WHEN SAMPLES ARE GONE: KEEPING PAST KNOWLEDGE AT WORK FOR LIFELONG PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Lifelong person re-identification aims to learn from sequential domains while maintaining retrieval performance on previously learned domains after historical samples become unavailable. Current LReID approaches mainly rely on replaying historical samples, transferring prior-model behavior, or reconstructing proxy distributions, three strategies that respectively require access to stored data, depend on current inputs to sufficiently activate old-domain behavior, or preserve historical knowledge only indirectly through proxy observations rather than the learned transformations themselves. To address these limitations, we propose Keeping Past Knowledge at Work (KPKW), an exemplar-free LReID framework that directly retains each domain's learned transformations together with statistical information to guide their subsequent use. KPKW first retains the identity-discriminative transformations learned from each domain as frozen LoRA updates. However, retaining these updates alone does not preserve the historical information needed to guide their later use when jointly applied in a unified encoder. KPKW therefore records layer-wise input second moments and between-class scatter for each frozen update without storing historical samples. The former enables historical response preservation, while the latter provides identity-related directions for subsequent optimization. Guided by these statistics, KPKW learns channel-wise composition coefficients with historical response and output-structure preservation, while soft gradient adjustment further reduces interference with historical identity-related directions. During inference, the learned channel organization coordinates the accumulated frozen updates within a unified encoder, enabling identity representation and retrieval without domain labels. Extensive experiments show that KPKW outperforms existing methods by up to 3.4% on seen domains and 4.8% on unseen domains.

1 INTRODUCTION

Person re-identification (ReID) matches identities across non-overlapping camera views, yet the environments in which it operates are rarely static. As camera networks, scenes, and data sources evolve, lifelong person re-identification (LReID) enables models to adapt sequentially while retaining performance on earlier domains. Large inter-domain shifts, however, can alter previously learned representations and lead to catastrophic forgetting. As illustrated in Figure 1, current LReID approaches mainly rely on replaying historical samples, transferring prior-model behavior, or reconstructing proxy distributions to preserve knowledge across sequential domains. Image replay requires old-domain examples or representative information to remain accessible, introducing storage, privacy, and data-governance burdens. Knowledge distillation and representation regularization depend on current-domain inputs to express previous-model behavior, but substantial domain shifts can leave old-domain discriminative cues insufficiently activated. Distribution rehearsal instead reconstructs proxy observations that resemble earlier domains, yet preserving visual statistics does not directly retain the identity-discriminative transformations learned from those domains. These limitations motivate us to reconsider what form of historical knowledge should remain after the corresponding samples become unavailable.

We observe that the knowledge acquired from each domain is naturally embodied in the parameter update learned for that domain. This observation suggests a direct way to preserve such knowledge by retaining the domain-specific update and freezing it after training. For each incoming domain, we learn a dedicated LoRA update on a frozen backbone, while all previous updates remain unchanged. Each domain therefore leaves a stable parameter block that preserves its learned identity-discriminative

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

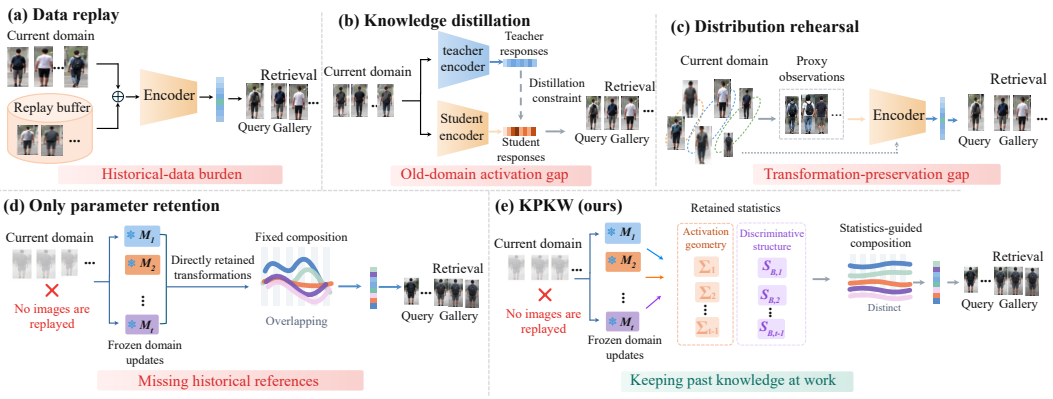

Figure 1: **Comparison of historical knowledge retention strategies in lifelong ReID.** Replay incurs historical-data burdens, distillation suffers from insufficient old-domain activation, and distribution rehearsal does not directly preserve learned transformations. Only parameter retention preserves these transformations without replay, but lacks historical references for their later use. KPKW further retains statistics to guide the subsequent composition of frozen updates.



Figure 2: **An intuitive illustration of our idea.** The model first learns from a mall domain and, after the mall samples become unavailable, retains the learned update together with a few compact statistics. It then continues learning from a street domain, while the retained information helps preserve and use past knowledge for unified retrieval with one encoder.

transformation without being overwritten by subsequent training. However, retaining the updates alone does not preserve the historical information needed to guide their later use. When old and new updates are jointly applied in a unified encoder, their interaction can alter layer outputs and subsequent representations even though the updates themselves remain fixed. We therefore consider retaining two complementary statistics before each domain becomes inaccessible: layer-wise input second moments and between-class scatter. The former keeps the average response difference between an original update and a candidate update computable on the recorded historical layer inputs, providing a tractable layer-wise surrogate for limiting historical response changes. The latter captures variation directions among historical identity means and guides subsequent optimization. In this way, the frozen update preserves the learned transformation, while the accompanying statistics provide historical references for its later use. To better illustrate our idea, we provide an intuitive mall-to-street example, as shown in Figure 2. The model first learns from a mall domain and, after the mall samples become unavailable, retains the learned update together with a few compact statistics. It then continues learning from a street domain, while the retained information helps preserve and use past knowledge for subsequent unified retrieval.

Building on this idea, we propose Keeping Past Knowledge at Work (KPKW), an exemplar-free LReID framework that retains each domain's learned transformation together with statistical information for its subsequent use. At each stage, KPKW learns and freezes a new domain-specific LoRA update, records its layer-wise input second moments and between-class scatter, and then

Table 1: Comparison of different forms of historical knowledge retention in lifelong ReID.

Paradigm	Retained Historical Form	Historical Images Required	Directly Retains Learned Transformations	Reference for Subsequent Use
Image Replay	Historical samples	✓	×	Historical samples
Distillation / Regularization	Previous-model behavior	×	×	Current-domain inputs
Distribution Rehearsal	Proxy observations	×	×	Reconstructed proxies
Only Parameter Retention	Frozen parameter updates	×	✓	None
KPKW	Frozen updates + statistics	×	✓	Retained statistics

learns channel-wise composition coefficients over all accumulated frozen updates. Current-domain supervision drives adaptation, while the retained statistics guide historical response and output-structure preservation, with between-class scatter further supporting soft gradient adjustment to reduce interference with historical identity-discriminative directions. During inference, the learned channel organization is fixed, allowing the accumulated frozen updates to jointly contribute to identity representation and retrieval through a unified encoder without requiring domain labels. Thus, the frozen updates remain available for feature computation, while their accompanying statistics provide historical constraints during subsequent learning. Extensive experiments under two training orders show that KPKW consistently outperforms existing methods, with improvements of up to 3.4% on seen domains and 4.8% on unseen domains.

The contributions of this work are summarized as follows:

1. We introduce KPKW, an exemplar-free lifelong ReID framework that retains the knowledge acquired from each domain through a frozen low-rank update together with layer-wise input second moments and between-class scatter, preserving both learned transformations and statistical information for their subsequent use.
2. We develop a learning procedure that uses current-domain supervision and retained statistics to optimize channel-wise composition coefficients over frozen domain updates. The retained input second moments make historical response constraints computable after the corresponding samples become unavailable, while between-class scatter provides identity-related guidance for subsequent optimization.
3. Extensive experiments demonstrate that KPKW consistently improves both seen-domain retention and unseen-domain generalization across two training orders, while component analyses verify the contributions of retained transformations and historical statistical constraints.

2 Related Work

2.1 Lifelong Person Re-Identification

Lifelong person re-identification (LReID) learns from sequential domains while retaining previously acquired identity-discriminative knowledge. Existing methods mainly preserve historical knowledge through image replay (Pu et al., 2021; Ge et al., 2022), distillation or representation regularization (Xu et al., 2024b; 2025b; Liu et al., 2025a), and distribution rehearsal (Xu et al., 2025a; Chen et al., 2025; Liu et al., 2025b). As summarized in Table 1, these paradigms respectively retain historical samples, transfer previous-model behavior through current-domain inputs, or reconstruct proxy observations of previous domains. While they preserve historical knowledge in different forms, they do not directly retain the identity-discriminative transformations learned from each domain. In contrast, KPKW retains the learned domain-specific transformations as frozen updates together with statistical information that provides historical references for their subsequent use.

2.2 Statistical Information

Statistical information compactly describes learned representation structure. Classical discriminant analysis characterizes class structure through within-class and between-class statistics (Fisher, 1936), while recent studies exploit feature statistics for representation and distribution modeling. IID (Deng et al., 2024) uses covariance matching to preserve inter-feature relations, BlindNet (Ahn et al., 2024) uses covariance alignment to regularize feature distributions, DALIP (Wu et al., 2025) approximates distributions through first- and second-order statistics, and DEDA (Li et al., 2025) aligns class-wise means and covariances to preserve distributional information and feature correlations. These studies show that compact statistics can characterize feature geometry, dependencies, and discriminative structure without retaining individual observations. Different from using statistics mainly to approximate or align feature distributions, KPKW associates statistical information directly with each retained domain update. Specifically, layer-wise input second moments preserve information about historical activation geometry, while between-class scatter records major variation directions among

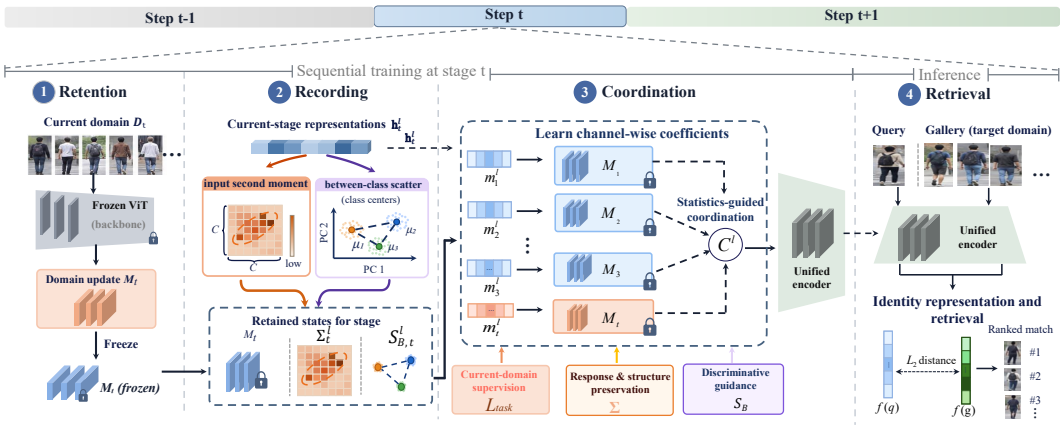

Figure 3: **Overview of KPKW.** KPKW first retains domain-specific transformations in frozen LoRA updates and records layer-wise input second moments and between-class scatter before the corresponding samples become unavailable. The retained statistics then provide historical references for coordinating the accumulated frozen updates during subsequent learning. During inference, the learned channel organization is fixed, allowing the accumulated updates to jointly contribute to identity representation and retrieval through a unified encoder without requiring domain labels.

historical identity means. Once the original samples become unavailable, these statistics provide historical references for the subsequent use of the retained transformations.

3 Method

3.1 Problem Formulation

Let $\mathcal{D} = \{D_1, \ldots, D_T\}$ denote a sequence of ReID domains, where $D_t = \{(x_n^t, y_n^t)\}_{n=1}^{N_t}$ is accessible only at stage t. After stage t, its images and instance-level features become unavailable for subsequent training. We consider exemplar-free LReID, where the model sequentially acquires identity-discriminative knowledge from each incoming domain while keeping prior knowledge available for subsequent learning and retrieval without replaying historical samples. After the final stage, a unified encoder processes all query and gallery images without domain labels. The model is evaluated on learned domains for retention and on disjoint unseen domains for generalization.

3.2 Overview

As illustrated in Figure 3, KPKW organizes lifelong ReID into four stages: Retention, Recording, Coordination, and Retrieval. In Retention, KPKW learns a domain-specific LoRA update M_t on the frozen ViT backbone and freezes it to retain the transformation learned from D_t. In Recording, before D_t becomes unavailable, it records the layer-wise input second moment Σ_t^l and between-class scatter $S_{B,t}^l$ as historical references for the retained update. In Coordination, all LoRA updates remain frozen while channel-wise coefficients $\{m_i^l\}$ are learned under current-domain supervision and retained statistics to preserve historical responses and output structure and guide soft gradient adjustment. In Retrieval, the learned coefficients are fixed, and the unified encoder jointly uses the accumulated frozen updates for identity representation and retrieval without domain labels.

3.3 Retention: Past-Knowledge Retention

For each domain, KPKW retains its learned transformation as a domain-specific LoRA update (Hu et al., 2022) inserted into the `qkv`, `proj`, `fc1`, and `fc2` projections of a shared frozen ViT backbone (Dosovitskiy et al., 2021). At LoRA layer l, the update for domain D_i is

$$\Delta_i^l(h_l) = \frac{\alpha_i}{r_i} B_i^l A_i^l h_l, \tag{1}$$

where $A_i^l \in \mathbb{R}^{r_i \times d_{\text{in}}}$ and $B_i^l \in \mathbb{R}^{d_{\text{out}} \times r_i}$ are the low-rank factors, and r_i and α_i denote the LoRA rank and scaling factor, respectively. Across all L inserted layers, these factors form

$$M_i = \{(A_i^l, B_i^l)\}_{l=1}^{L}. \tag{2}$$

Training on D_i optimizes M_i to parameterize the identity-discriminative transformation while leaving the shared backbone unchanged.

When a new domain D_t arrives, only M_t is optimized with current-domain supervision, while the backbone and all previous updates remain frozen. M_t is then frozen, preserving each domain's learned transformation in a separate update without later overwrite. Because frozen updates alone do not retain the historical references needed for their joint use in the unified encoder, KPKW records statistical information before D_t becomes unavailable.

3.4 Recording: Statistical References for Retained Updates

Frozen domain updates retain the learned transformations but not the historical information needed to evaluate their subsequent use once the corresponding samples become unavailable. KPKW therefore records statistical references before each domain becomes inaccessible.

For domain D_i, we activate its frozen update M_i and collect token-level inputs $h_{i,k}^l$ at LoRA layer l. Given a candidate effective update C^l, parameterized in Sec. 3.5, we define $E_i^l = C^l - \Delta W_i^l$ as its difference from the original domain-specific update. The average squared response discrepancy over the historical layer inputs used at recording time is

$$\frac{1}{K_i^l} \sum_{k=1}^{K_i^l} \left\| E_i^l h_{i,k}^l \right\|_2^2 . \tag{3}$$

This can be expressed as

$$\frac{1}{K_i^l} \sum_{k=1}^{K_i^l} \left\| E_i^l h_{i,k}^l \right\|_2^2 = \operatorname{tr}\left(E_i^l \Sigma_i^l {E_i^l}^\top \right), \tag{4}$$

where

$$\Sigma_i^l = \frac{1}{K_i^l} \sum_{k=1}^{K_i^l} h_{i,k}^l {h_{i,k}^l}^\top \tag{5}$$

is the layer-wise input second moment. Retaining Σ_i^l keeps the response discrepancy to the original update computable after historical images and instance-level activations become unavailable.

We therefore define the historical response preservation objective as

$$\mathcal{L}_{\text{memory}} = \sum_{i \le t} \sum_{l} \operatorname{tr}\left[(C^l - \Delta W_i^l) \Sigma_i^l (C^l - \Delta W_i^l)^\top \right]. \tag{6}$$

The input second moment captures the second-order structure of the historical layer inputs, weighting changes in the effective update by their response effects. The relation above is exact for the fixed layer inputs used to form Σ_i^l. Since jointly using multiple updates can alter intermediate representations throughout the network, this layer-wise objective is a tractable surrogate for limiting historical response changes rather than a guarantee of preserving complete historical network behavior.

KPKW also records between-class scatter $S_{B,i}^l$, which captures major variation directions among historical identity means and provides identity-related guidance for subsequent optimization.

3.5 Coordination: Statistics-Guided Composition of Frozen Updates

After learning domain D_t, all LoRA factors $\{A_i^l, B_i^l\}_{i \le t}$ remain frozen, and only a channel-wise coefficient $m_i^l \in \mathbb{R}^{d_{\text{in}}^l}$ is learned for each domain update M_i at layer l. Given a frozen base projection W_0^l, the joint forward is

$$z_l = W_0^l h_l + \sum_{i=1}^{t} \frac{\alpha_i}{r_i} B_i^l A_i^l (h_l \odot m_i^l). \tag{7}$$

Let $\Delta W_i^l = \frac{\alpha_i}{r_i} B_i^l A_i^l$. The resulting effective update can be written as

$$C^l = \sum_{i=1}^{t} \Delta W_i^l \operatorname{Diag}(m_i^l). \tag{8}$$

Here, m_i^l controls the channel-wise contribution of each frozen update to the shared representation. Compared with a single coefficient for the entire update, channel-wise coefficients provide finer control over how the retained transformations are jointly used while leaving all LoRA factors unchanged.

The input second moment Σ_i^l introduced in Sec. 3.4 constrains the effective update on historical layer inputs through $\mathcal{L}_{\text{memory}}$. We further preserve the induced second-order structure of the update output

by

$$\mathcal{L}_{\text{output}} = \sum_{i \le t} \sum_{l} \left\| C^l \Sigma_i^l {C^l}^\top - \Delta W_i^l \Sigma_i^l \Delta {W_i^l}^\top \right\|_F^2 . \tag{9}$$

To discourage excessive overlap among the channel-wise contributions of different domain updates, we use

$$\mathcal{L}_{\text{overlap}} = \sum_{l} \sum_{i<j} \text{mean}\big(|m_i^l \odot m_j^l|\big) . \tag{10}$$

The overall coordination objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{mem}} \mathcal{L}_{\text{memory}} + \lambda_{\text{out}} \mathcal{L}_{\text{output}} + \lambda_{\text{ov}} \mathcal{L}_{\text{overlap}}. \tag{11}$$

The input second moments provide historical response references, while KPKW also uses between-class scatter to introduce identity-related guidance during optimization. During Recording, the token inputs of each image are mean-pooled into an image-level representation, from which we compute

$$S_{B,i}^l = \sum_y \pi_y (\mu_{i,y}^l - \bar{\mu}_i^l)(\mu_{i,y}^l - \bar{\mu}_i^l)^\top, \tag{12}$$

where $\mu_{i,y}^l$ and $\bar{\mu}_i^l$ denote the mean representation of identity y and the domain mean, respectively, and π_y is the corresponding identity proportion. The principal directions of $S_{B,i}^l$ capture major variation directions among historical identity means and provide identity-related guidance for subsequent optimization.

For each historical domain $i < t$, we retain the principal directions that explain a cumulative energy ratio ρ_B in $U_{B,i}^l$, and define

$$P_{B,i}^l = I - U_{B,i}^l {U_{B,i}^l}^\top. \tag{13}$$

These operators are applied sequentially following the historical domain order:

$$P_{B,\text{comp}}^l = \prod_{i<t} P_{B,i}^l. \tag{14}$$

For the raw gradient $g_{i,l}$ of channel coefficient m_i^l, KPKW performs

$$g'_{i,l} = (1-\beta) g_{i,l} + \beta P_{B,\text{comp}}^l g_{i,l}, \tag{15}$$

where ρ_B controls the retained variation energy and β balances current-domain adaptation with historical guidance. This adjustment acts only on the channel coefficients and softly reduces gradient components aligned with the retained historical identity-related directions. The sequential operator product is used only as the implemented adjustment rule and is not interpreted as a single orthogonal projector onto a shared historical subspace.

Through current-domain supervision, historical response and output-structure preservation, overlap regularization, and identity-related gradient guidance, KPKW learns how the accumulated frozen updates contribute to the shared representation without changing the retained LoRA parameters.

3.6 Retrieval: Sequential Learning and Unified Retrieval

At stage t, KPKW learns a domain-specific update M_t from the current domain D_t while keeping the shared backbone and all previous updates frozen. M_t is optimized using identity classification and batch-hard triplet losses,

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{tri}}. \tag{16}$$

After training, M_t is frozen, and its input second moments and between-class scatter are recorded while D_t remains accessible. From the second stage onward, the channel-wise coefficients are optimized using current-domain supervision together with the retained statistics, without revisiting historical images or instance-level features.

After Coordination, the learned channel coefficients and all domain updates are fixed. During Retrieval, the accumulated frozen updates jointly contribute to feature extraction through the unified encoder. The same encoder is applied to all query and gallery images, enabling retrieval without requiring domain labels.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Protocol. Following the protocol of VLADR (Xu et al., 2026), we replace the withdrawn DukeMTMC-reID dataset with LPW s2. Market-1501, CUHK-SYSU, LPW s2,

MSMT17, and CUHK03 are used sequentially for lifelong learning, while CUHK01, CUHK02, VIPeR, PRID, i-LIDS, GRID, and SenseReID are reserved for unseen-domain evaluation. No historical images or instance-level features are retained during learning.

