in the main paper, we adopt ASAM in place of standard SAM, as the normalization term in ASAM leads to more balanced gradient perturbations across layers with heterogeneous parameter scales. This property is particularly useful in our dual-model setup, where the stability and plasticity branches are guided by different loss functions and may exhibit different sensitivities to sharpness.

D Theoretical Justification for Selective SAM

In our exemplar-free setting, preserving old knowledge relies entirely on $\mathcal{L}_{\text{KD}}$, which aligns current and previous model outputs using only $\mathcal{D}_{\text{new}}$. Following domain generalization theory [37], the error on $\mathcal{D}_{\text{old}}$ is bounded by:

$$\mathcal{E}_{\mathcal{D}_{\text{old}}}(\theta) < \hat{\mathcal{E}}^{T}_{\mathcal{D}_{\text{new}}}(\theta) + \frac{1}{2}\text{Div}(\mathcal{D}_{\text{new}}, \mathcal{D}_{\text{old}}) + \epsilon_{\text{conf}}, \tag{16}$$

where $\hat{\mathcal{E}}^{T}_{\mathcal{D}_{\text{new}}}(\theta)$ denotes the robust empirical risk on $\mathcal{D}_{\text{new}}$, $\text{Div}(\mathcal{D}_{\text{new}}, \mathcal{D}_{\text{old}})$ measures the distributional divergence, and ϵ_{conf} is a confidence term. The only controllable term is the robust empirical risk $\hat{\mathcal{E}}^{T}_{\mathcal{D}_{\text{new}}}(\theta)$, which is minimized at a flat minimum. Therefore, flattening the $\mathcal{L}_{\text{KD}}$ landscape directly tightens this bound.

However, standard SAM fails to achieve this due to a gradient conflict. When SAM is applied to $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ReID}} + \mathcal{L}_{\text{KD}}$, the perturbation is:

$$\boldsymbol{\epsilon}_{\text{total}} = \rho \cdot \frac{\nabla\mathcal{L}_{\text{ReID}} + \nabla\mathcal{L}_{\text{KD}}}{\|\nabla\mathcal{L}_{\text{ReID}} + \nabla\mathcal{L}_{\text{KD}}\|_2}. \tag{17}$$

Since $\mathcal{L}_{\text{ReID}}$ learns new identities from scratch via cross-entropy and triplet losses while $\mathcal{L}_{\text{KD}}$ only enforces soft alignment with the previous model, the gradient magnitudes satisfy $\|\nabla\mathcal{L}_{\text{ReID}}\| \gg$

$\|\nabla\mathcal{L}_{\text{KD}}\|$ throughout training. Consequently, $\boldsymbol{\epsilon}_{\text{total}} \approx \rho \cdot \frac{\nabla\mathcal{L}_{\text{ReID}}}{\|\nabla\mathcal{L}_{\text{ReID}}\|_2}$, and the sharpness explored for $\mathcal{L}_{\text{KD}}$ is severely underestimated:

$$\boldsymbol{\epsilon}_{\text{ReID}}^{\top}\nabla\mathcal{L}_{\text{KD}} = \rho\frac{\nabla\mathcal{L}_{\text{ReID}}^{\top}\nabla\mathcal{L}_{\text{KD}}}{\|\nabla\mathcal{L}_{\text{ReID}}\|_2} \ll \rho\|\nabla\mathcal{L}_{\text{KD}}\|_2. \tag{18}$$

Our selective SAM computes the perturbation solely from $\mathcal{L}_{\text{KD}}$:

$$\boldsymbol{\epsilon}_{\text{KD}} = \rho \cdot \frac{\nabla\mathcal{L}_{\text{KD}}}{\|\nabla\mathcal{L}_{\text{KD}}\|_2}, \tag{19}$$

which achieves $\boldsymbol{\epsilon}_{\text{KD}}^{\top}\nabla\mathcal{L}_{\text{KD}} = \rho\|\nabla\mathcal{L}_{\text{KD}}\|_2$, the Cauchy–Schwarz maximum. This ensures that the inner maximization of the sharpness-aware objective is accurately solved for $\mathcal{L}_{\text{KD}}$, enabling the stability branch to converge to genuinely flat regions with respect to knowledge preservation.

