
HypMoE-ReID: Hyperspherical Mixture-of-Experts for Large Scale Person Re-Identification

Kunlun Xu*

Wangxuan Institute of Computer Technology
Peking University
xkl@stu.pku.edu.cn

Haotong Cheng*

Wangxuan Institute of Computer Technology
Peking University
chenght9923@mails.jlu.edu.cn

Yufei Guo

Intelligent Science & Technology Academy
China Aerospace Science and Industry Corporation
yfguo@pku.edu.cn

Jiahuan Zhou[†]

Wangxuan Institute of Computer Technology
Peking University
jiahuanzhou@pku.edu.cn

Abstract

With the rapid growth of surveillance data, Person Re-Identification (ReID) is evolving toward large-scale scenarios, where datasets exhibit substantial increases in both size and diversity. Under this trend, the widely adopted dense backbone architectures, such as Transformers, suffer from an inherent limitation: Shared parameters couple all input images, forcing the model to learn generic representations that are suboptimal for heterogeneous data sources. Although Mixture-of-Experts (MoE) architectures offer a promising solution for handling diverse data, they encounter a fundamental challenge in ReID: Person images exhibit high inter-instance similarity at the coarse level, while discriminative cues lie in subtle fine-grained details that are easily overlooked. As a result, existing MoE methods often suffer severe routing collapse, leading to insufficient expert specialization and limited diverse knowledge acquisition capacity. To address these problems, we propose a novel framework, termed Hyperspherical Mixture-of-Experts for Large Scale Person Re-Identification (HypMoE-ReID). Specifically, motivated by theoretical analysis in hyperspherical space, HypMoE-ReID incorporates two key components: (1) a Token Purification mechanism that suppresses noisy tokens detrimental to expert specialization, and (2) a Routing Diversity Learning scheme that explicitly separates experts on a unit hypersphere to enforce specialization and prevent collapse. Furthermore, to facilitate the investigation in large-scale ReID, we construct a large-scale ReID benchmark comprising 809,701 images by unifying 13 public datasets with significant distribution gaps. Furthermore, extensive experiments demonstrate that HypMoE-ReID outperforms state-of-the-art MoE-based methods by at least 11.3%/10.3% in Average mAP/R@1, and surpasses strong dense backbone models by 2.2%/2.0% in Average mAP/R@1. Our code is available at <https://github.com/PKU-OV3-LAB/NeurIPS2026-HypMoE-ReID>.

*Equal Contribution

[†]Corresponding author

1 Introduction

Person Re-Identification (ReID) aims to retrieve images of the same identity across non-overlapping camera views [1; 2; 3; 4; 5]. With the explosive growth of surveillance data, ReID is inevitably moving toward the large-scale era, where datasets expand not only in identity volume but also in distribution diversity [6; 7; 8; 9]. Despite the strong performance of transformer-based models on small-scale ReID benchmarks [10; 11; 12], such dense architectures suffer from a fundamental limitation. Specifically, the shared parameter space processes all input images uniformly, forcing the model to learn generic representations [13; 14] that are often suboptimal for heterogeneous data sources.

In contrast, Mixture-of-Experts (MoE) architectures [15; 16; 17; 18; 19] substantially expand model capacity, providing a principled solution for breaking the rigidity of generic representations through expert specialization. However, existing studies in vision tasks have shown that MoE architectures are primarily effective in coarse-grained scenarios with naturally separated domains (e.g., Visible-Infrared [20] and Text-Image [21]). Under single-domain visual data, MoE models often suffer from severe routing collapse [22; 23], leading to limited capacity for diverse knowledge acquisition. Current attempts on MoE-based ReID [24; 25; 26; 27] further reinforce these concerns. Specifically, they are mainly evaluated on settings with inherent modality gaps [25] or relatively moderate-scale ReID benchmarks [28; 6]. Therefore, a fundamental question naturally arises: **Can MoE architectures be effectively applied to classic single-modality ReID when the training data reaches a large scale?**

To overcome these challenges, we propose a novel Hyperspherical Mixture-of-Experts for Large Scale Person Re-Identification (HypMoE-ReID) model, a sparse framework aimed at breaking the rigidity of generic representations in dense architectures through sparse computing, while simultaneously resolving the routing collapse that MoE suffers from on fine-grained ReID data, as shown in Figure 1. Specifically, we introduce a Token Purification mechanism that reweights tokens based on their affinity with experts, suppressing noise that pollutes expert specialization. Further, we propose a Routing Diversity Learning scheme that explicitly separates the experts on a unit hypersphere, which includes a Global Separation Loss and a Local Diversification Loss. The former aligns each cluster center with its corresponding expert, while the latter maximizes the angular margin against the hardest negative cluster. Both mechanisms are derived from a hyperspherical formulation, offering