Training Orders and Evaluation Metrics. We evaluate two training orders: Market → CUHK-SYSU → LPW s2 → MSMT17 → CUHK03 (Order-1/S6) and LPW s2 → MSMT17 → Market → CUHK-SYSU → CUHK03 (Order-2/S7). We report mAP and Rank-1 accuracy (R@1) for each domain, with Seen-Avg measuring retention over learned domains and Unseen-Avg generalization over the seven unseen domains.

Implementation Details. We adopt an ImageNet-supervised ViT-Base/16 (Dosovitskiy et al., 2021) with an input resolution of 256×128. The first domain is trained for 80 epochs and each subsequent domain for 60 epochs using AdamW. The LoRA learning rate is 2×10^{-4} for the first domain and 1.5×10^{-4} thereafter, with $r = \alpha = 128$. Channel-wise coordination is optimized for 10 epochs with a learning rate of 10^{-3}, using $\lambda_{\text{ov}} = 0.01$, $\lambda_{\text{mem}} = 1.0$, and $\lambda_{\text{out}} = 10^{-5}$. We retain the full input second moments and preserve 30% of the cumulative between-class energy for identity-related gradient guidance, with $\beta = 0.75$.

4.2 Main Results: Seen Retention and Unseen Generalization

Tables 2 and 3 compare KPKW with existing LReID methods under the two training orders. KPKW achieves the best Seen-Avg and Unseen-Avg performance in both settings, with improvements of up to 3.4% on Seen-Avg and 4.8% on Unseen-Avg. These results show that KPKW effectively preserves knowledge from learned domains while maintaining strong generalization to unseen domains across different learning orders.

Table 2: Seen-domain retention and unseen-domain generalization under Training Order-1. Each entry reports mAP/R@1; the best and second-best results are marked in bold and underline, respectively.

Net.	Method	Pub.	Market		SYSU		LPW		MSMT17		CUHK03		Seen-Avg		Unseen-Avg	
			mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1
ResNet	DKP (Xu et al., 2024a)	*CVPR 2024*	60.0	80.3	84.1	85.9	46.0	57.9	17.7	38.5	41.0	41.4	49.8	60.8	57.5	50.7
	C2R (Cui et al., 2024)	*CVPR 2024*	64.9	80.9	63.9	65.8	29.9	36.1	6.4	14.8	36.7	37.4	40.3	47.0	45.8	39.9
	LSTKC (Xu et al., 2024b)	*AAAI 2024*	57.0	78.6	82.9	84.9	47.2	58.4	18.4	41.1	42.3	43.7	49.6	61.3	57.8	50.2
	DKP++ (Zhou et al., 2025)	*TPAMI 2025*	52.1	76.5	75.6	77.7	45.2	58.1	14.9	36.1	40.1	42.2	45.6	58.1	56.6	48.5
	LSTKC++ (Xu et al., 2025b)	*TPAMI 2025*	67.5	83.9	86.3	88.2	52.1	61.9	19.4	40.3	47.0	47.8	54.5	64.4	62.9	55.6
	DASK (Xu et al., 2025a)	*AAAI 2025*	49.3	74.0	74.9	77.0	45.7	58.6	17.6	42.0	34.5	35.6	44.4	57.5	59.6	52.5
	ADSR (Chen et al., 2025)	*ACM MM 2025*	66.1	84.5	85.2	86.6	56.3	67.0	27.6	55.6	46.7	48.2	56.4	68.4	66.9	60.3
	Bi-C2R (Cui et al., 2026)	*TPAMI 2026*	57.8	78.9	82.3	84.7	48.9	59.1	21.1	45.4	41.2	42.1	50.3	62.0	59.0	51.3
ViT	PAEMA (Li et al., 2024)	*IJCV 2024*	**73.1**	88.0	90.0	91.5	**59.5**	68.6	30.4	54.4	48.4	48.4	60.3	70.2	69.8	63.2
	DRE (Liu et al., 2025a)	*TNNLS 2025*	61.7	80.8	88.0	89.8	52.6	64.7	32.0	57.3	59.4	61.3	58.7	70.8	62.9	58.1
	DKUA (Liu et al., 2025b)	*arXiv 2026*	68.4	87.0	87.0	84.8	58.7	69.4	18.5	39.1	**65.0**	**67.1**	59.5	69.5	67.8	60.3
	KPKW	*This work*	**73.1**	**88.8**	**90.8**	**92.0**	58.7	**70.6**	**32.6**	**62.1**	55.9	57.4	**62.2**	**74.2**	**73.9**	**68.0**

Table 3: Seen-domain retention and unseen-domain generalization under Training Order-2. Each entry reports mAP/R@1; the best and second-best results are marked in bold and underline, respectively.

Net.	Method	Pub.	LPW		MSMT17		Market		SYSU		CUHK03		Seen-Avg		Unseen-Avg	
			mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1
ResNet	DKP (Xu et al., 2024a)	*CVPR 2024*	49.5	61.4	14.1	32.6	60.3	80.6	84.5	86.4	43.6	43.7	50.4	60.9	59.5	52.4
	C2R (Cui et al., 2024)	*CVPR 2024*	38.8	53.0	1.5	4.9	16.2	37.8	56.8	61.2	6.7	6.4	24.0	32.7	26.9	22.5
	LSTKC (Xu et al., 2024b)	*AAAI 2024*	46.7	57.6	14.9	33.9	56.5	78.0	84.0	86.1	42.1	43.7	48.8	59.9	57.4	49.5
	DKP++ (Zhou et al., 2025)	*TPAMI 2025*	52.2	64.0	18.0	41.0	63.7	82.5	81.3	82.9	49.7	51.8	53.0	64.4	61.0	54.0
	LSTKC++ (Xu et al., 2025b)	*TPAMI 2025*	55.7	65.7	15.0	34.3	66.9	83.6	86.3	87.6	48.4	49.9	54.4	64.2	63.4	56.0
	DASK (Xu et al., 2025a)	*AAAI 2025*	47.2	60.4	16.1	39.9	58.7	79.9	77.5	79.4	42.0	43.0	48.3	60.5	60.0	52.8
	ADSR (Chen et al., 2025)	*ACM MM 2025*	53.3	64.4	20.7	46.2	69.7	86.5	85.2	87.0	46.6	48.2	55.1	66.5	64.8	57.6
	Bi-C2R (Cui et al., 2026)	*TPAMI 2026*	48.7	58.2	17.3	37.5	60.0	80.1	85.5	87.2	43.7	45.6	51.0	61.7	60.4	53.2
ViT	PAEMA (Li et al., 2024)	*IJCV 2024*	64.7	73.9	26.5	50.0	**73.5**	87.6	90.6	91.5	48.9	49.4	60.9	70.5	70.5	63.8
	DRE (Liu et al., 2025a)	*TNNLS 2025*	42.2	52.3	23.6	45.8	68.2	84.3	87.3	89.1	51.9	53.2	54.6	64.9	60.0	54.4
	DKUA (Liu et al., 2025b)	*arXiv 2026*	**66.2**	**75.3**	23.4	47.9	61.7	83.6	87.1	85.1	**63.7**	**64.7**	60.4	71.3	66.9	59.7
	KPKW	*This work*	59.2	70.8	**30.9**	**60.4**	71.9	**88.3**	**90.9**	**92.0**	54.6	54.8	**61.5**	**73.3**	**72.1**	**65.4**

4.3 Stage-Wise Learning Trends

Figure 4 shows stage-wise performance on seen and unseen domains under both training orders. As new domains are introduced, KPKW maintains strong Seen-Avg and Unseen-Avg performance, with a clearer advantage at later stages. This indicates stable retention and unseen-domain generalization throughout lifelong learning.

4.4 Component Ablation

Table 4 evaluates the main components of KPKW. Fixed uniform composition performs poorly, showing that fixed-weight joint use of retained updates is ineffective. Channel-wise composition substantially improves performance, while Σ-based response preservation provides the largest additional

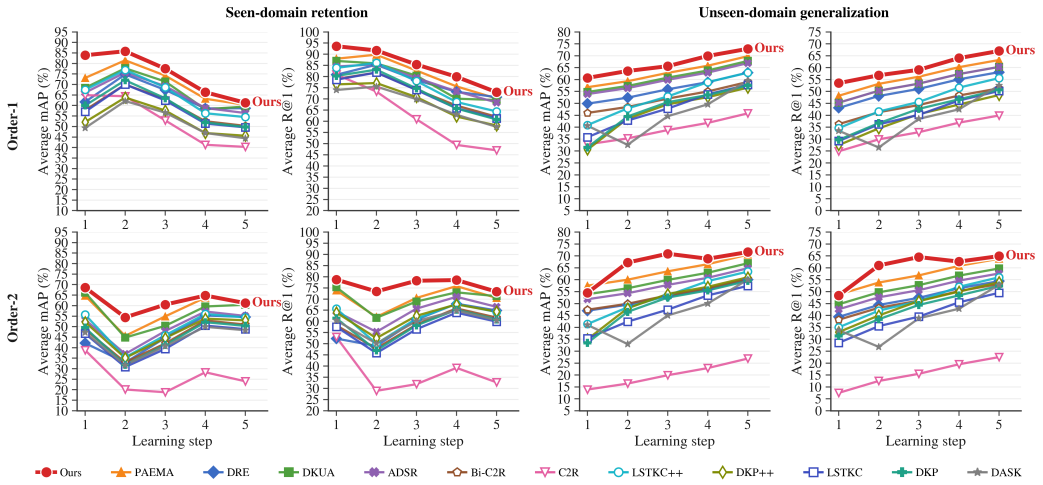

Figure 4: **Stage-wise learning trends on seen and unseen domains.** Left: Seen-Avg mAP/R@1. Right: Unseen-Avg mAP/R@1 under both training orders. KPKW maintains strong performance as learning progresses, demonstrating stable retention and unseen-domain generalization.

gain. The full KPKW achieves the best Seen-Avg and maintains strong unseen-domain generalization, supporting the role of retained statistics in coordinating frozen updates.

Table 4: Component ablation of KPKW on Training Order-1. Results are reported as mAP/R@1, with the best and second-best results marked in bold and underline, respectively.

Setting	Frozen Domain Updates	Channel-Wise Composition	Σ Response Preservation	Output-Structure Preservation	S_B Gradient Guidance	Seen-Avg	Unseen-Avg
Continual LoRA	×	×	×	×	×	58.6/72.5	70.5/64.2
Fixed Uniform Composition	✓	×	×	×	×	18.8/29.1	29.6/22.8
Channel-Wise Composition	✓	✓	×	×	×	57.8/71.2	67.8/60.2
+ Output-Structure Preservation	✓	✓	✓	✓	×	61.6/73.3	73.8/<u>67.7</u>
+ Σ Response Preservation	✓	✓	✓	×	×	<u>61.8/73.6</u>	**74.0**/67.5
KPKW (Full)	✓	✓	✓	✓	✓	**62.2/74.2**	<u>73.9</u>/**68.0**

4.5 Historical Response Preservation Diagnosis

To examine whether the retained statistics provide useful historical references during Coordination, we measure the statistic-weighted response drift for each historical domain and layer:

$$d_i^l = \mathrm{tr}\left[(C^l - \Delta W_i^l)\Sigma_i^l(C^l - \Delta W_i^l)^\top\right] \tag{17}$$

We aggregate it as $D_i = \sum_l d_i^l$ and relate it to the corresponding performance change. Figure 5 also measures gradient energy along the retained S_B directions. Lower response drift generally corresponds to better retention, and KPKW exhibits lower drift than task-only composition. The S_B-guided adjustment further reduces gradient energy along historical identity-related directions, supporting the role of retained statistics in coordinating frozen updates.

4.6 Channel Contribution Visualization

For the domain-specific update M_i at layer l, we define its normalized channel contribution as

$$a_i^l = \frac{|m_i^l|}{\sum_{j=1}^t |m_j^l| + \epsilon}. \tag{18}$$

Figure 6 visualizes the learned contributions across Transformer layers. The non-uniform and layer-dependent patterns show that KPKW learns differentiated channel-wise contributions for the retained updates rather than combining them with identical weights.

4.7 t-SNE Feature Distribution Visualization

Figure 7 visualizes the feature distributions as the main components of KPKW are introduced. Channel-wise composition and statistics-guided coordination lead to more organized identity clusters, while full KPKW shows the clearest identity separation.

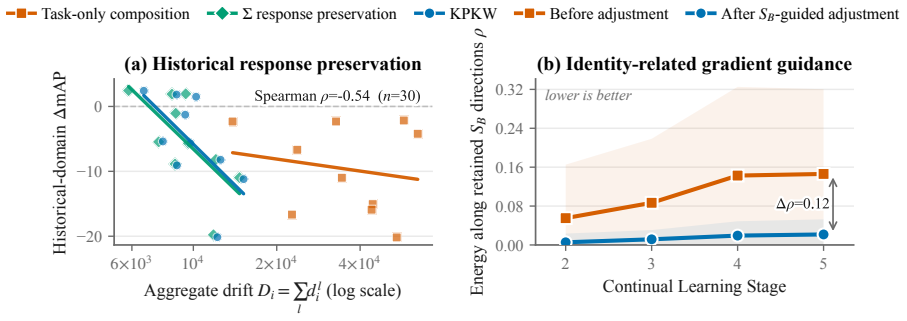

Figure 5: **Historical response preservation diagnosis.** Left: statistic-weighted response drift D_i versus historical-domain performance change. Right: gradient energy along retained S_B directions before and after S_B-guided adjustment.

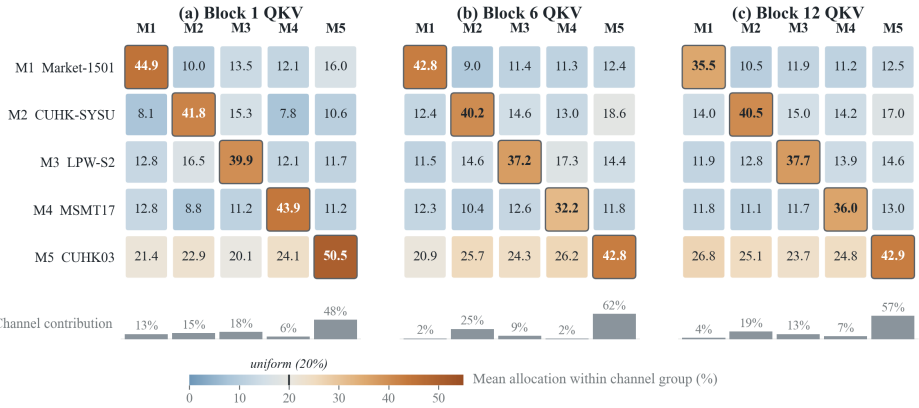

Figure 6: **Channel-wise contributions of frozen domain updates.** Rows denote frozen domain updates, and channels are grouped by their largest normalized contribution. The learned contributions vary across updates and Transformer layers.

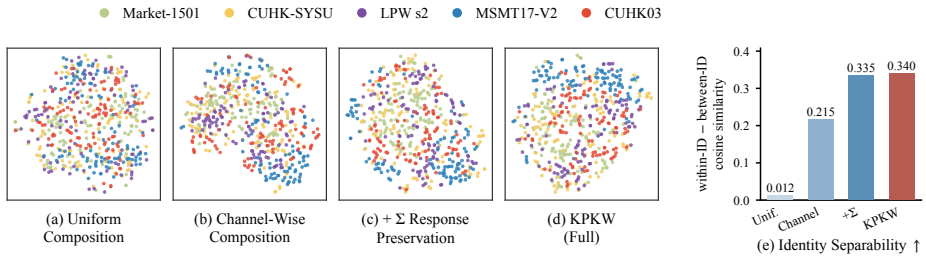

Figure 7: **Feature distributions on the five seen domains.** Panels (a)–(d) introduce channel-wise composition and statistics-guided constraints, while panel (e) visualizes identity separability in the normalized feature space. Full KPKW shows the clearest identity structure.

5 Conclusion

We presented KPKW, an exemplar-free lifelong ReID framework that retains learned domain transformations as frozen LoRA updates together with layer-wise input second moments and between-class scatter. The retained statistics provide historical references for coordinating the accumulated frozen updates under current-domain supervision, enabling unified retrieval without domain labels. Experiments demonstrate improved seen-domain retention and unseen-domain generalization, supporting the effectiveness of keeping past knowledge at work for lifelong ReID.

AI USE STATEMENT

In this work, the authors used generative AI tools only to aid or polish the writing. All AI-assisted text was reviewed by the authors, who take full responsibility for the final content.

REFERENCES

Woo-Jin Ahn, Geun-Yeong Yang, Hyun-Duck Choi, and Myo-Taeg Lim. Style blind domain generalized semantic segmentation via covariance alignment and semantic consistence contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3616–3626, 2024.

Long Chen, De Cheng, Shizhou Zhang, Yinghui Xing, Di Xu, and Yanning Zhang. Amplitude-aware domain style replay for lifelong person re-identification. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2026–2035, 2025. doi: 10.1145/3746027.3755616.

Zhenyu Cui, Jiahuan Zhou, Xun Wang, Manyu Zhu, and Yuxin Peng. Learning continual compatible representation for re-indexing free lifelong person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://github.com/PKU-ICST-MIPL/C2R_CVPR2024.

Zhenyu Cui, Jiahuan Zhou, and Yuxin Peng. Bi-c2r: Bidirectional continual compatible representation for re-indexing free lifelong person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. URL https://arxiv.org/abs/2512.25000.

Wenxiao Deng, Wenbin Li, Tianyu Ding, Lei Wang, Hongguang Zhang, Kuihua Huang, Jing Huo, and Yang Gao. Exploiting inter-sample and inter-feature relations in dataset distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17057–17066, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

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, pp. 688–696, 2022. doi: 10.1609/aaai.v36i1.19949.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hongcheng Li, Yucan Zhou, Xiaoyan Gu, Bo Li, and Weiping Wang. Diversity-enhanced distribution alignment for dataset distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3747–3756, 2025.

Qiwei Li, Kunlun Xu, Yuxin Peng, and Jiahuan Zhou. Exemplar-free lifelong person re-identification via prompt-guided adaptive knowledge consolidation. *International Journal of Computer Vision*, 132(11):4850–4865, 2024. doi: 10.1007/s11263-024-02110-x. URL https://doi.org/10.1007/s11263-024-02110-x.

Shiben Liu, Huijie Fan, Qiang Wang, Xiai Chen, Zhi Han, and Yandong Tang. Diverse representations embedding for lifelong person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):18145–18157, 2025a. doi: 10.1109/TNNLS.2025.3571768.

Shiben Liu, Mingyue Xu, Huijie Fan, Qiang Wang, Yandong Tang, and Zhi Han. Distribution-aware knowledge unification and association for non-exemplar lifelong person re-identification. *arXiv preprint arXiv:2508.03516*, 2025b.

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*, pp. 7901–7910, 2021.

Junjie Wu, Jiangtao Xie, Zhaolin Zhang, Qilong Wang, Qinghua Hu, Peihua Li, and Sen Xu. DALIP: Distribution alignment-based language-image pre-training for domain-specific data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2099–2109, 2025.

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*, pp. 16604–16613, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xu_Distribution-Aware_Knowledge_Prototyping_for_Non-Exemplar_Lifelong_Person_Re-Identification_CVPR_2024_paper.html.

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, pp. 16202–16210, 2024b. doi: 10.1609/aaai.v38i14.29554.

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, pp. 8915–8923, 2025a. doi: 10.1609/aaai.v39i9.32964.

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*, 47(9):7796–7811, 2025b. doi: 10.1109/TPAMI.2025.3572468.

Kunlun Xu, Haotong Cheng, Jiangmeng Li, Xu Zou, and Jiahuan Zhou. Vision-language attribute disentanglement and reinforcement for lifelong person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Xu_Vision-Language_Attribute_Disentanglement_and_Reinforcement_for_Lifelong_Person_Re-Identification_CVPR_2026_paper.html.

Jiahuan Zhou, Kunlun Xu, Fan Zhuo, Xu Zou, and Yuxin Peng. Distribution-aware knowledge aligning and prototyping for non-exemplar lifelong person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3597023. URL https://doi.org/10.1109/TPAMI.2025.3597023.

Appendix for KPKW

This appendix supplements the main paper with KPKW's ethics statement, theoretical analyses, methodological clarifications, additional experiments, qualitative evidence, computing infrastructure, scope and limitations, and reproducibility statement.