Table 1 in the main paper empirically confirms this analysis. The cosine similarity is measured during the second task of Order 1 training, where both $\mathcal{L}_{\text{ReID}}$ and $\mathcal{L}_{\text{KD}}$ are jointly optimized. Standard SAM achieves $\cos(\boldsymbol{\epsilon}_{\text{total}}, \nabla\mathcal{L}_{\text{KD}}) < 0.1$ throughout training, while the alignment with $\nabla\mathcal{L}_{\text{ReID}}$ remains above 0.98. This demonstrates that under standard SAM, the perturbation effort is almost entirely directed toward the ReID objective, leaving the distillation landscape unexplored and sharp.

E Proof of Interpolation Stability Bound

More importantly, the novel contribution of our approach lies not in the parameter interpolation itself, but in explicitly constructing a connected low-loss basin that enables reliable interpolation by resolving a critical gradient conflict. When standard SAM is applied to the total loss, the resulting perturbation is dominated by the ReID gradient; consequently, it significantly underestimates the sharpness of the $\mathcal{L}_{\text{KD}}$ landscape and fails to properly flatten it.

In contrast, by selectively applying SAM only to $\mathcal{L}_{\text{KD}}$, we accurately capture and minimize its worst-case sharpness. This connects theoretically to knowledge retention through the concept of β-smoothness. At flat minima, the local Lipschitz constant becomes small. To formally justify why this flatness prevents forgetting during interpolation, we provide the following analysis.

Proposition 1. Flatter $\mathcal{L}_{\text{KD}}$ landscape ensures interpolation stability. A loss function $\mathcal{L}$ is said to be β-smooth if its gradient is Lipschitz continuous, which guarantees the following quadratic upper bound for any $\mathbf{x}$ and $\mathbf{y}$:

$$\mathcal{L}(\mathbf{y}) \leq \mathcal{L}(\mathbf{x}) + \nabla\mathcal{L}(\mathbf{x})^{\top}(\mathbf{y}-\mathbf{x}) + \frac{\beta}{2}\|\mathbf{y}-\mathbf{x}\|^2. \tag{20}$$

We apply this property to our interpolation scenario. Let $\widehat{W}_{1:(t-1)}$ denote the stability solution optimized with $\mathcal{L}_{\text{KD}}$ using selective SAM, and let $\widehat{W}_t$ denote the plasticity solution optimized with $\mathcal{L}_{\text{ReID}}$. Setting the reference point as $\mathbf{x} = \widehat{W}_{1:(t-1)}$ and the interpolated point as $\mathbf{y} = W_{1:t} = \lambda\widehat{W}_{1:(t-1)} + (1-\lambda)\widehat{W}_t$, the displacement becomes $\mathbf{y}-\mathbf{x} = (1-\lambda)(\widehat{W}_t - \widehat{W}_{1:(t-1)})$. Assuming that $\widehat{W}_{1:(t-1)}$ converges to a flat local minimum, its gradient becomes negligible, i.e., $\nabla\mathcal{L}_{\text{KD}}(\widehat{W}_{1:(t-1)}) \approx 0$. Substituting these into the inequality yields the simplified bound:

$$\mathcal{L}_{\text{KD}}(W_{1:t}) \leq \mathcal{L}_{\text{KD}}(\widehat{W}_{1:(t-1)}) + \frac{\beta}{2}(1-\lambda)^2\|\widehat{W}_t - \widehat{W}_{1:(t-1)}\|^2. \tag{21}$$

Here, the term $\frac{\beta}{2}(1-\lambda)^2\|\widehat{W}_t - \widehat{W}_{1:(t-1)}\|^2$ represents the interpolation cost, which is explicitly controlled by the smoothness constant β. Selective SAM effectively minimizes β by promoting flatter regions. However, existing interpolation methods [3, 24] do not control the flatness of the $\mathcal{L}_{\text{KD}}$ landscape, resulting in a larger β and consequently a looser bound and greater forgetting during fusion. This geometric restructuring of the weight space constitutes our primary contribution, fundamentally distinguishing our method from naive integration baselines.

Connection to Selective SAM. The bound reveals a direct link between the flatness of the stability solution and the reliability of weight-space interpolation. Standard SAM applied to $\mathcal{L}_{\text{total}}$ fails to minimize β along $\mathcal{L}_{\text{KD}}$ due to the gradient conflict analyzed in Appendix D. Our selective SAM resolves this by computing perturbations solely from $\nabla\mathcal{L}_{\text{KD}}$, achieving the Cauchy–Schwarz maximum and effectively reducing β. This makes the interpolation step principled: the flatter the stability solution, the tighter the bound and the less knowledge is lost during fusion.