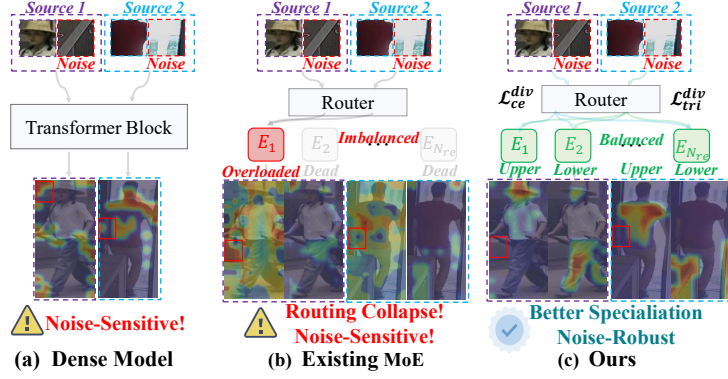


Figure 1: Dense models introduce shared knowledge to process all input content, failing to handle diverse person and background noise features. Existing MoE confronts severe routing collapse under fine-grained ReID data. Our method learns better expert specialization by Routing Diversity Learning, while being robust to diverse fine-grained person and background noise information.

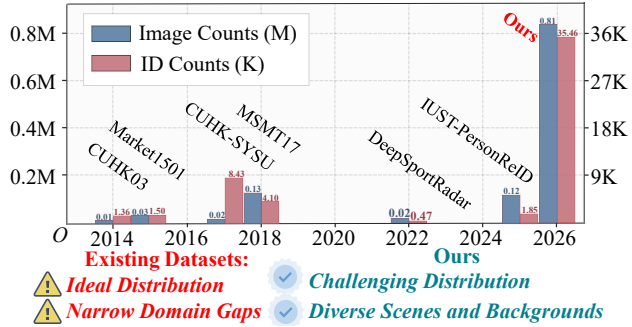


Figure 2: Existing curated ReID datasets feature ideal distributions. We build a challenging large-scale ReID benchmark by reorganizing several ReID datasets.

clean geometric interpretability. Furthermore, as shown in Figure 2, we build a new ReID benchmark with overwhelming data volume and scene diversity compared to existing ones. To this end, we unify 13 publicly available datasets, resulting in over 800 K images. This benchmark allows us to systematically study the behavior of architectures in a realistic large-scale ReID setting where both the limitations of dense backbones and the routing collapse of standard MoE are fully exposed. Extensive experiments demonstrate that HypMoE-ReID outperforms state-of-the-art MoE-based methods and dense backbone models.

In summary, our contributions are fourfold: (1) We present a pioneering and systematic investigation of standard MoE in classic ReID at scale, and reveal that the extreme homogeneity of ReID data causes severe routing collapse in sparse architectures. (2) We propose HypMoE-ReID, a novel framework combining Token Purification and Routing Diversity Learning with strong geometric interpretability which simultaneously breaks the rigidity of dense generic representations and resolves MoE routing collapse. (3) To facilitate the investigation in this area, we build a large-scale ReID dataset, making it a challenging benchmark reflecting real-world distribution. (4) Extensive experiments validate the effectiveness of HypMoE-ReID, which significantly surpasses existing MoE-based backbones and dense architectures, achieving new state-of-the-art on large-scale ReID.

2 Related Works

2.1 Person Re-Identification

Person Re-Identification aims to retrieve the correct individual from non-overlapping camera views, including classic ReID [10; 29], Cloth-Changing ReID [30], Multi-Modal ReID [31], and others. Since TransReID [10] pioneered the implementation of Transformers in ReID, subsequent works have extended the advantages of pretrained models into various sub-fields. For instance, CLIP-ReID [11] enhances multi-modal learning using trainable prompt tokens. Despite these advances, the majority of existing ReID models remain built on dense architectures, which learn limited generic representations and is easily disrupted by distribution shifts. Recent efforts have attempted to mitigate this limitation by training data at scale [32; 33]. However, these works focus primarily on data quantity while leaving the architectural capacity largely unchanged, and building large-scale datasets from scratch remains labor-intensive.