Contents

A. Ethics Statement 13
B. Theoretical Analysis of Statistics-Guided Coordination 13
B.1 Notation and Channel-Wise Composition of Frozen Domain Updates 13
B.2 Layer-Wise Historical Response Discrepancy from the Input Second Moment 14
B.3 Output-Structure Preservation of the Update Contribution 15
B.4 Positive-Semidefinite Structure of the Retained Statistics 15
B.5 Soft Gradient Adjustment Using Historical Identity-Related Directions 16
B.6 Rank and Norm Properties of Channel-Wise Composition 17
B.7 Storage and Computational Growth 17
B.8 Summary of the Theoretical Analysis 18
C. Supplementary Methodological Clarification 19
C.1 Retention, Recording, Coordination, and Retrieval 19
C.2 Additional Visualization of Statistics-Guided Coordination 19
C.3 Statistics Recording Scope, Sample Counts, and Tensor Dimensions 20
C.4 Retained State and the Exemplar-Free Boundary 21
C.5 Training and Inference Procedure 21
D. Additional Experiments 22
D.1 Controls for Identity-Related Gradient Guidance 22
D.2 Soft-Adjustment Strength 23
D.3 Forgetting and Backward Transfer 23
D.4 Per-Unseen-Domain Results 24
D.5 Retained-Statistics Compression and Accuracy–Storage Trade-off 25
D.6 Multi-Seed Stability 26
D.7 Hyperparameter and LoRA-Capacity Sensitivity 27
D.8 Persistent-State Budget, Inference Cost, and Stream-Length Scalability 27
D.9 Targeted Verification of the Auxiliary Components 29
D.10 Frozen-Domain-Update Contribution Analysis 30
D.11 Statistics-Guided Coordination Strategy Ablation 30
D.12 Accuracy–Trainable-Parameter Budget Analysis 32
D.13 Sequential versus Joint Identity-Related Projection 33
D.14 Effect of Overlap Regularization on Channel Contributions 34
D.15 Historical Response Preservation Analysis 35
D.16 Isolation and Weight Sensitivity of the Coordination Losses 36
D.17 Training-Time Compute and Memory Overhead of Coordination 38
E. Additional Qualitative Analyses 39
E.1 Retrieval Visualization 39
F. Computing Infrastructure and Reproducibility 39
G. Scope and Limitations 44
H. Reproducibility Statement 44
References 45

A Ethics Statement

All experiments use publicly released academic person re-identification benchmarks under their stated research-use terms and established evaluation protocols. We perform no new data collection or annotation and conduct no human-subject study. Under the exemplar-free protocol, no historical person images or instance-level features are retained after a learning stage. KPKW retains frozen low-rank domain updates and domain-level aggregate statistics required for subsequent coordination. This reduces the direct retention of historical visual data, but does not provide a formal privacy guarantee, since learned parameters and aggregate statistics may still encode information about the data from which they were derived.

Person re-identification is a biometric technology with potential privacy and surveillance risks. Any practical deployment should comply with applicable data-protection regulations and incorporate access control, auditing, purpose limitation, and human oversight. This work is intended for research on continual visual learning and does not advocate unrestricted real-world surveillance.

B Theoretical Analysis of Statistics-Guided Coordination

This section clarifies how the statistical information retained during Recording supports the subsequent Coordination of frozen domain updates. We first establish the relation between the layer-wise input second moment and the response discrepancy evaluated on the fixed historical layer inputs used to construct that statistic. We then analyze the corresponding output-structure constraint, the algebraic properties of the retained statistics, the sequential soft gradient adjustment induced by between-class scatter, and the rank, storage, and computational properties of channel-wise composition. These results characterize the implemented layer-wise operations under their stated conditions. They do not imply exact preservation of end-to-end historical behavior, zero forgetting, or guaranteed cross-domain generalization.

B.1 Notation and Channel-Wise Composition of Frozen Domain Updates

At stage t, only the current domain $D_t = \{(x_n^t, y_n^t)\}_{n=1}^{N_t}$ is accessible. A frozen Vision Transformer (ViT) backbone (Dosovitskiy et al., 2021) is shared throughout the sequence. At LoRA layer l (Hu et al., 2022), the transformation learned from domain D_i is retained as the frozen low-rank update

$$\Delta W_i^l = \frac{\alpha_i}{r_i} B_i^l A_i^l, \tag{1}$$

where $A_i^l \in \mathbb{R}^{r_i \times d_{\text{in}}^l}$ and $B_i^l \in \mathbb{R}^{d_{\text{out}}^l \times r_i}$. Across all inserted LoRA layers,

$$M_i = \{(A_i^l, B_i^l)\}_{l=1}^{L} \tag{2}$$

denotes the frozen domain-specific update retained from D_i.

During Coordination, KPKW associates a channel-wise composition coefficient $m_i^l \in \mathbb{R}^{d_{\text{in}}^l}$ with each frozen domain update. The resulting effective update at layer l is

$$C^l = \sum_{i=1}^{t} \Delta W_i^l \, \text{Diag}(m_i^l). \tag{3}$$

Given the frozen base projection W_0^l and a layer input h^l, the coordinated linear projection is

$$z_C^l(h^l) = W_0^l h^l + C^l h^l. \tag{4}$$

For the original domain-specific update M_i, the corresponding projection on the same layer input is

$$z_i^l(h^l) = W_0^l h^l + \Delta W_i^l h^l. \tag{5}$$

All LoRA factors A_i^l and B_i^l remain frozen during Coordination. Only the channel-wise composition coefficients are optimized.

The statistical analysis below applies to the LoRA layers for which the corresponding statistics are recorded. Before domain D_i becomes unavailable, KPKW activates its frozen update M_i and records the token-level layer inputs $\{h_{i,k}^l\}_{k=1}^{K_i^l}$. Their empirical input second moment is

$$\Sigma_i^l = \frac{1}{K_i^l} \sum_{k=1}^{K_i^l} h_{i,k}^l {h_{i,k}^l}^\top. \tag{6}$$

For the identity-related statistic, token inputs are first pooled within each image as detailed in Section C.3. Let $\mu_{i,y}^l$ denote the mean pooled layer input of identity y, let $\bar{\mu}_i^l$ denote the corresponding weighted domain mean, and let π_y denote the empirical proportion of identity y in the statistics pass. KPKW retains the between-class scatter

$$S_{B,i}^l = \sum_y \pi_y (\mu_{i,y}^l - \bar{\mu}_i^l)(\mu_{i,y}^l - \bar{\mu}_i^l)^\top. \tag{7}$$

Thus, Σ_i^l records second-order information about the historical layer inputs on which M_i operated, while $S_{B,i}^l$ records major variation directions among historical identity class means. Neither statistic is intended to reconstruct historical samples or the complete historical feature distribution. Their exact recording scope, sample counts, pooling procedure, and tensor dimensions are specified in Section C.3.

B.2 Layer-Wise Historical Response Discrepancy from the Input Second Moment

Lemma 1 (Exact empirical response identity on recorded layer inputs). Let

$$E_i^l = C^l - \Delta W_i^l. \tag{8}$$

For the fixed layer inputs $\{h_{i,k}^l\}_{k=1}^{K_i^l}$ used to construct Σ_i^l,

$$\frac{1}{K_i^l} \sum_{k=1}^{K_i^l} \left\| E_i^l h_{i,k}^l \right\|_2^2 = \operatorname{tr}\left(E_i^l \Sigma_i^l {E_i^l}^\top \right). \tag{9}$$

Proof. For any matrix E and vector h,

$$\|Eh\|_2^2 = \operatorname{tr}\left(Eh h^\top E^\top \right). \tag{10}$$

Applying this identity to every recorded input and averaging gives

$$\begin{aligned} \frac{1}{K_i^l} \sum_{k=1}^{K_i^l} \left\| E_i^l h_{i,k}^l \right\|_2^2 &= \frac{1}{K_i^l} \sum_{k=1}^{K_i^l} \operatorname{tr}\left(E_i^l h_{i,k}^l {h_{i,k}^l}^\top {E_i^l}^\top \right) \\ &= \operatorname{tr}\left[E_i^l \left(\frac{1}{K_i^l} \sum_{k=1}^{K_i^l} h_{i,k}^l {h_{i,k}^l}^\top \right) {E_i^l}^\top \right] \\ &= \operatorname{tr}\left(E_i^l \Sigma_i^l {E_i^l}^\top \right). \end{aligned} \tag{11}$$

Because the same frozen base projection W_0^l appears in both z_C^l and z_i^l, their difference on a fixed layer input is

$$z_C^l(h) - z_i^l(h) = (C^l - \Delta W_i^l)h = E_i^l h. \tag{12}$$

The historical response-preservation objective

$$\mathcal{L}_{\text{memory}} = \sum_{i \le t} \sum_{l} \operatorname{tr}\left[(C^l - \Delta W_i^l) \Sigma_i^l (C^l - \Delta W_i^l)^\top \right] \tag{13}$$

therefore exactly represents the average squared discrepancy between the coordinated and original domain-specific linear projection responses on the fixed layer inputs recorded from each domain. Once Σ_i^l has been retained, this quantity can be evaluated without historical images or individual historical activations.

The scope of this exactness is limited to these fixed recorded layer inputs. During subsequent Coordination, jointly applying multiple frozen updates can change the representations produced by preceding layers. The layer inputs that a historical image would produce under the later unified encoder are therefore not generally identical to those recorded when Σ_i^l was constructed. Because the historical images are no longer available, $\mathcal{L}_{\text{memory}}$ cannot directly evaluate this evolving end-to-end response. It should instead be interpreted as an exactly computable constraint on the recorded historical layer inputs and, at the network level, as a tractable layer-wise surrogate for limiting historical response changes.

B.3 Output-Structure Preservation of the Update Contribution

The same retained input second moment also determines the empirical second moment of the contribution produced by a linear update on the recorded inputs. For

$$z_{i,k}^l(W) = W h_{i,k}^l, \tag{14}$$

we have

$$\frac{1}{K_i^l} \sum_{k=1}^{K_i^l} z_{i,k}^l(W) z_{i,k}^l(W)^\top = W \Sigma_i^l W^\top. \tag{15}$$

This follows directly by substituting $z_{i,k}^l(W) = W h_{i,k}^l$ and moving W and $W^\top$ outside the empirical sum.

KPKW therefore defines output-structure preservation as

$$\mathcal{L}_{\text{output}} = \sum_{i \le t} \sum_{l} \left\| C^l \Sigma_i^l C^{l\top} - \Delta W_i^l \Sigma_i^l \Delta W_i^{l\top} \right\|_F^2. \tag{16}$$

This term compares the uncentered second-order structure of the update contributions produced by the coordinated update and the original domain-specific update on the same recorded input statistic. It differs from $\mathcal{L}_{\text{memory}}$ in what is aggregated. Historical response preservation averages squared per-input response differences, while output-structure preservation compares the aggregate second moments induced by the two update contributions.

Both relations are exact with respect to the fixed empirical inputs represented by Σ_i^l. Neither relation implies exact preservation of the complete layer-output distribution or the end-to-end network representation after subsequent stages. The output-structure term reuses Σ_i^l and introduces no additional retained statistic. It therefore serves as a complementary structural constraint rather than an independent guarantee of historical preservation.

B.4 Positive-Semidefinite Structure of the Retained Statistics

The two retained statistics are symmetric positive semidefinite. For any vector v,

$$v^\top \Sigma_i^l v = \frac{1}{K_i^l} \sum_{k} (v^\top h_{i,k}^l)^2 \ge 0, \tag{17}$$

and

$$v^\top S_{B,i}^l v = \sum_{y} \pi_y \left[v^\top (\mu_{i,y}^l - \bar{\mu}_i^l) \right]^2 \ge 0, \tag{18}$$

where $\pi_y \ge 0$ by construction. Both matrices are symmetric because they are sums of symmetric outer products.

Consequently, Σ_i^l defines a nonnegative quadratic form in the historical response objective. Likewise, $S_{B,i}^l$ admits an eigendecomposition with nonnegative eigenvalues, which provides a well-defined spectral basis for selecting leading directions according to cumulative retained energy. This algebraic property does not imply that the finite-sample statistic recovers the population distribution. It also does not imply that the selected eigendirections span a complete identity-discriminative space. They characterize principal variation among the recorded historical identity class means.

B.5 Soft Gradient Adjustment Using Historical Identity-Related Directions

For each historical domain $j < t$, let $U_{B,j}^l$ contain orthonormal leading eigenvectors of $S_{B,j}^l$ selected according to the cumulative retained-energy ratio ρ_B. KPKW defines

$$P_{B,j}^l = I - U_{B,j}^l {U_{B,j}^l}^\top. \tag{19}$$

Each $P_{B,j}^l$ is the orthogonal projector onto the complement of the selected domain-specific directions. The implementation applies these operators sequentially following the historical-domain order. For a raw gradient $g_{i,l}$ of channel-wise coefficient m_i^l, let

$$q_{i,l}^{(0)} = g_{i,l}, \tag{20}$$

and recursively apply

$$q_{i,l}^{(j)} = P_{B,j}^l q_{i,l}^{(j-1)}, \qquad j = 1, \ldots, t-1. \tag{21}$$

The resulting sequentially adjusted gradient can be written as

$$q_{i,l}^{(t-1)} = P_{B,\mathrm{comp}}^l g_{i,l}, \tag{22}$$

where $P_{B,\mathrm{comp}}^l$ denotes the ordered product induced by the sequential application above. KPKW then uses the soft interpolation

$$g'_{i,l} = (1-\beta) g_{i,l} + \beta P_{B,\mathrm{comp}}^l g_{i,l}, \qquad \beta \in [0,1]. \tag{23}$$

Because each $P_{B,j}^l$ is an orthogonal projector,

$$\|P_{B,j}^l\|_2 \le 1. \tag{24}$$

Submultiplicativity therefore gives

$$\|P_{B,\mathrm{comp}}^l\|_2 \le 1. \tag{25}$$

The soft adjustment is consequently non-expansive,

$$\begin{aligned} \|g'_{i,l}\|_2 &\le (1-\beta)\|g_{i,l}\|_2 + \beta \|P_{B,\mathrm{comp}}^l g_{i,l}\|_2 \\ &\le \|g_{i,l}\|_2. \end{aligned} \tag{26}$$

For a single historical subspace,

$$P_B = I - U_B U_B^\top, \tag{27}$$

and the adjustment satisfies

$$U_B^\top g' = (1-\beta) U_B^\top g \tag{28}$$

and

$$P_B g' = P_B g. \tag{29}$$

Only in this single-subspace case can β be interpreted as an exact attenuation factor for the selected directions. For example, $\beta = 0.75$ retains 25% of the raw gradient component in that selected subspace.

For multiple historical domains, the projectors need not commute. Their ordered product is therefore generally neither symmetric nor idempotent and should not be interpreted as a single orthogonal projector onto a common historical subspace. A later projection associated with one historical domain can also alter components relative to an earlier non-orthogonal subspace. Consequently, no identical per-direction attenuation factor is guaranteed across all historical domains. In the multi-domain setting, β represents the interpolation strength between the raw gradient and the result of the implemented sequential adjustment. The procedure provides soft identity-related gradient guidance rather than exact invariance of historical directions.

B.6 Rank and Norm Properties of Channel-Wise Composition

For every domain i and layer l, channel modulation cannot increase the rank of an individual frozen update,

$$\operatorname{rank}\left(\Delta W_i^l \operatorname{Diag}(m_i^l)\right) \le \operatorname{rank}(\Delta W_i^l) \le r_i. \tag{30}$$

This follows from the standard rank inequality for matrix products and the rank-r_i factorization of ΔW_i^l.

The effective update C^l is the sum of the modulated contributions from all retained domains. Therefore,

$$\operatorname{rank}(C^l) \le \min\left(d_{\text{out}}^l, d_{\text{in}}^l, \sum_{i=1}^{t} r_i\right). \tag{31}$$

KPKW does not assume that C^l remains rank r_i, nor does it physically merge the accumulated LoRA updates into a new single adapter. The matrix C^l is an algebraic representation of their joint contribution in the forward computation.

The spectral norm satisfies

$$\begin{aligned} \|C^l\|_2 &= \left\| \sum_{i=1}^{t} \Delta W_i^l \operatorname{Diag}(m_i^l) \right\|_2 \\ &\le \sum_{i=1}^{t} \|\Delta W_i^l\|_2 \|m_i^l\|_\infty. \end{aligned} \tag{32}$$

This provides a finite upper bound for a fixed collection of finite frozen updates and channel-wise coefficients. It is not a stream-independent stability bound and does not imply that $\|C^l\|_2$ remains constant as additional domains are learned.

These algebraic properties also do not imply a hard partition of channels across domains. Different frozen updates may contribute through the same channels. The overlap regularization used during Coordination only discourages excessive overlap and does not impose mutually exclusive or orthogonal channel allocation.

B.7 Storage and Computational Growth

To separate the theoretical state considered here from implementation-specific checkpoint contents, let $\mathcal{L}_{\text{comp}}$ denote the inserted LoRA layers equipped with channel-wise composition coefficients, and let $\mathcal{L}_{\text{stat}} \subseteq \mathcal{L}_{\text{comp}}$ denote the layers for which Σ_i^l and $S_{B,i}^l$ are retained.

For domain D_i, the frozen LoRA updates and channel-wise coefficients over $\mathcal{L}_{\text{comp}}$ require

$$\sum_{l \in \mathcal{L}_{\text{comp}}} \left[r_i(d_{\text{in}}^l + d_{\text{out}}^l) + d_{\text{in}}^l \right] \tag{33}$$

scalars. The full-matrix retained statistics require an additional

$$2 \sum_{l \in \mathcal{L}_{\text{stat}}} (d_{\text{in}}^l)^2 \tag{34}$$

scalars. Therefore, the storage associated with these core KPKW states after t domains is

$$O\left(\sum_{i=1}^{t} \left\{ \sum_{l \in \mathcal{L}_{\text{comp}}} \left[r_i (d_{\text{in}}^l + d_{\text{out}}^l) + d_{\text{in}}^l \right] + 2 \sum_{l \in \mathcal{L}_{\text{stat}}} (d_{\text{in}}^l)^2 \right\} \right). \tag{35}$$

When the architecture and LoRA rank are fixed across domains, this part of the persistent state grows linearly with the number of learned domains.

This expression describes only the core KPKW states analyzed above. The complete serialized training checkpoint can contain additional implementation-specific persistent state and training metadata. Section D.8 reports the measured complete checkpoint budget and its decomposition. The expression above should therefore not be interpreted as an exact accounting of the full serialized checkpoint.

The current joint forward explicitly evaluates the retained low-rank branches. Its adapter-dependent inference computation consequently grows with the number of accumulated domain updates. This is an implementation property of the current unified encoder and does not imply constant inference cost as the stream grows.

The retained statistical information serves as historical reference during subsequent Coordination. Once Coordination is complete, online Retrieval uses the frozen domain updates together with the learned and fixed channel-wise composition coefficients. The matrices Σ_i^l and $S_{B,i}^l$ are not required for the online retrieval forward.

With the full empirical Σ_i^l, Lemma 1 exactly reproduces the response discrepancy on the recorded historical layer inputs. If Σ_i^l is replaced by a diagonal or low-rank approximation, the resulting objective remains computable but does not in general equal the original full-matrix empirical discrepancy exactly. Section D.5 evaluates the corresponding accuracy-storage trade-off. KPKW is therefore exemplar-free, but it is neither state-free nor constant-cost.

B.8 Summary of the Theoretical Analysis

The analysis supports the following interpretation of KPKW.

1. Retention preserves each learned domain-specific transformation as a frozen low-rank update. Coordination leaves these updates unchanged and learns how their contributions are organized through channel-wise coefficients.
2. Recording retains layer-wise input second moments that make the empirical response discrepancy and update-output second moments exactly computable on the fixed historical layer inputs represented by those statistics. These layer-wise identities provide tractable historical references for subsequent Coordination, but they do not guarantee preservation of complete end-to-end historical behavior.
3. Between-class scatter records principal variation directions among historical identity class means. These directions guide a sequential soft adjustment of the channel-wise coefficient gradients. The multi-domain rule is an ordered sequence of domain-specific projections rather than a single orthogonal projection onto a common historical subspace, and it does not provide exact invariance of historical directions.
4. Channel-wise composition does not increase the rank of any individual frozen update, but the combined effective update may have a higher rank and may contain overlapping contributions from different domains. The retained state and adapter-dependent computation of the current implementation grow with the number of learned domains.

After Coordination, the retained statistical information is no longer required for online Retrieval. The learned channel-wise coefficients and frozen domain updates are fixed and jointly used by the same unified encoder for all query and gallery images without domain labels. The results above establish the mathematical consistency and scope of the implemented mechanisms. Retention, generalization, and efficiency remain empirical properties rather than consequences guaranteed by these identities.

C SUPPLEMENTARY METHODOLOGICAL CLARIFICATION

C.1 RETENTION, RECORDING, COORDINATION, AND RETRIEVAL

This section complements the main-paper description by making the phase boundaries, retained information, and trainable states of KPKW explicit. It also clarifies how the four stages interact under the exemplar-free protocol and how the retained states used during continual learning differ from those required for online retrieval.

Consistent with the organization in the main paper, KPKW separates lifelong learning into four stages: **Retention**, **Recording**, **Coordination**, and **Retrieval**. The distinction is operational. Each stage determines what is learned, what remains frozen, what information is retained, and what is required after the corresponding domain becomes unavailable.

Retention. At stage t, KPKW attaches a new domain-specific LoRA update M_t to the frozen ViT backbone. Only the current update and current-domain classifier are optimized using identity classification and batch-hard triplet losses, while all previously retained domain updates remain frozen. After current-domain training, M_t is frozen and retained as the transformation learned from D_t.

Recording. Before the samples of D_t become unavailable, KPKW performs a statistics-recording pass with the frozen update M_t active. Layer-input observations are aggregated into the input second moments Σ_t^l and between-class scatter matrices $S_{B,t}^l$. The former records second-order information about the layer inputs on which the retained update operated, while the latter records major variation directions among historical identity class means. The sample-level and token-level quantities used to form these aggregates are discarded after recording.

Coordination. From the second stage onward, all accumulated domain-specific LoRA updates remain frozen and KPKW optimizes only their channel-wise composition coefficients. Current-domain supervision drives adaptation, while retained input second moments provide historical references for response and output-structure preservation. Between-class scatter further provides identity-related guidance for the soft gradient adjustment. Overlap regularization discourages excessive overlap among different domain contributions without enforcing a hard channel partition. Historical images and instance-level historical features are not revisited during this phase. The first stage contains only one domain-specific update and therefore uses that update directly without multi-update Coordination.