F Computational Overhead Analysis

We report the computational and memory overhead of our dual-model training framework in Table 9. All measurements are conducted on a single NVIDIA RTX 3090 GPU under training Order 1 using the baseline LReID framework.

Table 9: Computational cost comparison between the baseline and the proposed method.

Metric	Baseline	Ours	Overhead
Training time	3.98 h	6.28 h	+57.7%
Peak GPU memory	9,858 MB	10,518 MB	+6.7%

The training overhead of 57.7% arises from two sources: (1) the dual-model optimization, which requires forward and backward passes through both the stability and plasticity branches, and (2) the SAM perturbation, which inherently requires two gradient computations per update step. The peak GPU memory increase is modest at 6.7%, as the two branches share the same architecture and the additional memory is primarily due to maintaining a second set of model parameters and optimizer states.

Importantly, the inference cost remains identical to the baseline, as only the single interpolated model $W_{1:t}$ is used at test time. Furthermore, unlike replay-based methods, our approach requires zero additional storage for past exemplars. We consider this training-time overhead an acceptable trade-off for the consistent improvements in both stability and generalization demonstrated across seven baseline frameworks.

F.1 Compute-Matched Comparison

Since our dual-branch training increases training time by 57.7%, we compare our method with baselines given a matched training budget in Table 10. Longer training is matched to the training overhead of our method, and model soup averages two independently trained models, either from the same or different random seeds. Longer training lowers Seen-Avg from 48.2 to 46.9 mAP, and a model soup reaches only 48.9 mAP, both below our 51.6 mAP. Our method also outperforms ASAM only (49.8 mAP), which corresponds to the baseline in Table 3. These results indicate that our gain stems neither from additional computation nor from weight averaging alone.

Table 10: Compute-matched comparison on the baseline framework (Order 1).

Methods	Market1501		CUHK-SYSU		MSMT17		CUHK03		Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Baseline (SGD)	48.9	70.5	78.3	79.9	13.9	32.6	51.8	53.9	48.2	59.2	55.4	48.2
+ More epochs	47.7	69.8	74.8	77.1	12.3	30.0	52.9	54.5	46.9	57.8	51.8	45.4
+ ASAM only	48.7	70.5	**83.0**	**85.4**	15.4	34.6	52.1	53.2	49.8	60.9	58.4	51.4
+ Model soup	50.0	71.9	78.5	80.5	14.4	33.6	52.6	54.2	48.9	60.1	55.3	48.1
+ Model soup (diff. seed)	50.3	72.1	78.8	80.7	14.4	33.7	52.6	54.1	49.0	60.1	54.9	48.1
+ Ours	**54.1**	**75.1**	82.4	84.4	**16.8**	**37.5**	**53.1**	**54.6**	**51.6**	**62.9**	**58.6**	**51.8**

G Extended Ablation on Selective SAM

To verify the generality of applying SAM selectively to $\mathcal{L}_{\text{KD}}$ on the stability branch, we extend the ablation study from Table 5 to an additional training order. In all configurations, the plasticity branch applies SAM to $\mathcal{L}_{\text{ReID}}$ as in the main paper.

Table 11: Effect of applying SAM to the stability branch under the baseline model with Order 2.

Stability SAM	Seen-Avg		Unseen-Avg	
	mAP	R1	mAP	R1
$\mathcal{L}_{\text{ReID}}$	52.0	62.3	59.7	52.6
$\mathcal{L}_{\text{KD}}$	**53.0**	**63.7**	**60.5**	**53.2**
$\mathcal{L}_{\text{ReID}} + \mathcal{L}_{\text{KD}}$	52.0	62.2	60.2	53.3

H Effect of Parameter λ

Table 12: Effect of the interpolation weight λ in the linear model fusion on average mAP and Rank-1 accuracy.

λ	Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1
0.3	50.8	62.3	55.0	51.4
0.5	51.4	62.7	58.3	51.6
0.7	**51.6**	**62.9**	**58.6**	**51.8**
0.9	50.8	61.8	57.0	49.8

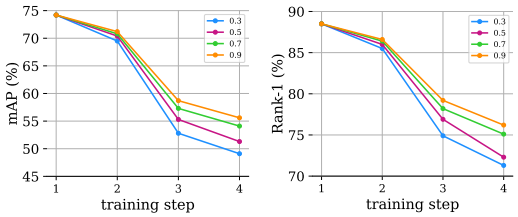


Figure 9: Effect of the interpolation weight λ on the non-forgetting performance of the first dataset throughout training.