Retrieval. After Coordination, both the frozen domain updates and their learned channel-wise composition coefficients are fixed. The same unified encoder jointly applies the accumulated updates to every query and gallery image for identity representation and retrieval, without requiring domain labels. The retained statistics are training-time references for subsequent Coordination and are not evaluated in the online retrieval forward.

C.2 ADDITIONAL VISUALIZATION OF STATISTICS-GUIDED COORDINATION

Figure A1 provides a detailed visualization of the Coordination stage in KPKW. All retained domain-specific updates remain frozen, while their channel-wise composition coefficients are optimized using current-domain supervision together with historical references from the retained input second moments and identity-related guidance from the between-class scatter. After optimization, the learned channel-wise organization is fixed and used together with the accumulated frozen updates for unified Retrieval.

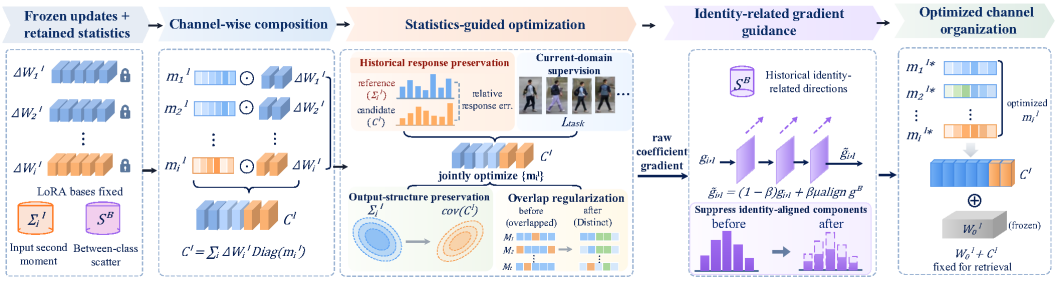

Figure A1: **Detailed illustration of statistics-guided Coordination in KPKW.** Frozen domain-specific updates remain fixed while their channel-wise composition coefficients are optimized using current-domain supervision, historical references from the retained input second moments, and identity-related guidance from the between-class scatter. The learned channel-wise organization is then fixed for unified Retrieval.

C.3 Statistics Recording Scope, Sample Counts, and Tensor Dimensions

This section makes the statistics recording procedure in the main paper operationally precise. The implementation inserts LoRA into the `qkv`, `proj`, `fc1`, and `fc2` linear projections of every Transformer block and registers a forward pre-hook on each of these LoRA layers to access its input. The convolutional patch embedding and identity classifier are not included in the statistics recording. Let $\mathcal{L}_{\mathrm{stat}}$ denote the set of LoRA layers for which the retained statistics are constructed. With the ViT-B/16 backbone used in our experiments, there are 12 Transformer blocks and four recorded projections per block, giving

$$|\mathcal{L}_{\mathrm{stat}}| = 12 \times 4 = 48. \tag{36}$$

Here, $\mathcal{L}_{\mathrm{stat}}$ refers specifically to the statistics-recording scope.

For domain D_i, statistics are recorded in evaluation mode with only its frozen domain-specific update M_i active. We first construct the layer-wise input second moment. Let the statistics loader yield batches indexed by b, where batch b contains B_b images, and let T_l denote the number of token-level input vectors entering layer l for each image. For $h^l_{i,b,n,s} \in \mathbb{R}^{d^l_{\mathrm{in}}}$, where n indexes images and s indexes token positions, the implementation flattens the batch and token dimensions before accumulation. Therefore,

$$K^l_i = \sum_{b \in \mathcal{B}^{\Sigma}_i} B_b T_l, \tag{37}$$

and

$$\Sigma^l_i = \frac{1}{K^l_i} \sum_{b \in \mathcal{B}^{\Sigma}_i} \sum_{n=1}^{B_b} \sum_{s=1}^{T_l} h^l_{i,b,n,s} h^l_{i,b,n,s}{}^{\top}. \tag{38}$$

Accordingly, K^l_i denotes the number of token-level layer-input vectors actually processed when constructing Σ^l_i. It is neither the number of identities nor simply the number of images in D_i.

Under the main configuration, input images have resolution 256×128, while both the patch size and stride are 16. Each Transformer block therefore receives $16 \times 8 = 128$ patch tokens together with one class token, giving $T_l = 129$ for all layers in $\mathcal{L}_{\mathrm{stat}}$. Hence,

$$K^l_i = 129 \sum_{b \in \mathcal{B}^{\Sigma}_i} B_b. \tag{39}$$

The main experiments set the input-statistics batch limit to zero, corresponding to no batch cap on this recording procedure. Thus, $\mathcal{B}^{\Sigma}_i$ contains all batches yielded by the statistics loader. The count K^l_i is accumulated directly from the observed tensors, so a smaller final batch is handled without approximation.

The between-class scatter is constructed using a separate sample-level aggregation. For every processed image, the implementation first mean-pools all T_l token-level inputs, including the class token,

$$\tilde{h}^l_{i,n} = \frac{1}{T_l} \sum_{s=1}^{T_l} h^l_{i,n,s}. \tag{40}$$

These pooled vectors are then accumulated by identity. Let $n_{i,y}$ denote the number of sampled image occurrences associated with identity y, let $N^B_i = \sum_{y \in \mathcal{Y}^B_i} n_{i,y}$, and let $\mathcal{Y}^B_i$ denote the set of identities observed during this statistics pass. The corresponding identity mean, empirical identity weight, and weighted domain mean are

$$\mu^l_{i,y} = \frac{1}{n_{i,y}} \sum_{n: y_n = y} \tilde{h}^l_{i,n}, \qquad \pi_y = \frac{n_{i,y}}{N^B_i}, \qquad \bar{\mu}^l_i = \sum_{y \in \mathcal{Y}^B_i} \pi_y \mu^l_{i,y}. \tag{41}$$

The retained between-class scatter is then

$$S_{B,i}^{l} = \sum_{y \in \mathcal{Y}_i^B} \pi_y (\mu_{i,y}^{l} - \bar{\mu}_i^{l})(\mu_{i,y}^{l} - \bar{\mu}_i^{l})^{\top}. \tag{42}$$

In the reported configuration, this aggregation uses at most 50 PK-sampled training batches with batch size 64, corresponding to at most 3,200 sampled image occurrences. The counts $n_{i,y}$ and weights π_y are computed from the samples actually observed during this pass. This sampling limit applies to the between-class statistic and is distinct from the uncapped input-second-moment recording described above.

After Σ_i^l and $S_{B,i}^l$ have been constructed, the token-level inputs, pooled sample vectors, identity labels, and per-identity temporary accumulators used during recording are discarded. Only the resulting domain-level aggregate statistics are retained.

The tensor dimensions are determined by the input and output widths of each LoRA projection:

Projection in each ViT-B/16 block	d_{in}^l	d_{out}^l	A_i^l	B_i^l	m_i^l	$\Sigma_i^l,\ S_{B,i}^l$
`qkv`	768	2304	$r_i \times 768$	$2304 \times r_i$	768	768×768
`proj`	768	768	$r_i \times 768$	$768 \times r_i$	768	768×768
`fc1`	768	3072	$r_i \times 768$	$3072 \times r_i$	768	768×768
`fc2`	3072	768	$r_i \times 3072$	$768 \times r_i$	3072	3072×3072

These four projection types repeat across all 12 Transformer blocks. Statistics are accumulated in FP32. Under the default full-statistics setting, each Σ_i^l and $S_{B,i}^l$ is retained as a full $d_{\text{in}}^l \times d_{\text{in}}^l$ matrix. The diagonal and low-rank alternatives evaluated in Section D.5 change the retained representation of these statistics but not the statistics-recording layer scope or the definitions above.

C.4 Retained State and the Exemplar-Free Boundary

After stage t, the core retained state of KPKW consists of the frozen ViT backbone, the frozen domain-specific LoRA updates $\{M_i\}_{i=1}^t$, their learned channel-wise composition coefficients, and the domain-level statistics required for subsequent Coordination. The latter include the layer-wise input second moments Σ_i^l and between-class scatter matrices $S_{B,i}^l$ described above.

The complete serialized training checkpoint additionally contains implementation-specific persistent state required by the training pipeline, including the retained feature covariance reported in Section D.8 and classifier or optimization state needed to resume training. These auxiliary checkpoint contents are distinguished from the core KPKW states above when analyzing the method, but are included when reporting the measured persistent-state budget.

KPKW does not retain historical images for replay, per-instance embeddings, or per-instance layer activations after the corresponding statistics have been formed. The sample-level and token-level quantities used during Recording are discarded after aggregation. The retained matrices are nevertheless persistent historical information rather than replayed examples, and their storage is explicitly accounted for in Section D.8.

Accordingly, the exemplar-free designation refers specifically to the absence of stored historical examples or instance-level features for replay. It does not imply that KPKW retains no historical state, that its storage cost is independent of the number of learned domains, or that the retained parameters and aggregate statistics provide a formal privacy guarantee.

C.5 Training and Inference Procedure

Algorithm C.5 summarizes the complete stage-wise training procedure of KPKW. The procedure explicitly separates the states required to continue lifelong learning from those required for online Retrieval.

Algorithm 1. KPKW Training over a Domain Stream

Input: domain sequence $\{D_t\}_{t=1}^T$; frozen backbone f_θ. **Continual-learning state:** $(f_\theta, \mathcal{M}, \mathcal{V}, \mathcal{S})$.

1. Initialize frozen domain updates $\mathcal{M} \leftarrow \varnothing$, channel-wise composition coefficients $\mathcal{V} \leftarrow \varnothing$, and retained statistics $\mathcal{S} \leftarrow \varnothing$.
2. For $t = 1, \ldots, T$:
   (a) **Retention:** attach a new domain-specific LoRA update M_t; freeze f_θ and all updates in $\mathcal{M}$; optimize M_t and the current classifier using D_t; then freeze M_t and append it to $\mathcal{M}$.
   (b) **Recording:** record Σ_t and $S_{B,t}$ while D_t remains accessible, then append them to $\mathcal{S}$.
   (c) If $t > 1$, **Coordination:** initialize or extend $\mathcal{V}$ for the accumulated frozen updates and activate their joint forward. Optimize only $\mathcal{V}$ using current-domain task supervision, overlap regularization, historical

response preservation, and output-structure preservation; apply soft S_B-based adjustment to the gradients of $\mathcal{V}$. Otherwise, use the single frozen update M_1 directly.

(d) Discard the images and instance-level features of D_t.

3. Return $(f_\theta, \mathcal{M}, \mathcal{V}, \mathcal{S})$.

Inference. After the final Coordination stage, the frozen domain updates and learned channel-wise composition coefficients are fixed. For every query or gallery image, the unified encoder jointly applies the accumulated frozen updates with their fixed coefficients and produces a single representation for standard feature matching. No domain label is required. The retained statistics $\mathcal{S}$, including Σ_i^l and $S_{B,i}^l$, are required for subsequent Coordination during lifelong training but are not evaluated in the online retrieval forward. Therefore, the deployed retrieval state consists of $(f_\theta, \mathcal{M}, \mathcal{V})$, whereas continuing the domain stream additionally requires the retained statistics $\mathcal{S}$.

D Additional Experiments

Unless stated otherwise, controlled variants use Training Order-1 and seed 3407, while rows representing the selected final KPKW result mirror the seed-8380 main-table result. Results are mAP/R@1.

The complete KPKW recipe consequently appears in this appendix under two different roles, and the two should not be read as disagreeing measurements of one number. Where the full recipe serves as a *reference* row alongside prior methods or per-domain results—Tables A1, A9, A12, and A15—it reproduces the selected seed-8380 main result, $62.2/74.2$ Seen-Avg and $73.9/68.0$ Unseen-Avg. Where it is instead the endpoint *arm of a controlled ablation*, it is the seed-3407 run of that ablation, measured under that section's own protocol and diagnostic configuration, and the deltas reported in the same table are computed from it. Those arms read $62.0/73.8$ and $73.3/66.9$ in Table A10, $61.8/73.3$ and $72.9/66.0$ in Tables A14, A16, and A17, and $62.0/73.8$ and $73.0/66.9$ in Table A18. Relative to the selected result the spread reaches 0.4 on Seen-Avg mAP and 2.0 on Unseen-Avg R@1, reflecting both the change of seed and the run-to-run variation quantified in Section D.16. Comparisons are therefore valid within a table and not across tables, and no arm has been substituted by the main-table value.

D.1 Controls for Identity-Related Gradient Guidance

Table A1: Controls for the source and form of identity-related gradient guidance. All projected variants start from the same historical-response and output-structure preservation configuration. The best and second-best performance results are marked in bold and underline, respectively.

Variant	Projection source	Adjustment	Seen-Avg	Unseen-Avg
Task-only composition	none	none	57.8/71.2	67.8/60.2
$+\Sigma$ historical response preservation	none	none	<u>61.8/73.6</u>	**74.0**/67.5
+ output-structure preservation	none	none	61.6/73.3	73.8/<u>67.7</u>
Rank-matched random projection	random orthogonal basis	hard, $\beta = 1.00$	<u>61.9</u>/73.5	73.0/66.9
Input-moment projection	Σ	hard, $\beta = 1.00$	61.4/73.2	73.2/66.3
Hard identity-related projection	S_B	hard, $\beta = 1.00$	61.6/73.4	72.7/66.0
KPKW	S_B	soft, $\beta = 0.75$	**62.2/74.2**	<u>73.9</u>/**68.0**

Historical response preservation provides a substantial improvement over task-only composition, confirming the value of the retained input second moments as historical references during Coordination. Adding output-structure preservation produces a more modest change, consistent with its role as a complementary structural constraint.

The rank-matched random and input-moment controls use the same hard-adjustment mechanism as the hard S_B variant. These controls examine whether the choice of guidance directions matters beyond generic random suppression or directions derived only from activation energy. Their different behaviors indicate that the projection source affects the resulting retention and generalization trade-off. Hard S_B adjustment is also weaker than the final soft configuration, suggesting that preserving plasticity while incorporating historical identity-related guidance is preferable to fully suppressing the adjusted gradient component.

The Σ-only variant achieves the highest Unseen mAP, whereas full KPKW achieves the strongest Seen-Avg and Unseen R@1. This metric-dependent behavior indicates that the different coordination mechanisms contribute differently to historical retention and cross-domain generalization. Considering the overall balance between seen-domain retention and unseen-domain generalization, we therefore adopt the S_B-guided soft adjustment with $\beta = 0.75$ as the default configuration of KPKW.

Table A2: Sensitivity to the S_B soft-adjustment strength. The second column reports the interpolation weight $(1-\beta)$ assigned to the raw gradient in the update rule. $\beta = 0$ leaves the raw gradient unchanged, whereas $\beta = 1$ uses the fully sequentially adjusted gradient. The selected setting is highlighted in the β column, and the best and second-best performance results are marked in bold and underline, respectively.

β	**Raw-gradient interpolation weight** $(1-\beta)$	**Seen-Avg**	**Unseen-Avg**
0.00	100%	61.6/73.3	<u>73.8/67.7</u>
0.25	75%	<u>62.0/73.7</u>	73.2/67.2
0.50	50%	61.9/73.3	73.0/66.6
0.75	25%	**62.2/74.2**	**73.9/68.0**
1.00	0%	61.6/73.4	72.7/66.0

D.2 Soft-Adjustment Strength

The results show a clear preference for an intermediate adjustment strength. Without identity-related gradient adjustment, corresponding to $\beta = 0$, the model already benefits from historical response and output-structure preservation, but does not achieve the strongest overall performance. Increasing β introduces progressively stronger guidance from the historical between-class statistics. The best results are obtained at $\beta = 0.75$, which achieves the highest Seen-Avg and Unseen-Avg among the reported settings. In contrast, the hard setting $\beta = 1$ decreases all four metrics, indicating that fully relying on the sequentially adjusted gradient can overly restrict the plasticity required for the current domain.

For multiple historical domains, $(1-\beta)$ should be interpreted as the interpolation weight of the raw gradient rather than the exact fraction of every historical identity-related gradient component that is retained, since the sequential domain-specific projectors need not commute. Overall, the results support a soft balance between historical guidance and current-domain plasticity. We therefore use $\beta = 0.75$ as the default soft-adjustment strength in KPKW.

D.3 Forgetting and Backward Transfer

Let $R_{t,j}$ denote the performance on domain j after learning stage t. For a sequence of T stages, we compute average forgetting (AF) and backward transfer (BWT) directly from the raw stage-by-domain performance matrix:

$$\mathrm{AF} = \frac{1}{T-1}\sum_{j=1}^{T-1}\left(\max_{k\in\{j,\ldots,T-1\}} R_{k,j} - R_{T,j}\right), \tag{43}$$

$$\mathrm{BWT} = \frac{1}{T-1}\sum_{j=1}^{T-1}\left(R_{T,j} - R_{j,j}\right). \tag{44}$$

Table A3: Stage-by-domain performance and forgetting analysis of KPKW. Each performance entry and each AF/BWT value reports mAP/R@1. AF and BWT are reported on the final-stage row of each training order and are computed directly from the corresponding stage-by-domain measurements.

Order	**Stage**	D_1	D_2	D_3	D_4	D_5	**AF** $\downarrow$	**BWT** $\uparrow$
Order-1	Sequence	Market	SYSU	LPW	MSMT17	CUHK03	–	–
	1	84.1/93.7	–	–	–	–	–	–
	2	82.6/93.5	89.2/90.8	–	–	–	–	–
	3	78.4/91.3	90.5/91.5	63.9/74.4	–	–	–	–
	4	76.1/90.1	91.4/92.4	62.2/73.5	37.5/66.4	–	–	–
	5	73.1/88.8	90.8/92.0	58.7/70.6	32.6/62.1	55.9/57.4	5.4/3.4	-4.9/-3.0
Order-2	Sequence	LPW	MSMT17	Market	SYSU	CUHK03	–	–
	1	68.6/78.7	–	–	–	–	–	–
	2	66.4/76.5	42.4/70.4	–	–	–	–	–
	3	65.8/75.6	37.8/67.3	78.2/91.7	–	–	–	–
	4	60.9/71.8	31.3/60.7	75.0/89.2	91.6/92.6	–	–	–
	5	59.2/70.8	30.9/60.4	71.9/88.3	90.9/92.0	54.6/54.8	7.0/5.5	-7.0/-5.5

Order-1 exhibits lower forgetting and less negative backward transfer under both metrics. Its AF differs from $-\mathrm{BWT}$ because the performance on SYSU improves after the domain is first learned: it increases from 89.2/90.8

at stage 2 to 90.8/92.0 at the final stage and reaches 91.4/92.4 at stage 4. This intermediate improvement affects the peak-based AF and the acquisition-stage-based BWT differently. In Order-2, all four historical domains finish below their performance when first learned, and AF consequently equals $-\mathrm{BWT}$.

The evaluator records per-domain scores to one decimal place. We therefore compute AF and BWT directly from these measured per-domain entries and round only the resulting aggregate values to one decimal place. No cumulative average is inverted to reconstruct the stage-level results.

Figure A2 reorganizes the same stage-by-domain matrix into a two-dimensional retention map, making the temporal evolution of each historical domain easier to inspect. Under Order-1, the largest decline occurs on the first domain, Market-1501, whose mAP decreases from 84.1 to 73.1. In contrast, CUHK-SYSU improves after acquisition and reaches a maximum gain of $+2.2$ mAP at stage 4. Under Order-2, the two earliest domains, LPW s2 and MSMT17, both exhibit larger declines, with MSMT17 decreasing by 11.5 mAP points by the final stage. The larger losses on these early domains contribute to the higher AF and more negative BWT observed under Order-2.

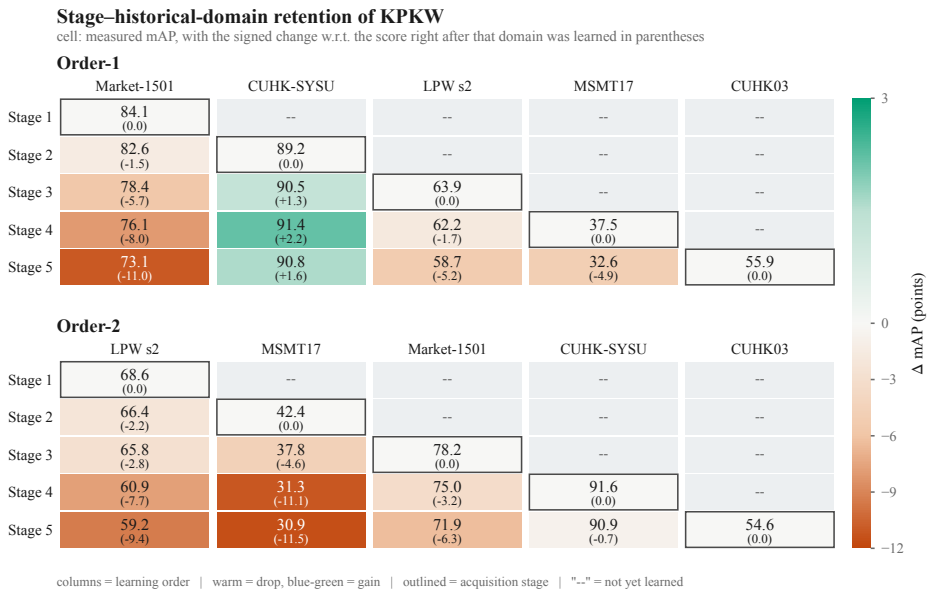

Figure A2: **Stage–historical-domain retention of KPKW under both training orders.** Each cell reports the measured mAP after a learning stage, with the signed change relative to the performance obtained immediately after that domain was learned shown in parentheses. Color encodes this relative change rather than absolute mAP so that differences in dataset scale do not obscure the retention pattern. Both panels share a diverging scale centered at zero and fixed to $[-12, +3]$, with warm colors denoting decreases and blue-green colors denoting increases. Outlined cells indicate the stage at which each domain is learned, and light-gray cells indicate domains not yet observed. The map is constructed directly from the stage-by-domain measurements in Table A3 rather than reconstructed from cumulative averages. Since protocol-matched stage-by-domain matrices are not available for the compared methods, this visualization characterizes the retention dynamics of KPKW rather than providing a baseline comparison.