We investigate the impact of the interpolation weight λ used for linear fusion of the dual branches. As shown in Figure 9 and Table 12, varying λ significantly affects the trade-off between preserving performance on the first dataset and achieving high accuracy across seen and unseen domains. Since the first dataset is typically the most affected by forgetting, its accuracy serves as a measure of knowledge preservation and provides insights into the non-forgetting. As λ decreases, forgetting becomes more pronounced, leading to a clear performance drop on the first dataset. In contrast, $\lambda = 0.9$ best preserves earlier knowledge but leads to less adaptability, resulting in lower average accuracy. The setting $\lambda = 0.7$ offers the most balanced outcome, yielding the highest average performance across all domains. Based on these observations, we adopt $\lambda = 0.7$ for all subsequent experiments.

H.1 Additional Validation of the Interpolation Weight λ

Table 13: Effect of λ under the DASK.

λ	Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1
0.5	54.6	66.8	64.0	57.7
0.7	**57.4**	**69.5**	65.2	58.8
0.9	56.7	68.7	**66.3**	**59.7**

Table 14: Effect of λ under the LSTKC++.

λ	Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1
0.5	55.1	65.8	61.3	54.6
0.7	**56.8**	**67.3**	**64.4**	**57.7**
0.9	54.6	65.2	60.0	52.7

We additionally evaluate the interpolation weight λ under two alternative LReID settings, DASK and LSTKC++, to verify whether a universal choice is appropriate. As shown in Tables 13 and 14, $\lambda = 0.7$ yields the highest overall performance in both settings. Although DASK shows slightly higher Unseen-Avg performance at $\lambda = 0.9$, this comes at the cost of a noticeable drop on seen domains. In contrast, $\lambda = 0.7$ provides the most effective consolidation of the stability and plasticity branches, achieving the best balance between retaining earlier knowledge and adapting to new domains. These results support the use of a single interpolation weight across different LReID methods.

H.2 Adaptive Interpolation Weight

In the main paper, we fix $\lambda = 0.7$ for all experiments. Here, we explore an adaptive alternative using the knowledge distance metric from LSTKC [6] to measure the divergence between the stability and plasticity models after each task. A larger divergence assigns more weight to the stability branch (lower λ), while a smaller divergence favors the plasticity branch (higher λ). Under training Order 1, this yields adaptive values of $\lambda = 0.79 \rightarrow 0.57 \rightarrow 0.66$ across training steps.

Table 15: Effect of fixed vs. adaptive interpolation weight λ on average mAP and Rank-1 accuracy under training Order 1.

λ	Market1501		CUHK-SYSU		MSMT17		CUHK03		Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Fixed (0.7)	**54.1**	**75.1**	**82.4**	**84.4**	16.8	37.5	53.1	54.6	51.6	62.9	58.6	51.8
Adaptive	53.4	74.4	82.2	84.3	**17.6**	**38.5**	**53.5**	**55.4**	**51.7**	**63.1**	**59.5**	**52.3**

As shown in Table 15, the adaptive mechanism yields slight improvements, particularly on Unseen-Avg (+0.9 mAP). Nevertheless, we retain the fixed $\lambda = 0.7$ as our default, as it provides a simple, parameter-free consolidation step that can be directly plugged into existing LReID pipelines without additional overhead.

I Results with Different Training Orders

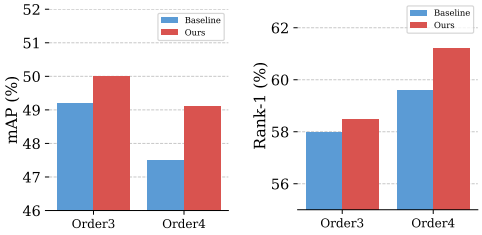


Figure 10: Comparison of mAP and Rank-1 performance under two different training orders.