Overall, although KPKW does not completely eliminate forgetting, the stage-wise results show that it preserves substantial performance on previously learned domains throughout the learning sequence and can even produce positive backward transfer on some historical domains. The controlled degradation across successive stages indicates that retaining domain-specific transformations and coordinating them with historical statistical references effectively mitigates catastrophic forgetting while preserving the plasticity required to learn new domains. These results provide direct stage-wise evidence for the strong anti-forgetting capability of KPKW in lifelong ReID.

D.4 Per-Unseen-Domain Results

Tables A4 and A5 further break down unseen-domain generalization across CUHK01, CUHK02, VIPeR, PRID, i-LIDS, GRID, and SenseReID. We compare KPKW with the strong ViT-based methods PAEMA, DRE, and DKUA to examine whether the aggregate unseen-domain improvement reported in the main paper is consistently distributed across individual target domains.

Table A4: Per-unseen-domain performance on Training Order-1. Each entry reports mAP/R@1. The best and second-best results are marked in bold and underline, respectively.

Method	CUHK01	CUHK02	VIPeR	PRID	i-LIDS	GRID	SenseReID	Unseen-Avg
PAEMA (Li et al., 2024)	82.0/81.6	79.9/79.1	64.8/54.5	72.9/64.6	83.4/77.3	56.1/46.0	49.5/39.3	69.8/63.2
DRE (Liu et al., 2025a)	75.1/76.5	73.0/74.0	57.9/49.4	66.0/59.5	76.5/72.2	49.2/40.9	42.6/34.2	62.9/58.1
DKUA (Liu et al., 2025b)	80.0/78.7	77.9/76.2	62.8/51.6	70.9/61.7	81.4/74.4	54.1/43.1	47.5/36.4	67.8/60.3
KPKW	**85.7/86.3**	**85.6/86.4**	**68.0/57.6**	**75.1/69.0**	**87.3/80.0**	**59.9/49.6**	**55.7/47.3**	**73.9/68.0**

Under Order-1, KPKW achieves the best mAP and R@1 on all seven unseen domains. Compared with the strongest reference method, PAEMA, the gains are observed consistently across CUHK01, CUHK02, VIPeR, PRID, i-LIDS, GRID, and SenseReID rather than being concentrated on a single target dataset. This results in an improvement of 4.1/4.8 points in Unseen-Avg, from 69.8/63.2 to 73.9/68.0.

Table A5: Per-unseen-domain performance on Training Order-2. Each entry reports mAP/R@1. The best and second-best results are marked in bold and underline, respectively.

Method	CUHK01	CUHK02	VIPeR	PRID	i-LIDS	GRID	SenseReID	Unseen-Avg
PAEMA (Li et al., 2024)	82.5/82.6	81.7/80.9	64.5/53.3	**73.7/66.0**	84.6/**77.7**	55.7/44.8	50.8/41.3	70.5/63.8
DRE (Liu et al., 2025a)	72.0/73.2	71.2/71.5	54.0/43.9	63.2/56.6	74.1/68.3	45.2/35.4	40.3/31.9	60.0/54.4
DKUA (Liu et al., 2025b)	78.9/78.5	78.1/76.8	60.9/49.2	70.1/61.9	81.0/73.6	52.1/40.7	47.2/37.2	66.9/59.7
KPKW	**84.0/84.0**	**84.6/84.3**	**68.2/58.5**	69.9/63.0	**85.7**/76.7	**57.4/45.6**	**55.0/45.8**	**72.1/65.4**

Under Order-2, KPKW again shows broadly strong generalization, achieving the best mAP on six of the seven unseen domains and the best R@1 on five. PAEMA remains stronger on PRID under both metrics and on i-LIDS R@1. Nevertheless, KPKW improves the overall Unseen-Avg from 70.5/63.8 to 72.1/65.4, corresponding to gains of 1.6/1.6 points over PAEMA.

Taken together, the two training orders show that the unseen-domain advantage of KPKW is broadly distributed across heterogeneous target domains rather than being driven by a particular benchmark. The consistent improvements across most individual domains indicate that retaining domain-specific transformations and coordinating them with historical statistical references helps form a unified representation with strong transferability beyond the domains observed during lifelong learning. These results further support the strong and robust cross-domain generalization capability of KPKW.

D.5 Retained-Statistics Compression and Accuracy–Storage Trade-off

The default implementation retains the full input second moments and between-class scatter matrices. This preserves the complete recorded second-order information used by the statistics-guided Coordination objectives, but these matrices also constitute the dominant part of the persistent state. We therefore evaluate whether structured approximations can preserve the benefit of statistical references while substantially reducing storage.

For a positive-semidefinite statistic S, the rank-k representation retains its leading eigensystem,

$$S \approx U_k \Lambda_k U_k^\top, \tag{45}$$

whereas the diagonal representation retains only its coordinate-wise energy. The same representation is used consistently during statistics recording, Coordination, and the extraction of identity-related directions from S_B.

Table A6: Accuracy–storage trade-off for the retained statistics on Training Order-1. Statistics storage includes only the Σ and S_B states affected by the compression experiment. The fixed 2.251 MiB/domain feature covariance is excluded from this column but included in the final checkpoint. The best and second-best performance results are marked in bold and underline, respectively.

Statistics representation	Retained rank	Statistics storage per domain	Final checkpoint	Seen-Avg	Unseen-Avg
No retained statistics	–	0	702.7 MiB	57.8/71.2	67.8/60.2
Diagonal	–	0.519 MiB	716.6 MiB	60.7/72.4	73.4/67.1
Low-rank	16	7.932 MiB	753.6 MiB	59.9/71.9	68.7/61.5
Low-rank	32	15.813 MiB	793.0 MiB	60.9/72.9	70.4/63.5
Low-rank	64	31.575 MiB	871.9 MiB	61.6/73.5	72.8/66.4
Full matrices	full	1026.00 MiB	5844.1 MiB	**62.2/74.2**	**73.9/68.0**

Removing the retained statistics causes a substantial degradation, reducing Seen-Avg from 62.2/74.2 to 57.8/71.2 and Unseen-Avg from 73.9/68.0 to 67.8/60.2. This confirms that the historical statistical references provide important information for coordinating the accumulated frozen domain updates rather than serving merely as auxiliary stored state.

The full-matrix representation achieves the strongest overall retention and generalization. Nevertheless, a large fraction of this benefit remains under substantial compression. Among the low-rank variants, rank 64 reduces

the retained $\Sigma + S_B$ storage from 1026.00 MiB to 31.575 MiB per domain, corresponding to a $32.5\times$ reduction, while remaining within 0.6/0.7 points on Seen-Avg and 1.1/1.6 points on Unseen-Avg of the full representation.

The diagonal representation provides an even more aggressive operating point. It requires only 0.519 MiB of retained statistics per domain while achieving 73.4/67.1 Unseen-Avg, only 0.5/0.9 points below the full representation. Its Seen-Avg decreases more noticeably by 1.5/1.8 points, indicating that the information removed by diagonalization is more important for historical retention than for unseen-domain generalization. Since diagonalization also changes the eigenspectrum used to derive the S_B guidance directions, it should be regarded as an approximation of the retained statistical information rather than a serialization-only compression technique.

Figure A3 visualizes these operating points by plotting seen- and unseen-domain mAP against the retained $\Sigma + S_B$ storage per domain. The full representation occupies the highest-accuracy operating point, whereas the compressed variants reveal a broad range of lower-storage alternatives. The operating points are not strictly monotonic with retained size, as illustrated by the rank-16 variant falling below the diagonal representation despite using more storage. This indicates that preserving the appropriate statistical structure is more important than storage capacity alone.

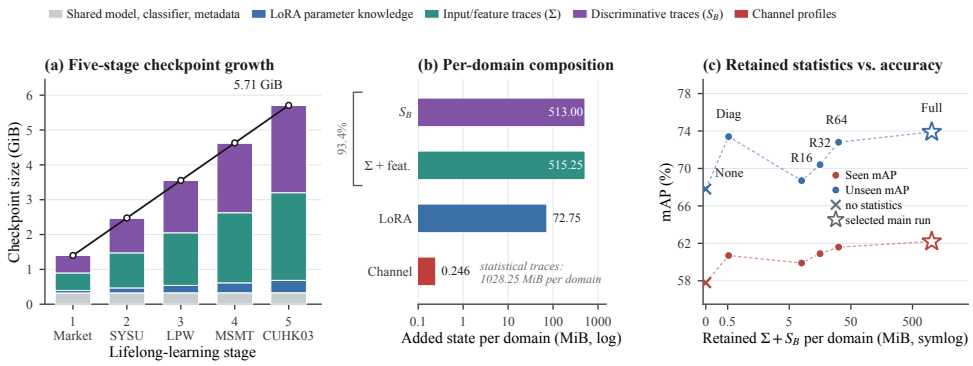

Figure A3: **Accuracy–storage trade-off of the retained statistics on Training Order-1.** Seen- and unseen-domain mAP are plotted against the retained $\Sigma + S_B$ storage per domain, corresponding to the same statistics budget reported in Table A6. The horizontal axis uses a symmetric-log scale so that the zero-storage configuration and highly compressed variants remain visible together with the full representation. Connecting lines are included only as visual guides between the evaluated operating points and do not represent a continuously tunable curve.

D.6 Multi-Seed Stability

To evaluate the stability of KPKW under stochastic optimization, we repeat the complete KPKW training procedure using five pre-specified random seeds: 2723, 6198, 8380, 14018, and 25696. Table A7 reports the protocol-matched result used in the main paper together with the sample standard deviation computed over all completed runs. Complete results from all five seeds are available for Order-2, while four complete final evaluations are available for Order-1.

Table A7: Multi-seed stability of the final KPKW configuration. The value before $\pm$ follows the protocol-matched result reported in the main paper, while the value after $\pm$ denotes the sample standard deviation across the completed runs for the corresponding training order.

Training order	Number of runs	Seen mAP	Seen R@1	Unseen mAP	Unseen R@1
Order-1	4	62.2 ± 0.2	74.2 ± 0.4	73.9 ± 0.5	68.0 ± 0.9
Order-2	5	61.5 ± 0.3	73.3 ± 0.3	72.1 ± 0.4	65.4 ± 0.6

Across the completed runs, the standard deviation remains at most 0.5 points for mAP and 0.9 points for R@1. Small variations are consistently observed under both training orders, covering both seen-domain retention and unseen-domain generalization. These results indicate that the overall performance of KPKW is stable under different stochastic training conditions.

Importantly, the variation remains small for both central objectives of lifelong ReID. Across different random initializations and stochastic optimization trajectories, KPKW maintains a similar level of historical-domain retention while preserving strong generalization to unseen domains. The multi-seed results therefore support the robustness of the overall performance trend reported in the main experiments and provide a reference scale for interpreting small endpoint differences in subsequent controlled analyses.

D.7 HYPERPARAMETER AND LORA-CAPACITY SENSITIVITY

We complement the soft-adjustment analysis with two factors that directly control the strength and capacity of KPKW. The retained-energy ratio ρ_B determines how many leading directions of the historical between-class scatter are used for identity-related gradient guidance, while the LoRA rank controls the capacity of each retained domain-specific transformation. In each sweep, only the corresponding factor is changed while the remaining KPKW configuration is kept fixed.

Table A8: Sensitivity to the retained-energy ratio ρ_B (top) and LoRA domain-update capacity (bottom) on Training Order-1. Each sweep changes only one factor while keeping the remaining KPKW configuration fixed. The best and second-best performance results are marked in bold and underline, respectively.

ρ_B	**Mean retained rank**	**Seen-Avg**	**Unseen-Avg**
0.10	1.131	<u>62.0</u>/73.6	<u>73.4</u>/67.5
0.20	2.019	61.9/73.7	72.3/65.7
0.30	3.344	**62.2/74.2**	**73.9/68.0**
0.40	5.254	61.9/73.4	73.0/66.6
0.50	8.117	61.9/<u>73.9</u>	73.1/66.6

LoRA rank $r = \alpha$	**Frozen-update size per domain**	**Seen-Avg**	**Unseen-Avg**
16	9.1 MiB	58.0/70.3	70.4/63.7
32	18.2 MiB	59.8/71.8	71.5/65.3
64	36.4 MiB	<u>61.5</u>/73.3	<u>73.0</u>/66.5
128	72.8 MiB	**62.2/74.2**	**73.9/68.0**

Increasing ρ_B expands the set of historical identity-related directions used for gradient guidance, with the mean retained rank increasing from 1.131 at $\rho_B = 0.10$ to 8.117 at $\rho_B = 0.50$. Performance does not improve monotonically as more directions are included. Instead, $\rho_B = 0.30$ achieves the strongest Seen-Avg and Unseen-Avg across all reported settings. A smaller value provides more limited historical guidance, whereas further increasing ρ_B introduces a broader constraint on coefficient optimization without yielding additional performance gains. This behavior suggests that retaining a compact set of principal identity-related directions provides a better balance between historical guidance and current-domain plasticity. We therefore use $\rho_B = 0.30$ as the default setting.

The LoRA-rank sweep exhibits a different trend. Increasing the rank consistently improves all four endpoint metrics, confirming that sufficient capacity in each frozen domain-specific update is important for retaining the transformation learned from that domain. Increasing the rank from 16 to 128 improves Seen-Avg by 4.2/3.9 points and Unseen-Avg by 3.5/4.3 points. The improvement from rank 64 to rank 128 narrows to 0.7/0.9 points on Seen-Avg and 0.9/1.5 points on Unseen-Avg, indicating diminishing returns as the retained update capacity increases.

The storage cost of the frozen domain update increases correspondingly from 9.1 MiB per domain at rank 16 to 72.8 MiB at rank 128. Thus, the rank sweep exposes a direct capacity–storage trade-off. Since rank 128 consistently achieves the strongest retention and generalization among the evaluated settings, we adopt $r = \alpha = 128$ in the default KPKW configuration. Each rank setting is trained independently with its corresponding domain-specific LoRA updates and channel-wise composition coefficients. A lower-rank configuration therefore represents a separately trained model rather than a post-hoc truncation of a rank-128 update.

Together, these results show that KPKW is not highly sensitive to simply increasing the number of retained identity-related directions, while sufficient domain-update capacity remains important. The selected $\rho_B = 0.30$ and rank 128 configuration provides the strongest overall performance among the evaluated settings and is therefore used throughout the main experiments.

D.8 PERSISTENT-STATE BUDGET, INFERENCE COST, AND STREAM-LENGTH SCALABILITY

We first examine the persistent-state budget of the final KPKW configuration. Figure A4 shows that the serialized checkpoint grows from 1.40 GiB after the first domain to 5.71 GiB after the fifth. Each additional domain contributes 72.75 MiB of rank-128 frozen LoRA updates and 0.246 MiB of channel-wise composition coefficients. The full input second moments together with the feature covariance require 515.25 MiB per domain, while the full between-class scatter requires another 513.00 MiB. The retained statistics therefore occupy 1028.25 MiB per domain and account for 93.4% of the newly retained persistent state.

The output-structure preservation term reuses the recorded input second moments and does not introduce an additional statistic. These measurements indicate that the current full-matrix configuration prioritizes preserving

the complete recorded statistical information rather than minimizing continual-learning storage. Importantly, this large persistent-state cost mainly comes from the statistics retained for subsequent Coordination rather than from the online Retrieval path. Section D.5 further shows that this storage can be reduced substantially through structured approximations while retaining a large fraction of the performance benefit.

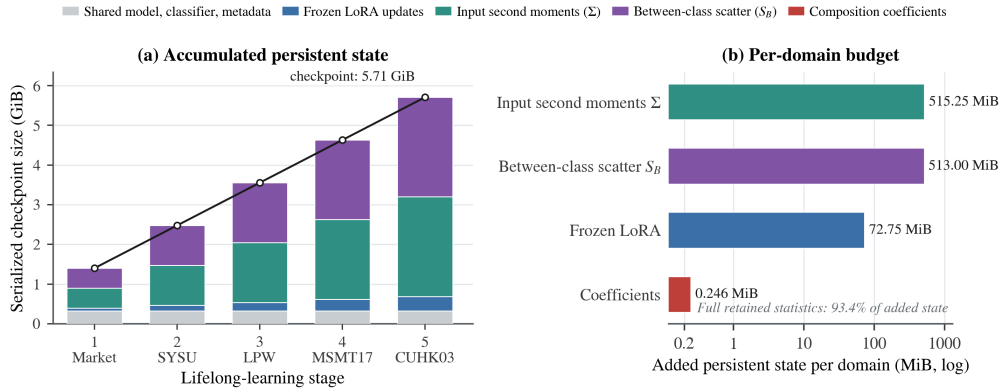

Figure A4: **Frozen-domain-update and retained-statistics budget of KPKW.** (a) Measured serialized checkpoint growth across the five Order-1 stages, decomposed into shared state, frozen LoRA updates, input/feature statistics, between-class statistics, and channel-wise composition coefficients. (b) Exact FP32 persistent-state increment per domain on a logarithmic axis. Full retained statistics account for 93.4% of the added persistent state.

Because the current implementation explicitly evaluates every retained low-rank branch during Retrieval, its adapter-dependent inference cost increases with the number of learned domains. We therefore profile the complete online inference path using FP32 and batch size 1. Persistent state is reported separately from online peak GPU memory because the retained historical statistics are required for subsequent continual learning but are not loaded by the Retrieval path. Latency and throughput are averaged over five independent repetitions, each containing 50 warm-up iterations and 500 measured iterations.

Table A9: Resource comparison at the final Order-1 stage (top) and empirical stream-length scalability of KPKW from one to five learned domains (bottom). Persistent state is reported separately from online peak GPU memory. The upper block compares the final-stage resource and performance characteristics of the evaluated methods, with the best and second-best values marked in bold and underline, respectively. The lower block reports the measured scaling behavior of KPKW and is therefore not ranked. Training-time Coordination cost is reported separately in Table A18.

Method	Params (M) ↓	Persistent state (GiB) ↓	FLOPs (G/img) ↓	Latency (ms/img) ↓	Throughput (img/s) ↑	Peak GPU memory (GiB) ↓	Seen-Avg ↑	Unseen-Avg ↑
PAEMA (Li et al., 2024)	**85.84**	0.00034	23.1	13.72 ± 0.20	73.9 ± 0.4	**0.33**	60.3/70.2	69.8/63.2
DRE (Liu et al., 2025a)	107.85	**0.00000**	**22.2**	13.91 ± 0.44	72.5 ± 1.9	0.83	58.7/70.8	62.9/58.1
DKUA (Liu et al., 2025b)	248.31	0.28614	61.9	27.25 ± 0.03	36.7 ± 0.1	1.36	59.5/69.5	67.8/60.3
KPKW	182.19	5.71	42.48	42.70 ± 1.69	22.83 ± 0.18	0.716	**62.2/74.2**	**73.9/68.0**

Learned domains t	Persistent state (GiB)	FLOPs (G/img)	Latency (ms/img)	Peak GPU memory (GiB)
1	1.399	22.682	8.64 ± 1.15	0.416
2	2.476	27.632	17.20 ± 0.51	0.491
3	3.553	32.582	26.76 ± 1.18	0.563
4	4.630	37.532	34.72 ± 1.40	0.641
5	5.707	42.482	42.70 ± 1.69	0.716

The baseline resource entries follow the corresponding final-stage Order-1 resource measurements, with the reported memory-bank sizes converted from MB to GiB by division by 1024. The same resource definitions are used throughout Table A9.

Inference cost and stream-length scalability. Across the evaluated stream from one to five learned domains, the measurements exhibit approximately linear growth in adapter-dependent parameters and computation. Each additional frozen domain update adds approximately 19.1M inference parameters and 4.95 G FLOPs. As the number of learned domains increases from one to five, latency rises from 8.64 ± 1.15 to 42.70 ± 1.69 ms/image, while peak GPU memory increases only from 0.416 to 0.716 GiB. Throughput correspondingly decreases from 130.59 ± 5.18 to 22.83 ± 0.18 images/s.

At the final stage, KPKW achieves the strongest Seen-Avg and Unseen-Avg among the compared methods. Its 42.48 G FLOPs remain below DKUA's 61.9 G, and its 0.716 GiB online peak GPU memory is the second lowest among the compared methods, despite jointly applying all accumulated frozen domain updates. The main

efficiency costs of the current implementation are therefore the accumulated persistent state and the branch-wise inference latency rather than online GPU-memory consumption.

The persistent-state budget should not be interpreted as deployment-time memory consumption. Most of this state consists of historical statistics retained for future Coordination, and these statistics are not required by the unified encoder during online Retrieval. Moreover, the compression results in Section D.5 show that the dominant statistics budget can be reduced aggressively when storage is constrained. Thus, the current full-matrix configuration favors maximum retention and generalization accuracy, while the compressed alternatives provide a practical path toward substantially lower continual-learning storage.