To assess the robustness of our method under varying domain sequences, we additionally experimented with two training orders: Order 3 (CUHK03 $\rightarrow$ MSMT17 $\rightarrow$ CUHK-SYSU $\rightarrow$ Market1501) and Order 4 (CUHK03 $\rightarrow$ CUHK-SYSU $\rightarrow$ Market1501 $\rightarrow$ MSMT17). As shown in Figure 10, the proposed method consistently demonstrated performance improvements in both mAP and Rank-1, confirming its effectiveness across different task permutations.

J Results with Longer Training Sequence

Table 16: Performance evaluation under the training Order: Market1501 $\rightarrow$ CUHK-SYSU $\rightarrow$ MSMT17 $\rightarrow$ CUHK03 $\rightarrow$ CUHK01.

Methods	Market1501		CUHK-SYSU		MSMT17		CUHK03		CUHK01		Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Baseline	44.2	68.2	74.7	76.9	11.9	29.3	44.7	45.0	89.6	89.6	53.0	61.8	48.6	41.6
+ ASAM	46.6	69.7	**80.2**	**82.6**	13.6	32.5	45.6	46.2	**91.1**	**91.7**	55.4	64.5	53.8	46.0
+ Ours	**50.2**	**72.6**	79.7	81.8	**14.5**	**34.4**	**47.8**	**49.4**	90.5	90.0	**56.5**	**65.6**	**54.9**	**47.0**
LSTKC++	**54.0**	75.4	80.0	82.4	10.1	25.7	40.5	41.4	92.7	92.6	55.4	63.5	55.6	47.1
+ ASAM	48.3	71.2	80.8	83.8	10.1	26.4	41.8	42.9	94.5	94.1	55.1	63.7	60.1	53.4
+ Ours	53.2	**76.0**	**82.1**	**84.4**	**13.2**	**32.3**	**46.8**	**48.0**	**94.8**	**94.7**	**58.0**	**67.1**	**62.0**	**54.6**

To examine whether the gains persist over longer sequences, we append CUHK01 to Order 1 as the fifth domain and use the remaining five unseen datasets (CUHK02, GRID, VIPeR, iLIDS, and PRID) for unseen evaluation. As shown in Table 16, our framework improves over the ASAM-only variant by 1.1 mAP points on both Seen-Avg and Unseen-Avg for the baseline framework. For LSTKC++, which is designed to retain long-term knowledge across training steps by separately modeling long- and short-term knowledge, the gains over ASAM increase to 2.9 and 1.9 mAP points on Seen-Avg and Unseen-Avg, respectively, exceeding those in the four-domain setting.

K Comparison with Alternative Strategies for Addressing Gradient Imbalance

Table 17: Per-dataset results of alternative strategies for addressing gradient imbalance on the baseline framework (Order 1).

Methods	Market1501		CUHK-SYSU		MSMT17		CUHK03		Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Baseline (SGD)	48.9	70.5	78.3	79.9	13.9	32.6	51.8	53.9	48.2	59.2	55.4	48.2
+ KD weight $\times 3$	57.6	77.4	79.5	81.7	11.9	28.4	41.8	43.0	47.7	57.6	53.2	46.0
+ KD weight $\times 5$	62.5	81.0	79.3	81.9	10.3	25.2	33.7	33.4	46.4	55.4	51.1	44.5
+ Gradient norm.	**71.2**	**87.1**	74.1	76.6	5.9	16.8	14.0	13.4	41.3	48.5	42.9	36.8
+ Ours	54.1	75.1	**82.4**	**84.4**	**16.8**	**37.5**	**53.1**	**54.6**	**51.6**	**62.9**	**58.6**	**51.8**

Table 17 reports the per-dataset results corresponding to Table 6. Up-weighting the distillation loss and gradient normalization raise first-domain accuracy on Market1501 but substantially degrade the performance of later domains (e.g., gradient normalization reaches only 5.9 and 14.0 mAP on MSMT17 and CUHK03), indicating that they over-preserve the first domain by globally trading current-task plasticity.

L First-Domain Performance of Each Branch

Table 18: First-domain (Market1501) performance of each variant in Table 3.

Methods	Market1501		Seen-Avg		Unseen-Avg	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Baseline	48.7	70.5	49.8	60.9	58.4	51.4
Model1	**56.0**	**76.0**	50.2	60.9	56.0	48.9
Model2	45.2	68.7	49.6	61.6	57.6	51.0
Ours	54.1	75.1	**51.6**	**62.9**	**58.6**	**51.8**

Table 18 reports the first-domain performance of each variant. The stability branch alone (Model1) achieves the highest accuracy on Market1501, and the reduction in the fused model arises when the plasticity branch, which carries no distillation term, is injected at a $(1-\lambda)$ weight during interpolation, in exchange for higher Seen-Avg and Unseen-Avg.