Overall, KPKW exposes a clear accuracy–efficiency trade-off: the current implementation incurs increasing storage and branch-wise computation as the domain stream grows, but this additional state enables the strongest historical retention and unseen-domain generalization in the comparison. The measured scaling behavior also makes the source of the overhead explicit and provides concrete directions for future acceleration, such as compressed statistical references or more efficient joint execution of the retained updates.

D.9 Targeted Verification of the Auxiliary Components

The main component analysis shows that output-structure preservation produces only a modest change in endpoint accuracy. Since its primary purpose is to preserve the historical output structure rather than directly maximize the current retrieval objective, we further examine the quantity it explicitly constrains, namely historical output-second-moment drift. We separately evaluate overlap regularization using its optimization objective and the resulting concentration of channel-wise contributions.

Table A10: Targeted verification of the auxiliary components in KPKW. The upper block evaluates historical response preservation and output-structure preservation using both endpoint performance and historical output-second-moment drift. The KPKW (Full) row additionally includes S_B-based identity-related gradient guidance. The lower block evaluates overlap regularization through its direct objective and the distribution of channel-wise contributions. In the upper block, the best and second-best results are marked in bold and underline, respectively. In the lower two-row comparison, the better result is highlighted in bold.

Variant	Historical response preservation	Output-structure preservation	Seen-Avg	Unseen-Avg	Historical output-second-moment drift ↓
Task-only composition	×	×	57.6/71.0	68.3/61.3	2.4232×10^9
Output-structure preservation	×	✓	59.2/72.3	69.0/61.1	2.9654×10^8
Historical response preservation	✓	×	<u>61.8/73.6</u>	**74.0/67.5**	$\mathbf{2.3896 \times 10^8}$
KPKW (Full)	✓	✓	**62.0/73.8**	<u>73.3/66.9</u>	<u>2.4761×10^8</u>

Composition variant	Overlap objective ↓	Normalized contribution entropy ↓	Mean maximum contribution ↑	Seen-Avg	Unseen-Avg
Channel-wise composition	370.820	0.957	0.281	54.8/68.3	64.6/57.7
+ overlap regularization	**95.448**	**0.713**	**0.421**	**57.6/71.0**	**68.3/61.3**

Output-structure preservation directly affects the quantity it is designed to constrain. Compared with task-only composition, using output-structure preservation alone reduces historical output-second-moment drift from 2.4232×10^9 to 2.9654×10^8, corresponding to an 87.8% reduction, while also improving Seen-Avg from 57.6/71.0 to 59.2/72.3. This confirms that the mechanism substantially preserves historical output structure even though its marginal effect on endpoint retrieval accuracy is relatively modest.

Historical response preservation produces the lowest drift, corresponding to a 90.1% reduction from the task-only configuration, and achieves the strongest Unseen-Avg in this controlled comparison. Full KPKW maintains a similarly low level of output drift, with an 89.8% reduction relative to task-only composition, while achieving the strongest Seen-Avg. Its drift is 3.6% higher than that of historical response preservation alone. Therefore, the results do not suggest that output-structure preservation independently improves the historical response constraint itself. Instead, they show that it provides a complementary structural signal that directly stabilizes historical output statistics as part of the overall Coordination objective.

Overlap regularization exhibits an equally clear effect on its intended optimization behavior. It reduces the measured overlap objective from 370.820 to 95.448, a 74.3% reduction. At the same time, normalized contribution entropy decreases from 0.957 to 0.713 and the mean maximum contribution increases from 0.281 to 0.421. These changes indicate that excessive co-activation among accumulated domain updates is reduced and that their channel-wise contributions become more differentiated. Importantly, this structural change is accompanied by consistent performance gains: Seen-Avg improves from 54.8/68.3 to 57.6/71.0, while Unseen-Avg improves from 64.6/57.7 to 68.3/61.3.

Overall, these targeted diagnostics confirm that the auxiliary mechanisms in KPKW operate in the intended manner rather than merely adding optimization terms. Output-structure preservation substantially reduces historical output drift, while overlap regularization discourages excessive overlap among accumulated updates and improves both retention and generalization. Together with historical response preservation and identity-related gradient guidance, these mechanisms provide complementary signals for coordinating frozen domain-specific transformations.

D.10 FROZEN-DOMAIN-UPDATE CONTRIBUTION ANALYSIS

To examine whether the final coordinated model actually relies on the transformations retained from different learning stages, we perform a leave-one-frozen-update-out evaluation. For each frozen domain-specific update M_i, we suppress only its contribution during inference while keeping the backbone, all remaining frozen updates, and the learned channel-wise composition coefficients unchanged. We then measure the resulting change in each learned domain and in Unseen-Avg. This analysis is a post-hoc diagnosis of the fixed final model rather than a retrained ablation.

All conditions are evaluated using the same final Order-1 checkpoint and identical evaluation samples. The intact model obtains 62.0/73.7 on Seen-Avg and 73.4/67.0 on Unseen-Avg. Table A11 reports the effect of suppressing each frozen update, where every delta is computed as the masked result minus the intact result.

Table A11: Leave-one-frozen-update-out contribution analysis of KPKW. Each entry reports mAP/R@1. Deltas are computed relative to the intact final model under the same checkpoint and evaluation protocol. This analysis measures the contribution of each retained update within the fixed coordinated model and does not involve retraining after an update is removed.

Suppressed domain update (source domain)	**Seen-Avg**	**Δ Seen-Avg**	**Unseen-Avg**	**Δ Unseen-Avg**	**Δ source domain**
None (intact)	62.0/73.7	+0.0/+0.0	73.4/67.0	+0.0/+0.0	–
M_1 (Market-1501)	57.9/70.9	-4.1/-2.8	71.2/64.8	-2.2/-2.2	-14.4/-8.4
M_2 (CUHK-SYSU)	61.4/73.5	-0.6/-0.2	72.8/66.4	-0.6/-0.6	-0.7/-0.6
M_3 (LPW s2)	60.0/72.3	-2.0/-1.4	72.6/66.4	-0.8/-0.6	-7.7/-4.8
M_4 (MSMT17)	58.5/70.4	-3.5/-3.3	71.2/65.1	-2.2/-1.9	-10.5/-12.6
M_5 (CUHK03)	49.7/62.0	-12.3/-11.7	62.9/56.8	-10.5/-10.2	-39.1/-41.3

Suppressing any of the five frozen updates reduces both Seen-Avg and Unseen-Avg. This provides direct evidence that all retained domain-specific transformations make measurable contributions to the final coordinated representation rather than becoming inactive after subsequent stages. Their contributions are clearly asymmetric. Removing M_5 produces the largest degradation, followed by M_1 and M_4, whereas suppressing M_2 yields a smaller but consistently non-zero effect.

The per-domain maps further reveal a clear source-domain relationship. For each M_i, its suppression produces the largest or one of the largest performance decreases on the corresponding source domain. This behavior supports the role of frozen domain updates in retaining transformations learned from their respective stages. Importantly, their utility is not restricted to those source domains.

Suppressing every individual update also reduces Unseen-Avg, with decreases ranging from 0.6 to 10.5 mAP and from 0.6 to 10.2 R@1. Thus, the retained transformations are also reused when forming representations for domains that were never observed during lifelong learning. They therefore do not behave as isolated domain-specific experts. Instead, KPKW coordinates multiple retained transformations so that they jointly contribute to a single unified encoder.

Overall, the leave-one-update-out analysis provides direct functional evidence for the central design of KPKW. Retention preserves distinct stage-specific transformations, while subsequent Coordination keeps these frozen updates active and organizes their joint contribution. The resulting unified representation benefits both historical-domain retention and unseen-domain generalization without requiring domain labels at inference.

D.11 STATISTICS-GUIDED COORDINATION STRATEGY ABLATION

Table A12 compares different strategies for coordinating the accumulated frozen domain updates. Starting from direct summation, we progressively introduce learnable update-level weighting, channel-wise composition, overlap regularization, and finally retained-statistics guidance. This comparison examines not only whether the frozen updates should be preserved, but also how they should be organized to work together effectively.

Table A12: Statistics-guided Coordination strategy ablation. Results are reported as mAP/R@1. The best and second-best results are marked in bold and underline, respectively. "Static" indicates that the resulting composition is fixed during Retrieval and does not require domain-dependent routing at inference.

Strategy	**Granularity**	**Static**	**Overlap**	**Retained Statistics**	**Seen-Avg**	**Unseen-Avg**
Direct Sum	none	✓	×	×	18.8 / 29.1	29.6 / 22.8
Global-Scalar Composition	update	✓	×	×	51.0 / 66.0	67.0 / 59.8
Channel-Wise Composition	channel	✓	×	×	55.3 / 68.7	65.4 / 57.6
+ Overlap Regularization	channel	✓	✓	×	<u>57.8 / 71.2</u>	<u>67.8 / 60.2</u>
KPKW	channel	✓	✓	✓	**62.2 / 74.2**	**73.9 / 68.0**

Directly summing all frozen domain updates performs poorly, with only 18.8/29.1 Seen-Avg and 29.6/22.8 Unseen-Avg. This result shows that simply retaining and simultaneously applying previously learned transfor-

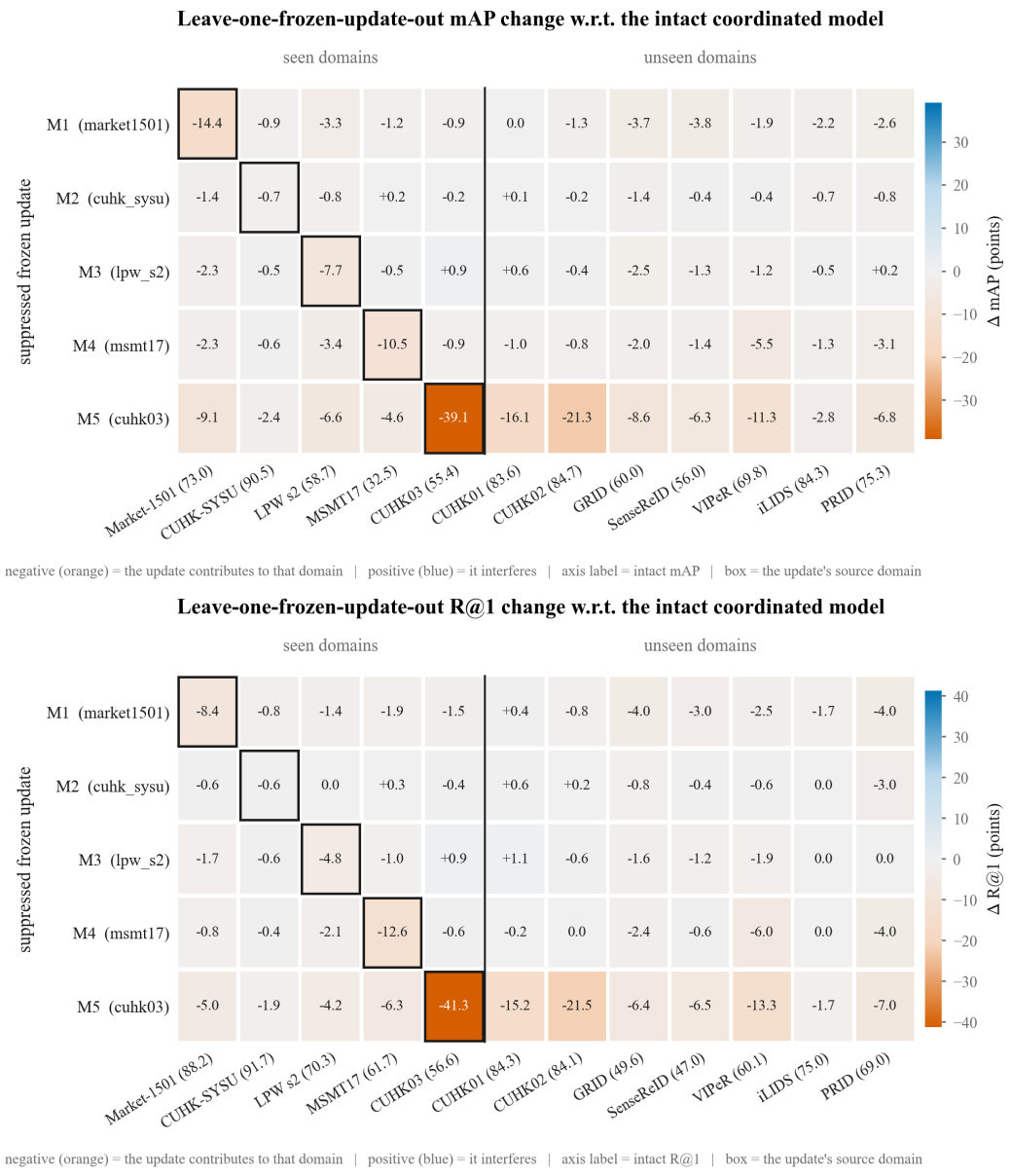

Figure A5: **Frozen-domain-update contribution maps.** The top and bottom panels report the per-domain changes in mAP and R@1, respectively, after suppressing one frozen domain-specific update at a time. Negative values indicate performance degradation relative to the intact model and therefore a positive contribution from the suppressed update. Outlined cells identify the source domain associated with each retained update.

mations is not sufficient. The accumulated updates require an explicit mechanism that determines how their contributions should be combined.

Introducing a learnable scalar for each frozen update substantially improves performance to 51.0/66.0 Seen-Avg and 67.0/59.8 Unseen-Avg. This confirms that even coarse update-level coordination is considerably more effective than direct summation. Replacing global scalar weights with channel-wise composition further improves Seen-Avg to 55.3/68.7, although its Unseen-Avg decreases slightly to 65.4/57.6. Therefore, finer-grained composition alone does not guarantee better generalization.

The benefit of channel-wise composition becomes clearer once overlap regularization is introduced. By discouraging excessive overlap among the channel contributions of different frozen updates, Seen-Avg increases to 57.8/71.2 and Unseen-Avg to 67.8/60.2, outperforming both unconstrained channel-wise composition and global-scalar composition on all four endpoint metrics. This result indicates that fine-grained coordination is most effective when the contributions of accumulated transformations are explicitly organized rather than allowed to overlap freely.

Retained-statistics guidance provides the largest additional improvement. Compared with channel-wise composition with overlap regularization, full KPKW improves Seen-Avg by 4.4/3.0 points and Unseen-Avg by 6.1/7.8 points, reaching 62.2/74.2 and 73.9/68.0, respectively. These gains show that the retained historical statistics provide substantially more informative guidance than the structural composition mechanism alone. They allow the frozen updates to be coordinated with reference to historical response behavior and identity-related structure even after the corresponding samples are no longer available.

Overall, the ablation reveals a clear distinction between retaining past transformations and making them work effectively together. Frozen updates alone are insufficient, and increasing the granularity of their combination is also not enough by itself. KPKW achieves its strongest retention and generalization by combining channel-wise composition, overlap regularization, and retained statistical references. This result directly supports the central design of KPKW, in which past transformations are first retained and then explicitly coordinated so that they remain useful within a unified encoder.

D.12 Accuracy–Trainable-Parameter Budget Analysis

Figure A6 compares seen-domain retention and unseen-domain generalization against the number of trainable parameters required when learning each subsequent domain. This analysis focuses on the optimization budget during continual learning rather than the total persistent state analyzed in Section D.8.

KPKW achieves a favorable accuracy–trainable-parameter trade-off, maintaining strong Seen-Avg and Unseen-Avg performance while optimizing substantially fewer parameters for each subsequent domain than the compared methods. This efficiency follows from the design of KPKW: previously learned domain-specific transformations remain frozen, while later Coordination optimizes only lightweight channel-wise composition coefficients rather than repeatedly updating the accumulated historical transformations.

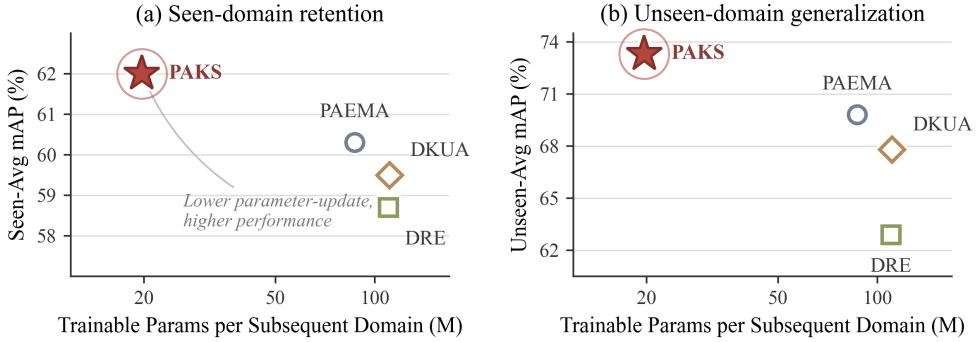

Figure A6: **Accuracy–trainable-parameter budget comparison.** The left and right panels report Seen-Avg and Unseen-Avg mAP, respectively, against the number of trainable parameters for each subsequent domain. The horizontal axis measures parameters actively optimized during subsequent learning rather than the complete persistent state. KPKW maintains strong retention and generalization with a comparatively small trainable-parameter budget, demonstrating an effective accuracy–optimization-efficiency trade-off.

Overall, these results show that the strong retention and generalization of KPKW do not require repeatedly optimizing the full accumulated model. By keeping previously learned transformations fixed and restricting subsequent optimization to lightweight coordination parameters, KPKW maintains a compact trainable budget while effectively reusing historical knowledge.

D.13 SEQUENTIAL VERSUS JOINT IDENTITY-RELATED PROJECTION

KPKW adjusts the gradients of its channel-wise composition coefficients using the principal identity-related directions extracted from the historical between-class scatter. In our default implementation, the domain-specific complement operators are applied sequentially,

$$P^l_{B,\mathrm{seq}} = \prod_{i<t} \left(I - U^l_{B,i} {U^l_{B,i}}^\top \right). \tag{46}$$

Each factor is an orthogonal projector onto the complement of the corresponding historical direction set. When the historical subspaces are not mutually orthogonal, however, their ordered product is generally neither symmetric nor idempotent. Therefore, the sequential operator used in KPKW should be interpreted as the implemented domain-wise gradient-adjustment rule rather than as a single orthogonal projector onto the complement of one common historical subspace.

To examine whether explicitly constructing such a common subspace materially changes the behavior, we additionally introduce a joint orthogonal variant as a diagnostic control. We concatenate the historical bases,

$$C^l = \left[U^l_{B,1}, \ldots, U^l_{B,t-1} \right], \tag{47}$$

obtain an orthonormal basis Q^l for $\mathrm{span}(C^l)$ using rank-revealing SVD, and define

$$P^l_{B,\mathrm{joint}} = I - Q^l {Q^l}^\top. \tag{48}$$

The joint operator is not used in the default KPKW configuration. It serves only as an alternative diagnostic that explicitly removes gradient components lying in the combined span of the retained identity-related directions. The sequential and joint variants use the same training protocol, augmentation, optimization budget, retained statistics, loss weights, and soft-adjustment strength, differing only in the gradient-adjustment operator. The no-adjustment configuration uses $\beta = 0$ under the same setup. Historical gradient energy measures the fraction of the adjusted gradient that remains along the retained historical directions, with lower values indicating weaker interference with these directions.

Table A13: Sequential versus joint identity-related gradient adjustment. Results are reported as mAP/R@1. The sequential product is the gradient-adjustment rule adopted by KPKW, whereas the joint orthogonal operator is evaluated only as an alternative diagnostic. For the operator diagnostics, smaller symmetry error, idempotence error, and historical gradient energy indicate closer orthogonal-projection behavior and weaker gradient energy along retained historical directions.

Operator	Retained Rank	Symmetry Error ↓	Idempotence Error ↓	Historical Gradient Energy ↓	Seen-Avg	Unseen-Avg
No adjustment	0.0	0	0	0.0662	61.9 / 73.5	72.9 / 66.6
Sequential product	3.8	5.07×10^{-2}	2.76×10^{-2}	<u>0.0167</u>	61.9 / <u>73.6</u>	<u>73.0 / 66.6</u>
Joint orthogonal	**13.1**	**0**	1.21×10^{-6}	**0.0157**	**61.9 / 73.7**	**73.3 / 67.0**

The results first verify that the sequential adjustment adopted by KPKW is already effective. Compared with no adjustment, it reduces historical gradient energy from 0.0662 to 0.0167, corresponding to a 74.8% reduction, while improving Seen R@1 from 73.5 to 73.6 and Unseen mAP from 72.9 to 73.0. This shows that the implemented domain-wise adjustment substantially suppresses gradient components along the retained historical directions while maintaining strong seen-domain retention and unseen-domain generalization.

The operator diagnostics also confirm the expected mathematical distinction. The sequential product exhibits non-zero symmetry and idempotence errors and therefore should not be interpreted as a single joint orthogonal projector. In contrast, the diagnostic joint construction is symmetric up to numerical precision, has an idempotence error of only 1.21×10^{-6}, and slightly further reduces historical gradient energy to 0.0157.

Despite this mathematical difference, the two adjustment strategies achieve very similar endpoint performance. Relative to the sequential rule used by KPKW, the joint diagnostic changes Seen R@1 only from 73.6 to 73.7 and Unseen-Avg from 73.0/66.6 to 73.3/67.0, while Seen mAP remains identical at 61.9. The small 0.1–0.4 point differences indicate that explicitly constructing a joint orthogonal projector does not materially change the overall retention and generalization behavior.