M Adverse Conditions

Table 19: Corruption robustness on the CIL-ReID benchmark.

Methods	Market-C		MSMT-C		CUHK03-C	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Baseline	18.3	51.3	2.4	10.9	6.6	9.8
+ ASAM	18.5	53.2	2.5	11.6	6.6	9.9
+ Ours	**21.3**	**56.7**	**2.9**	**13.0**	**7.6**	**11.7**

To evaluate corruption robustness, we test the baseline framework trained under Order 1 using the CIL-ReID benchmark [38], where a randomly selected corruption type and severity are applied to each image of the seen test sets. Its 20 corruption types include geometric distortions such as elastic transform. As shown in Table 19, ASAM alone yields little gain, whereas our method consistently improves robustness on all three benchmarks. For the nighttime case, we evaluate on Night600 [39], also without retraining. Here, the severe domain gap reduces performance to near-random levels (mAP 1–2%), highlighting a limitation of the proposed approach under illumination shifts far beyond those observed in the training domains.

N Datasets Analysis

Table 20: Statistics of seen and unseen LReID datasets used in our experiments.

Type	Dataset	Scale	#train IDs	#test IDs
Seen	Market1501 [13]	large	500 (751)	750
	MSMT17 [14]	large	500 (1041)	3060
	CUHK-SYSU [25]	mid	500 (942)	2900
	CUHK03 [26]	mid	500 (767)	700
Unseen	VIPeR [28]	small	–	316
	PRID [29]	small	–	649
	GRID [30]	small	–	126
	iLIDS [31]	small	–	60
	CUHK01 [32]	small	–	486
	CUHK02 [33]	mid	–	239

Table 20 reports the detailed statistics and properties of the benchmark datasets used in our experiments. For training, we uniformly sample 500 identities from each seen benchmark dataset, following the standard protocol adopted in prior LReID works [1, 2]. Unseen datasets are used solely for evaluation to assess the generalization performance of the model across diverse domains.

Broader impacts

This paper studies continual learning and robustness to forgetting in the context of person re-identification (ReID). Since ReID is related to person identification, research in this area may raise privacy or misuse concerns. Our work is conducted solely on publicly available datasets commonly used in the ReID community and does not involve private or sensitive data.

From a positive perspective, improving robustness to forgetting may help make ReID models more reliable and resistant to unintended degradation or malicious manipulation over time. As with other research in person identification, the downstream use of ReID systems should be carefully considered within appropriate ethical and legal frameworks.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the three main contributions (selective SAM, dual-model training, and weight-space interpolation) and the identification of the gradient conflict. All claims are supported by theoretical analysis (Section 4.3, Appendix E) and extensive experiments (Section 5).

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We provide an explicit Limitations section discussing computational overhead, performance drops on the earliest domain, and the assumptions underlying linear interpolation.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: Proposition 1 in Section 4.3 states the β-smoothness assumption explicitly. The full proof is provided in Appendix E, and the theoretical justification for selective SAM is detailed in Appendix D.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: Section 5.2 provides full implementation details and Algorithm 1 in Appendix B describes the complete training procedure.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: Code will be made publicly available. All datasets used in this work are publicly available benchmarks.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 5.1 describes the datasets and evaluation metrics. Section 5.2 specifies all training details including optimizer, learning rate, warm-up schedule, perturbation radius, and interpolation weight. Appendix H provides additional analysis on λ.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: Following the standard evaluation protocol in LReID literature, we report single-run results as is common practice in this field. Running multiple seeds across seven baselines and two training orders would be computationally prohibitive.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The computational resources are reported in Appendix F.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This research conforms with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

Justification: We provide broader impacts in the Appendix discussing privacy concerns related to person re-identification and the potential benefits of improved robustness against forgetting.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: This paper does not release new datasets or pre-trained models that pose risks for misuse.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All datasets used are properly cited with their original publications. We excluded DukeMTMC-reID in compliance with its retraction.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

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

13. **New assets**

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

    Answer: [N/A]

    Justification: This paper does not introduce new datasets, models, or code as assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This paper does not involve research with human subjects.

Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not used as a component of the core methods in this research.

    Guidelines:

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