We therefore retain the sequential domain-wise adjustment as the default mechanism in KPKW. It directly applies the identity-related guidance retained from each historical domain, substantially reduces historical gradient interference, and already achieves performance comparable to the stricter joint orthogonal alternative. The joint construction is reported only as a diagnostic control to clarify the mathematical distinction between sequential

domain-specific adjustment and projection onto the complement of an explicitly combined historical span. These results further support that the effectiveness of the identity-related gradient guidance in KPKW does not depend on treating the sequential operator as a single joint orthogonal projector.

D.14 Effect of Overlap Regularization on Channel Contributions

Table A10 verifies overlap regularization under the controlled channel-composition setting of Section D.9, where neither the preservation constraints nor S_B-based gradient guidance is present. Here, we examine a complementary question: whether overlap regularization continues to affect the channel-wise composition inside the complete KPKW configuration, where historical response preservation, output-structure preservation, and identity-related gradient guidance already jointly shape the composition coefficients.

We vary only $\lambda_{\text{overlap}} \in \{0, 0.01, 0.02\}$ while keeping the remaining training configuration fixed. The default KPKW configuration uses $\lambda_{\text{overlap}} = 0.01$. For frozen domain update M_i at layer l and input channel c, we define its normalized channel contribution as

$$a_i^l(c) = \frac{|m_i^l(c)|}{\sum_j |m_j^l(c)| + \epsilon}. \tag{49}$$

The overlap objective is computed directly from the raw channel-wise composition coefficients,

$$\sum_l \sum_{i<j} \text{mean}\left(|m_i^l \odot m_j^l|\right), \tag{50}$$

whereas normalized contribution entropy and mean maximum contribution are computed from $a_i^l(c)$. All three diagnostics are aggregated over the 49 composition layers shared by the evaluated configurations.

Table A14: Contribution diagnostics under increasing overlap regularization. The three diagnostics are aggregated over the 49 common composition layers, and accuracy is reported as mAP/R@1. Bold and underline indicate the best and second-best diagnostic values, respectively. The accuracy columns are left unranked because this experiment is intended to diagnose the effect of overlap regularization rather than to identify an accuracy-optimal setting.

λ_{overlap}	Overlap objective ↓	Normalized contribution entropy ↓	Mean maximum contribution ↑	Seen-Avg	Unseen-Avg
0	157.286	0.844	0.415	61.7 / 73.5	73.1 / 66.6
0.01 (default)	<u>127.648</u>	<u>0.809</u>	<u>0.436</u>	61.8 / 73.3	72.9 / 66.0
0.02	**117.005**	**0.789**	**0.447**	61.8 / 73.3	73.3 / 67.1

All three diagnostics change monotonically in the intended direction as λ_{overlap} increases. The overlap objective decreases by 25.6% from 157.286 at $\lambda_{\text{overlap}} = 0$ to 117.005 at $\lambda_{\text{overlap}} = 0.02$. At the same time, normalized contribution entropy decreases and the mean maximum contribution increases. These results directly confirm that overlap regularization remains active inside the complete KPKW configuration and progressively discourages excessive magnitude overlap among different frozen updates on the same channels.

The effect is smaller than in the controlled comparison of Table A10. Without overlap regularization, the complete KPKW configuration already reaches a mean maximum contribution of 0.415, compared with only 0.281 in the corresponding controlled channel-composition setting. This indicates that historical response preservation, output-structure preservation, and identity-related gradient guidance already encourage differentiated use of the accumulated frozen transformations. Overlap regularization acts on top of these mechanisms and provides an additional structural bias toward less overlapping channel contributions.

Endpoint accuracy does not follow the same monotonic trend. Seen-Avg remains nearly unchanged across the three settings, while Unseen-Avg varies moderately. We therefore do not interpret this sweep as evidence that stronger overlap regularization necessarily improves retrieval accuracy. Instead, its purpose is to verify the mechanism directly: increasing λ_{overlap} consistently reduces the measured overlap and produces more concentrated channel-wise contributions. The default $\lambda_{\text{overlap}} = 0.01$ is retained in the main KPKW configuration.

Importantly, more concentrated contributions should not be interpreted as a hard partition of channels among frozen domain updates. Even at $\lambda_{\text{overlap}} = 0.02$, the mean maximum contribution is only 0.447 rather than approaching 1. Thus, the coordinated representation remains a soft shared composition on average, while overlap regularization discourages excessive co-activation rather than enforcing mutually exclusive channels.

Overall, this analysis provides direct mechanistic evidence for the role of overlap regularization in KPKW. The regularizer does not force historical transformations into isolated branches. Instead, it reduces unnecessary overlap while preserving their joint participation in the unified representation, complementing the statistical references that guide how the accumulated frozen updates should work together.

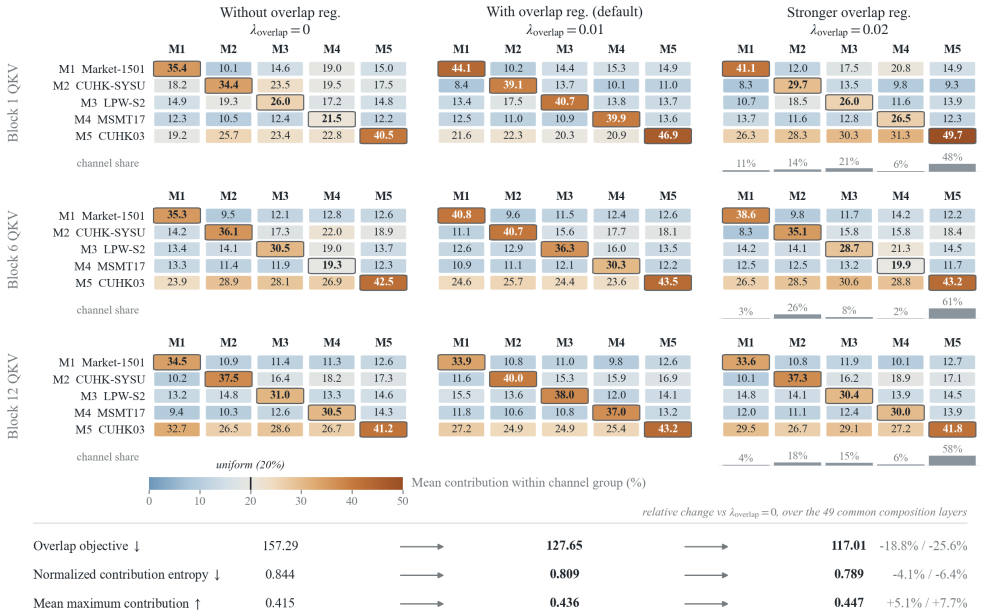

Figure A7: **Effect of overlap regularization on channel contributions.** The three columns visualize channel-wise compositions trained with $\lambda_{\text{overlap}} = 0$, 0.01 (default), and 0.02 in the complete KPKW configuration while keeping all other settings fixed. Rows correspond to early, middle, and late Transformer blocks, following the same layer selection as Figure 7 in the main paper. Within each panel, rows denote the five frozen domain updates and columns denote the five channel groups determined by the dominant update in the default model. The same grouping is reused across all three settings to enable direct comparison. Each cell reports the mean normalized contribution within the corresponding channel group. A shared diverging scale uses the uniform 20% contribution as its neutral point. Across the 49 common composition layers, stronger overlap regularization monotonically reduces the overlap objective from 157.29 to 117.01, decreases normalized contribution entropy from 0.844 to 0.789, and increases the mean maximum contribution from 0.415 to 0.447.

D.15 Historical Response Preservation Analysis

Table A15 provides a metric-wise analysis of the retained statistical references used during Coordination. Starting from task-only composition, we progressively introduce Σ-based historical response preservation, output-structure preservation, and S_B-based identity-related gradient guidance. While the main paper reports the corresponding component progression in compact mAP/R1 form, Table A15 separates the four aggregate metrics to clarify the contribution of each mechanism.

Table A15: Historical response preservation analysis on Training Order-1. Results are reported separately for seen-domain retention and unseen-domain generalization. The best and second-best results are marked in bold and underline, respectively.

Variant	Σ Response Preservation	Output-Structure Preservation	S_B Gradient Guidance	Seen mAP	Seen R1	Unseen mAP	Unseen R1
Task-Only Composition	×	×	×	57.8	71.2	67.8	60.2
+ Σ Response Preservation	✓	×	×	61.8	73.6	**74.0**	67.5
+ Output-Structure Preservation	✓	✓	×	61.6	73.3	73.8	67.7
KPKW (Full)	✓	✓	✓	**62.2**	**74.2**	73.9	**68.0**

Σ-based historical response preservation provides the largest individual improvement over task-only composition. It raises Seen mAP/R1 from 57.8/71.2 to 61.8/73.6 and Unseen mAP/R1 from 67.8/60.2 to 74.0/67.5. These gains show that the recorded layer-wise input second moments provide effective historical references for coordinating the accumulated frozen transformations after the corresponding samples are no longer available.

Adding output-structure preservation produces only small changes in endpoint accuracy, with Seen-Avg changing from 61.8/73.6 to 61.6/73.3 and Unseen-Avg from 74.0/67.5 to 73.8/67.7. This behavior is consistent with the targeted analysis in Section D.9, where output-structure preservation substantially reduces historical output-second-moment drift even though its marginal contribution to retrieval accuracy is modest. Its role is therefore better understood as a complementary structural constraint rather than an independent source of large endpoint gains.

Finally, introducing S_B-based identity-related gradient guidance yields the strongest overall configuration. Full KPKW reaches 62.2/74.2 on Seen-Avg and 73.9/68.0 on Unseen-Avg, achieving the best Seen mAP, Seen R1, and Unseen R1, while its Unseen mAP remains only 0.1 point below the best value obtained by the Σ-only variant. This indicates that the identity-related guidance further improves historical retention while preserving strong cross-domain generalization.

Overall, the results show that the three mechanisms play complementary roles. Historical response preservation provides the main performance gain, output-structure preservation stabilizes historical output statistics, and identity-related gradient guidance further improves the balance between retaining previously learned transformations and adapting the coordinated representation. Their combination gives KPKW the strongest overall retention–generalization trade-off among the evaluated configurations.

D.16 Isolation and Weight Sensitivity of the Coordination Losses

The Coordination objective combines the task loss with three auxiliary losses,

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{ov}}\mathcal{L}_{\text{overlap}} + \lambda_{\text{mem}}\mathcal{L}_{\text{memory}} + \lambda_{\text{out}}\mathcal{L}_{\text{output}}. \tag{51}$$

The preceding analyses show that output-structure preservation directly reduces historical output-second-moment drift and that overlap regularization remains active inside the complete KPKW configuration. We further investigate two questions. First, once historical response preservation is already present, how do output-structure preservation and S_B-based identity-related gradient guidance contribute individually and jointly to the final model? Second, how sensitive is KPKW to the three Coordination-loss weights?

All comparisons follow the same Order-1 protocol, data augmentation, optimization budget, LoRA capacity, $\rho_B = 0.30$, retained statistics, and evaluation setting, with only the factor under investigation changed. We additionally report two diagnostic quantities after the final Coordination stage. *Response drift* is the quantity directly constrained by λ_{mem}, while *historical output-second-moment drift* is directly constrained by λ_{out}. Both are accumulated in the raw units of their corresponding objectives, so their absolute magnitudes should only be compared within the same diagnostic column.

Table A16: Controlled 2×2 isolation of output-structure preservation and S_B-based identity-related gradient guidance. All four configurations keep $\lambda_{\text{ov}} = 0.01$ and $\lambda_{\text{mem}} = 1.0$, and accuracy is reported as mAP/R@1. The KPKW (Full) row corresponds to the endpoint of this controlled ablation. Bold and underline indicate the best and second-best values only for the diagnostic drift quantities. Accuracy values are intentionally left unmarked because the observed endpoint differences are small relative to the stochastic variation characterized in Section D.6 and by the repeated-run comparison below. The lower block reports the two main effects and their interaction, where $\Delta_{\text{out}} = Y_{1,0} - Y_{0,0}$, $\Delta_{S_B} = Y_{0,1} - Y_{0,0}$, and $\Delta_{\text{int}} = Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0}$.

Variant	λ_{out}	S_B guidance	Seen-Avg	Unseen-Avg	Response drift ↓	Output-moment drift ↓
Response Preservation	0	×	61.9 / 73.3	73.3 / 67.1	$\mathbf{4.275 \times 10^4}$	2.4957×10^8
Response + Output	10^{-5}	×	61.9 / 73.5	72.9 / 66.6	$\underline{4.294 \times 10^4}$	$\mathbf{2.1916 \times 10^8}$
Response + S_B	0	✓	61.9 / 73.8	73.0 / 66.5	4.347×10^4	2.4998×10^8
KPKW (Full)	10^{-5}	✓	61.8 / 73.3	72.9 / 66.0	4.457×10^4	$\underline{2.2604 \times 10^8}$

Effect	Seen-Avg mAP	Seen-Avg R@1	Unseen-Avg mAP	Unseen-Avg R@1	Output-moment drift
Δ_{out}	0.0	+0.2	−0.4	−0.5	-3.041×10^7
Δ_{S_B}	0.0	+0.5	−0.3	−0.6	$+4.075 \times 10^5$
Δ_{int}	−0.1	−0.7	+0.3	0.0	$+6.465 \times 10^6$

Output-structure preservation consistently affects the quantity it is designed to constrain. Without S_B guidance, it reduces historical output-second-moment drift by 12.2%, from 2.4957×10^8 to 2.1916×10^8. With S_B guidance enabled, the same constraint reduces the drift from 2.4998×10^8 to 2.2604×10^8. Thus, output-structure preservation remains effective at stabilizing historical output statistics when integrated into the complete Coordination setting.

Its effect on endpoint retrieval accuracy is much smaller. Adding output-structure preservation to response preservation changes Seen-Avg from 61.9/73.3 to 61.9/73.5 and Unseen-Avg from 73.3/67.1 to 72.9/66.6. Similarly, adding S_B guidance alone gives 61.9/73.8 Seen-Avg and 73.0/66.5 Unseen-Avg. The interaction effects are also small and mixed across the four endpoint metrics. These results indicate that the two mechanisms should not be interpreted as independent sources of large accuracy gains on top of historical response preservation.

Instead, their roles are better reflected by the quantities they directly regulate. Output-structure preservation provides a structural constraint on historical output statistics, while S_B-based guidance modifies the optimization

direction with respect to major historical identity-related directions. Response drift increases moderately from 4.275×10^4 under response preservation alone to 4.457×10^4 under the complete configuration. This shows that the mechanisms constrain different aspects of Coordination rather than simply minimizing the same historical-response objective more strongly.

We next evaluate the sensitivity to the three loss weights.

Table A17: Sensitivity of the three Coordination-loss weights. Each sweep changes one weight while keeping the remaining settings fixed at their default values. The three default rows correspond to the same controlled configuration. The final column reports the quantity directly constrained by the swept weight, namely response drift for λ_{mem}, historical output-second-moment drift for λ_{out}, and the overlap objective for λ_{ov}. Diagnostic quantities are therefore compared only within the corresponding sweep, where bold and underline indicate the best and second-best values. Endpoint accuracy is intentionally left unmarked because these sweeps are used to evaluate sensitivity and operating stability rather than to identify a sharp single-point optimum.

Weight	Value	Seen-Avg	Unseen-Avg	Directly constrained quantity ↓
λ_{mem}	0.5	61.8 / 73.5	73.0 / 66.3	4.698×10^4
	1.0 (default)	61.8 / 73.3	72.9 / 66.0	$\underline{4.457 \times 10^4}$
	2.0	62.0 / 73.4	72.6 / 65.7	$\mathbf{4.415 \times 10^4}$
λ_{out}	0	61.9 / 73.8	73.0 / 66.5	$\underline{2.4998 \times 10^8}$
	10^{-5} (default)	61.8 / 73.3	72.9 / 66.0	$\mathbf{2.2604 \times 10^8}$
	10^{-4}	61.9 / 73.7	73.0 / 67.1	2.5063×10^8
λ_{ov}	0	61.7 / 73.5	73.1 / 66.6	157.29
	0.01 (default)	61.8 / 73.3	72.9 / 66.0	$\underline{127.65}$
	0.02	61.8 / 73.3	73.3 / 67.1	**117.01**

The diagnostic quantities behave consistently with the intended roles of the corresponding losses. Increasing λ_{mem} from 0.5 to 2.0 progressively reduces response drift from 4.698×10^4 to 4.415×10^4, a reduction of approximately 6.0%. However, stronger response preservation does not monotonically improve retrieval accuracy. The three evaluated values instead produce similar endpoint performance, indicating a relatively broad operating region around the default setting.

A similar pattern is observed for overlap regularization. Increasing λ_{ov} from 0 to 0.02 monotonically reduces the overlap objective from 157.29 to 117.01, a reduction of approximately 25.6%. Endpoint accuracy, however, remains relatively flat across the three settings. This behavior is consistent with the intended role of overlap regularization as a soft constraint that discourages excessive co-activation among accumulated updates rather than forcing a hard channel partition.

The output-structure weight exhibits a different diagnostic trend. Historical output-second-moment drift reaches its minimum at $\lambda_{\mathrm{out}} = 10^{-5}$. Increasing the weight further to 10^{-4} raises the drift to 2.5063×10^8, slightly above the value obtained when the term is disabled. Thus, increasing the output constraint beyond the default does not lead to stronger preservation of its target statistic. The default value therefore provides a stable operating point for this structural constraint rather than reflecting an endpoint-accuracy optimum.

To put these small endpoint differences into context, Section D.6 reports multi-run standard deviations of at most 0.5 points for mAP and 0.9 points for R@1. We also repeat the default configuration under the same training-related settings while changing only the enabled read-only diagnostic exports. The two runs obtain 61.8/73.3 versus 61.9/73.6 Seen-Avg and 72.9/66.0 versus 73.0/66.6 Unseen-Avg, corresponding to a maximum difference of 0.6 points. Most endpoint differences in Tables A16 and A17 are of a comparable scale, while the largest pairwise deviations reach 1.1 points in Unseen R@1. We therefore do not interpret small differences among neighboring weight settings as evidence of a statistically meaningful single-point optimum.

Overall, the three sweeps indicate that KPKW is not highly sensitive to small changes in the Coordination-loss weights. Stronger constraints more aggressively control their corresponding diagnostic quantities in some cases, but this does not translate monotonically into higher retrieval accuracy. The selected values instead lie in a stable operating region that maintains a practical balance among historical response preservation, output-structure stability, and channel organization without requiring sharp hyperparameter tuning.

Together with the controlled isolation in Table A16, these results further clarify the complementary roles of the Coordination mechanisms. Historical response preservation provides the primary historical reference. Output-structure preservation explicitly stabilizes historical output statistics. Overlap regularization organizes the channel-wise contribution pattern. S_B-based guidance adjusts optimization with respect to major historical identity-related directions. Their joint integration allows KPKW to coordinate accumulated frozen transfor-

mations through a unified encoder while remaining robust to moderate variations of the corresponding loss weights.

D.17 Training-Time Compute and Memory Overhead of Coordination

Section D.8 analyzes the cost of KPKW after training, including persistent state, inference FLOPs, latency, and online peak memory. We further examine the additional computational cost incurred during continual learning, where historical statistics are recorded and accumulated frozen updates are coordinated. The results show that KPKW achieves substantial retention and generalization improvements with a manageable increase in training cost, while the recording of historical statistical references itself introduces only minor overhead.

We consider three configurations under Training Order-1 with the same backbone, LoRA capacity, batch size, epoch budget, Coordination budget, data order, augmentation, and evaluation protocol. **Task-only composition** optimizes the channel-wise composition coefficients without historical statistical references. **Response preservation** additionally records Σ_t and introduces historical response preservation. **KPKW (Full)** further incorporates overlap regularization, output-structure preservation, and S_B-based identity-related gradient guidance with $\lambda_{\text{ov}} = 0.01$, $\lambda_{\text{mem}} = 1.0$, $\lambda_{\text{out}} = 10^{-5}$, $\rho_B = 0.30$, and $\beta = 0.75$.

Each learning stage is divided into three phases. *Acquisition* trains the current-domain LoRA update and classifier. *Statistics recording* collects and stores Σ_t and $S_{B,t}$. *Coordination* optimizes the channel-wise composition coefficients while all accumulated domain updates remain frozen. Total training time is the sum of these three phases. Evaluation and checkpoint serialization are excluded from the training-time measurements. Each timing interval is bracketed by `torch.cuda.synchronize()`, and peak memory is measured after resetting the CUDA peak-memory statistics.

Stage 1 contains no previous domain updates to coordinate and is therefore excluded from the cumulative results over stages 2–5. Its acquisition times are 0.735, 0.723, and 0.732 h for task-only composition, response preservation, and full KPKW, respectively. Statistics recording at this stage requires 0, 18.7, and 27.3 s for the three configurations.

Table A18: Training-time compute and memory overhead of Coordination on Training Order-1. The upper block reports cumulative results over stages 2–5, where historical updates are available for Coordination. The lower block resolves full KPKW stage by stage. *Prior domains* denotes the number of accumulated frozen domain updates participating in Coordination. Per-step time is the Coordination wall-clock time divided by the corresponding number of optimization steps. Accuracy is reported as mAP/R@1, with the best and second-best results marked in bold and underline, respectively.

Variant	Acquisition (h)	Statistics recording (h)	Coordination (h)	Total training (h)	Coordination peak memory (GiB)	Seen-Avg	Unseen-Avg
Task-only composition	1.718	0.000	0.559	2.277	9.57	57.8 / 71.2	67.8 / 60.2
Response preservation	1.745	0.016	1.165	2.925	13.62	<u>61.8 / 73.6</u>	**74.0** / <u>67.5</u>
KPKW (Full)	1.833	0.028	1.889	3.750	21.09	**62.2 / 74.2**	<u>73.9</u> / **68.0**

Stage t	Prior domains	Coordination steps	Coordination (min)	Per-step (s)	Coordination peak memory (GiB)	Statistics recording (s)
2	1	340	5.66	0.999	10.67	16.5
3	2	1020	26.42	1.554	13.76	23.7
4	3	2060	54.86	1.598	17.18	37.9
5	4	760	26.40	2.084	21.09	22.1

Recording historical statistical references is lightweight. Across stages 2–5, full KPKW spends approximately 100 s on statistics recording, accounting for only 0.74% of its total measured training time. The recording cost remains approximately 3.6–4.9 ms per processed sample across stages. This indicates that retaining the layer-wise input second moments and between-class scatter provides historical references for later Coordination at only a small fraction of the overall training cost.

The additional computation is primarily devoted to Coordination, where the accumulated frozen transformations are organized for joint use. Coordination time increases from 0.559 h for task-only composition to 1.165 h with historical response preservation and to 1.889 h for full KPKW. Its peak memory increases from 9.57 GiB to 21.09 GiB. Importantly, this additional computation is introduced only during training and directly supports the central objective of KPKW: keeping previously learned transformations active and effectively coordinated as new domains arrive.

Measured over stages 2–5, total training time increases from 2.277 h for task-only composition to 3.750 h for full KPKW, corresponding to a 64.7% increase. Acquisition itself is unchanged by the Coordination mechanisms. Fixing acquisition time to that of task-only composition gives an additional training cost of approximately 59.6%. Overall, KPKW requires approximately 60–65% additional training time relative to task-only channel-wise composition without historical statistical references, providing a practical training-cost increase for the substantial gains obtained in continual retention and generalization.

The stage-wise measurements further show a gradual increase in Coordination cost as historical domains accumulate. Peak memory increases from 10.67 GiB with one prior domain to 21.09 GiB with four prior domains, with successive increases of approximately 3.09, 3.42, and 3.91 GiB. Per-step Coordination time similarly increases from 0.999 s to 2.084 s, corresponding to a $2.09\times$ increase from one to four prior domains. Even when four historical domain updates participate simultaneously, the complete Coordination procedure remains feasible within the available single-GPU memory budget.

The total Coordination time of an individual stage is jointly determined by the number of accumulated domains and the optimization budget. Stage 4 requires 54.86 min, whereas stage 5 requires 26.40 min despite involving one additional historical domain. This difference follows from their respective optimization budgets, with 2060 Coordination steps at stage 4 and 760 steps at stage 5. The observed overhead therefore reflects the actual amount of Coordination optimization rather than simply increasing with the stage index.

KPKW provides a favorable performance–cost trade-off. Under Training Order-1, task-only composition obtains 57.8/71.2 Seen-Avg and 67.8/60.2 Unseen-Avg, whereas full KPKW reaches 62.2/74.2 and 73.9/68.0. This corresponds to improvements of $+4.4/+3.0$ points on Seen-Avg and $+6.1/+7.8$ points on Unseen-Avg. In particular, the larger gains on unseen domains show that the additional Coordination cost does not merely improve retention of previously learned domains, but also results in a substantially stronger unified representation for cross-domain generalization.

Historical response preservation already accounts for a large fraction of the improvement, reaching 61.8/73.6 Seen-Avg and 74.0/67.5 Unseen-Avg. Full KPKW further improves Seen-Avg by $+0.4/+0.6$ and Unseen R@1 by $+0.5$, while maintaining essentially the same Unseen mAP with only a -0.1 difference. This demonstrates the complementary benefit of the complete Coordination design: historical response preservation provides the primary gain, while output-structure preservation, overlap regularization, and identity-related gradient guidance further organize the accumulated frozen transformations to achieve the strongest overall retention–generalization balance.

Overall, these results show that the computational overhead of KPKW is concentrated in a purposeful Coordination stage rather than in storing or replaying historical samples. With lightweight statistical recording and a moderate increase in training computation, KPKW substantially improves both seen-domain retention and unseen-domain generalization. This supports the practical value of keeping previously learned transformations available and coordinating them explicitly, rather than discarding or overwriting them as continual learning progresses.

E Additional Qualitative Analyses

E.1 Retrieval Visualization

We further visualize the final retrieval behavior of KPKW by comparing it with DKUA under Training Order-1. Both methods use exactly the same queries, gallery pools, filtering rules, and Top-5 ranking protocol, with no re-ranking. The visualization covers all five seen domains and seven unseen domains, providing a comprehensive qualitative view across the complete evaluation suite. For each dataset, four shared queries are deterministically selected from the exported rankings to expose informative retrieval differences while retaining both successful and challenging cases. Green and red borders indicate correct and incorrect identity matches, respectively.

Together, these examples provide a dataset-complete view of how the unified KPKW encoder behaves after sequential learning. The five seen-domain visualizations illustrate that previously learned identity retrieval capability remains available after subsequent domains have been introduced. The seven unseen-domain visualizations further demonstrate the behavior of the same unified encoder under substantial changes in camera style, background, pose, and image quality. This broad coverage is consistent with the quantitative results, showing that coordinating the accumulated frozen transformations enables KPKW to retain useful historical retrieval behavior while maintaining strong transfer to previously unseen domains.

Importantly, the same unified encoder is used throughout these evaluations without domain labels or domain-specific inference routing. The retrieval examples therefore provide an intuitive view of the central goal of KPKW: previously learned transformations are not merely stored after each stage, but remain jointly available to support retrieval across both historical and unseen domains. The complete Top-10 rankings are retained for every query, allowing the displayed retrieval results and match labels to be reproduced directly.

F Computing Infrastructure and Reproducibility

All experiments were conducted on an Arch Linux workstation equipped with two NVIDIA GeForce RTX 4090 GPUs (48 GB each), two Intel Xeon Gold 5218 CPUs with 32 physical cores in total, and 235 GiB of host memory. The software environment consists of Python 3.8.20 and PyTorch 2.0.1 with CUDA 11.7. The NVIDIA driver version is 595.71.05, reporting CUDA 13.2 compatibility.

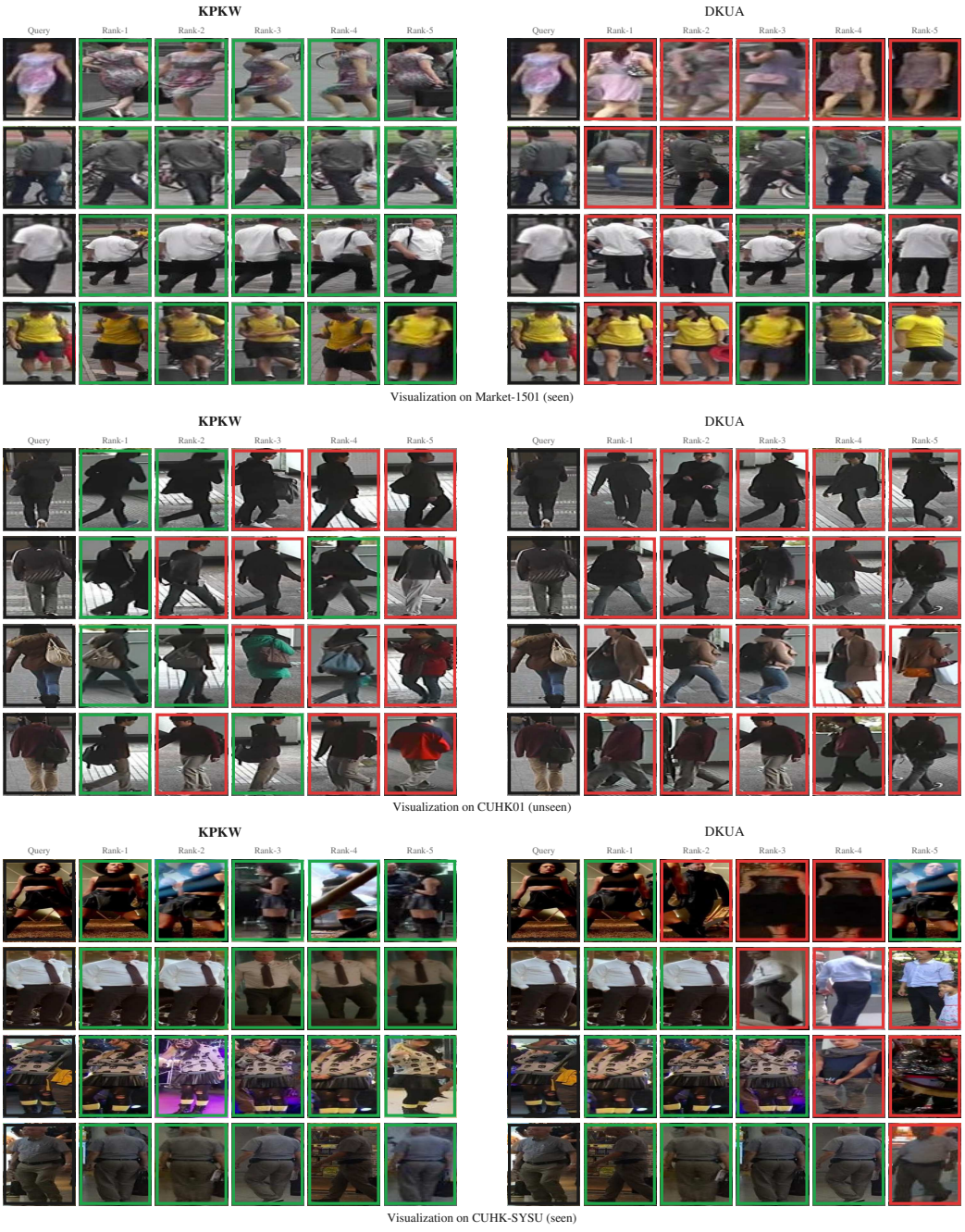

Figure A8: **Top-5 retrieval results of DKUA and KPKW on Market-1501, CUHK01, and CUHK-SYSU.** Green and red borders indicate correct and incorrect identity matches, respectively.

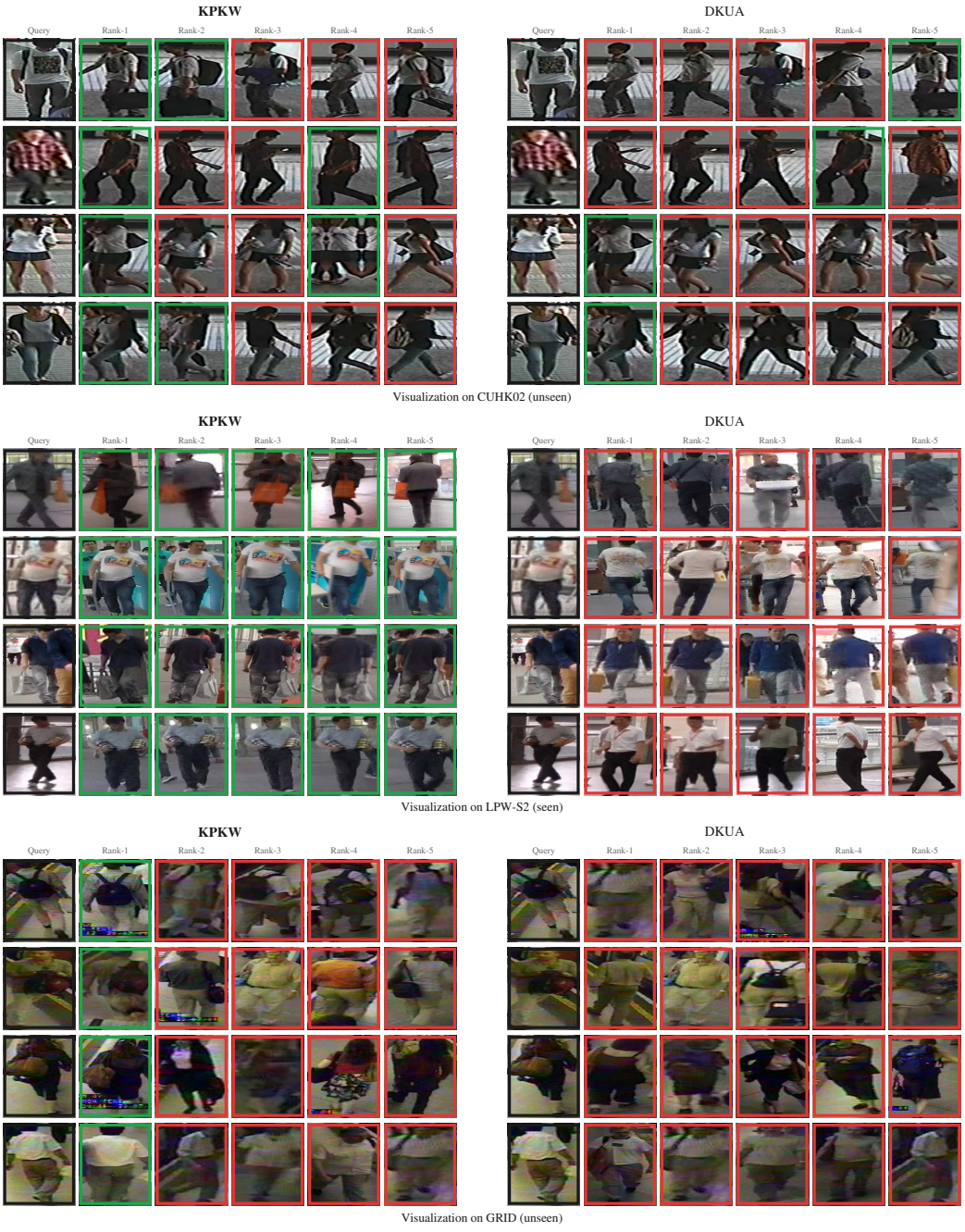

Figure A9: **Top-5 retrieval results of DKUA and KPKW on CUHK02, LPW s2, and GRID.** Green and red borders indicate correct and incorrect identity matches, respectively.

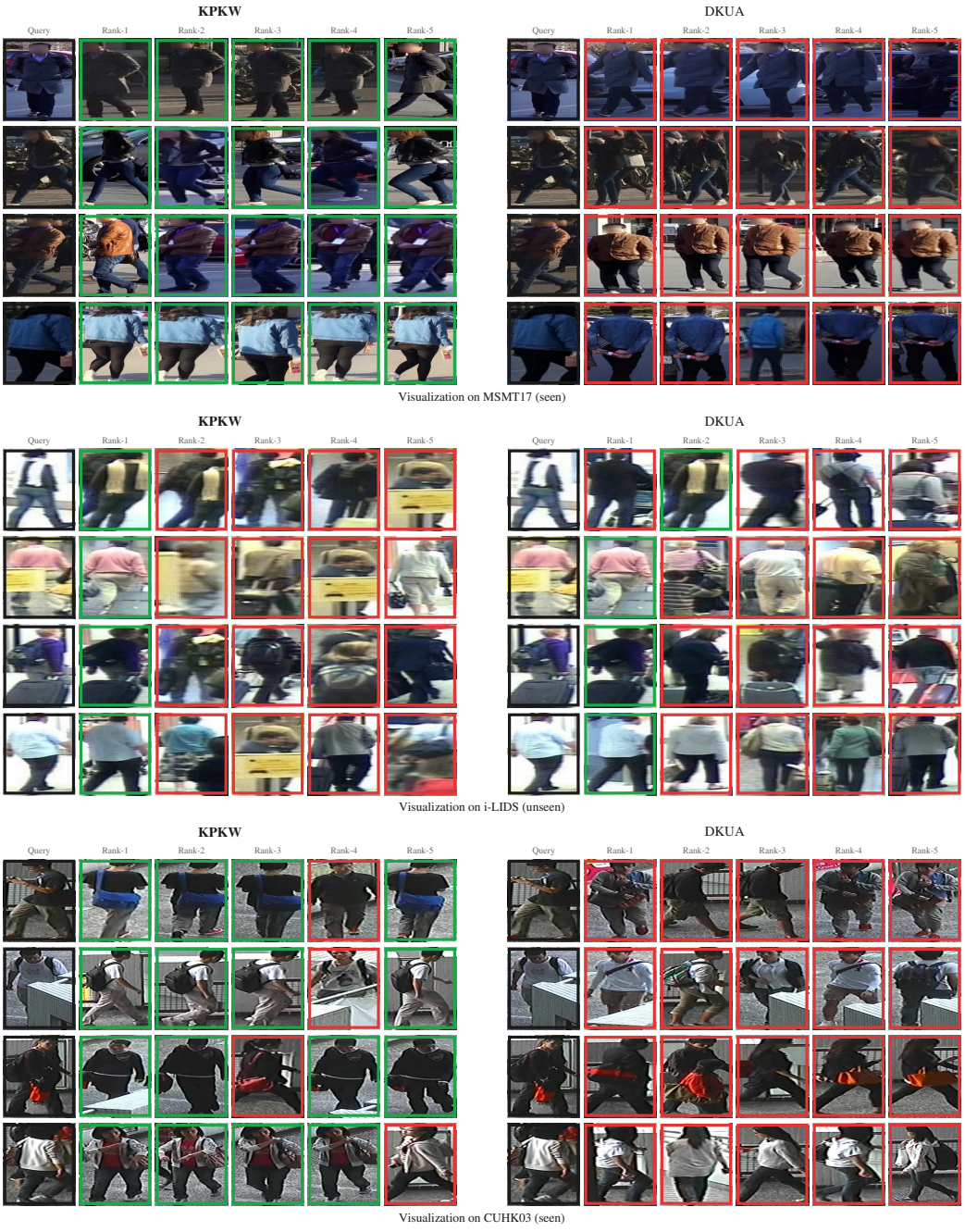

Figure A10: **Top-5 retrieval results of DKUA and KPKW on MSMT17, i-LIDS, and CUHK03.** Green and red borders indicate correct and incorrect identity matches, respectively.



Figure A11: **Top-5 retrieval results of DKUA and KPKW on PRID, SenseReID, and VIPeR.** Green and red borders indicate correct and incorrect identity matches, respectively.

Unless otherwise specified, all experiments use FP32 precision and 256×128 input images. Both single-GPU training and two-GPU data-parallel training are performed on the same workstation. The runtime measurements reported in the supplementary experiments follow the corresponding training protocols and evaluation splits described in the paper. Together, these settings provide a consistent computational environment for the reported continual-learning, ablation, and efficiency experiments.

G Scope and Limitations

KPKW is designed for exemplar-free lifelong learning, where historical samples are unavailable but previously learned transformations can be retained as model-internal states. This design allows historical knowledge to remain directly available for later Coordination and unified retrieval without replaying past images. Accordingly, KPKW is exemplar-free rather than state-free. The frozen domain-specific LoRA updates, channel-wise composition coefficients, and retained statistical references increase as additional domains are learned.

The major persistent-state cost in the default configuration comes from the full statistical matrices. Section D.5 shows that this cost can be substantially reduced through diagonal or low-rank approximations, providing a practical accuracy–storage trade-off. These compressed alternatives, however, approximate the full statistical objective rather than preserving its exact full-matrix form. Exploring more efficient representations of the retained statistical references therefore provides a natural direction for extending KPKW to substantially longer continual streams.

A second scaling factor comes from the explicit joint use of accumulated frozen domain updates. As the number of learned domains increases, more retained transformations participate in the unified forward process and in training-time Coordination. The efficiency analyses in Sections D.8 and D.17 quantify this growth over the evaluated stream and show that KPKW remains practical under the current five-stage protocol. Further reducing the cost of jointly using a larger number of retained transformations, for example through more efficient composition or equivalent deployment forms, is an important direction for longer-term continual learning.

The empirical evaluation covers the two training orders used in the main experiments, controlled component and mechanism analyses under Order-1, stage-wise forgetting and backward-transfer measurements, per-domain unseen evaluations, accuracy–storage trade-offs for compressed statistics, computational-cost analyses, and qualitative retrieval results across all evaluated seen and unseen domains. The baseline results reported in the comparative tables follow their corresponding reference sources rather than being uniformly rerun within our implementation. Taken together, the current evidence supports the central scope of KPKW: retaining previously learned transformations, using historical statistical references to coordinate them, and keeping these transformations jointly available through a unified encoder without historical sample replay.

H Reproducibility Statement

To facilitate reproducibility and allow direct inspection of the proposed method, we provide the core implementation code of KPKW in the **Supplementary Material** section on OpenReview. The provided code corresponds to the key implementation of the proposed framework and complements the methodological descriptions, implementation details, and experimental settings presented in the paper and supplementary material. Together, these materials provide additional implementation-level information for understanding and reproducing the main components of KPKW.

REFERENCES

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Qiwei Li, Kunlun Xu, Yuxin Peng, and Jiahuan Zhou. Exemplar-free lifelong person re-identification via prompt-guided adaptive knowledge consolidation. *International Journal of Computer Vision*, 132 (11):4850–4865, 2024. doi: 10.1007/s11263-024-02110-x. URL https://doi.org/10.1007/s11263-024-02110-x.

Shiben Liu, Huijie Fan, Qiang Wang, Xiai Chen, Zhi Han, and Yandong Tang. Diverse representations embedding for lifelong person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10): 18145–18157, 2025a. doi: 10.1109/TNNLS.2025.3571768.

Shiben Liu, Mingyue Xu, Huijie Fan, Qiang Wang, Yandong Tang, and Zhi Han. Distribution-aware knowledge unification and association for non-exemplar lifelong person re-identification. *arXiv preprint arXiv:2508.03516*, 2025b.