2.2 Mixture-of-Experts

Mixture-of-Experts, characterized by its sparsity, expands model capacity by selectively activating Top-K experts through routing mechanism [15]. Current applications of MoE fall primarily into two branches: NLP MoE and vision MoE. Early explorations in NLP date back to the incorporation of sparsely-gated MoE layers into LSTMs [34; 35], demonstrating the feasibility of scaling model capacity through sparse computing. Recent efforts, such as the DeepSeek family featuring fine-grained expert segmentation and shared expert isolation [36], have triggered the widespread revival of MoE in LLMs and VLMs [37; 38; 39; 40; 41; 42], establishing a new paradigm that outperforms dense architectures. In vision, however, the effectiveness of MoE remains less conclusive. While some studies report clear gains (e.g., Multi task learning [43], Few-shot learning [16], segmentation [23]), others observe limited gains over dense counterparts (e.g., Diffusion [44]), and certain sub-fields have yielded contradictory conclusions [22; 23]. Crucially, existing vision MoE research predominantly focuses on coarse-grained tasks where inter-class differences provide sufficient routing signals. On fine-grained tasks where such signals are inherently weak, the behaviour of MoE remains largely unstudied. FG-MoE [45] proposes a fine-grained classification MoE, but relies on strong priors and overly complex experts that contradict the core motivation of MoE. Similarly, [46] claims to apply MoE to fine-grained classification, yet cascades two ResNet backbones, which deviates from standard definition of MoE. In summary, vision MoE for fine-grained pattern recognition remains largely under-explored.

2.3 Mixture-of-Experts in Person Re-Identification

MoE-based ReID aims to leverage the principle of “Divide and Conquer” to expand model capacity while facilitating the learning of diverse fine-grained features. Existing MoE-based ReID primarily focuses on Multi-Modal ReID [24; 47; 48], Domain-Generalization ReID [22; 25; 49] and Video-

based ReID [50], most of which operate under significant domain gaps that inherently benefit expert specialization. For instance, MoE in DG-ReID allocates dedicated experts to each seen domain [49], while visible-Infrared ReID assigns specific experts to each modality [48; 51]. These studies successfully avoid the expert collapse issue by injecting strong routing priors. However, this pre-defined strategy contradicts the principle of MoE, where experts are expected to specialize from scratch. [52] made a meaningful attempt to diversify the experts through auxiliary loss [53; 54]. However, the actual effects of such losses on expert routing were not thoroughly analyzed, and the evaluation was conducted on multi-modal data. Consequently, the adaptability of standard MoE to classic ReID remains unexamined. A central challenge is whether the standard MoE suffers from routing collapse on extreme homogeneous data, where subtle inter-class differences prevent the router from learning meaningful expert specialization.

3 Preliminaries

Problem Formulation. Person Re-Identification aims to match individuals across non-overlapping camera views. Given query $\mathbf{x}_q$ and gallery set $\mathcal{G} = \{\mathbf{x}_g^i\}_{i=1}^{N_g}$, the model learns $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^D$ such that distance $d(f_\theta(\mathbf{x}_q), f_\theta(\mathbf{x}_g^i))$ is small for images of the same identity and large otherwise.

Transformer-based Dense Architecture. Standard dense backbones, such as ViT and its variants [55; 56; 14], process an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ by splitting it into N patches. These patches are linearly projected and augmented with [CLS] token and position embeddings, yielding token sequence $\mathbf{z}_i^{(0)} \in \mathbb{R}^{B \times (N+1) \times D}$, where B, N, D represent batch size, patch numbers and hidden dimensions, respectively. Tokens are processed through L transformer blocks, each of which consists of multi-head self-attention (MHSA), feed-forward network (FFN) and layer-wise normalizations:

$$\mathbf{z}_i^{(l)} = \text{MHSA}(\text{LN}(\mathbf{z}_i^{(l-1)})) + \mathbf{z}_i^{(l-1)}, \quad \mathbf{z}_o^{(l)} = \text{FFN}(\text{LN}(\mathbf{z}_i^{(l)})) + \mathbf{z}_i^{(l)}, \quad (1)$$

where $\text{FFN}(\cdot)$ is a two-layer MLP with activation $\sigma(\cdot)$. This dense design couples every parameter with every input, limiting model's capacity to capture fine-grained patterns, while making it susceptible to noisy or uninformative patches.

MoE-based Sparse Architecture. To expand model capacity without proportionally increasing overhead, MoE replaces the FFN in selected transformer blocks with a sparse layer, which consists of N_{re} routed experts $\{E_m^r\}_{m=1}^{N_{re}}$, N_{se} shared experts $\{E_n^s\}_{n=1}^{N_{se}}$ and a routing function $R(\cdot)$. Let $\hat{\mathbf{x}}_j^{(l)} \in \mathbb{R}^D$ denote the j -th normalized token. The router computes per-token gating scores as:

$$R^{(l)}(\hat{\mathbf{x}}_j^{(l)}) = \text{Softmax}(\text{Top-K}(\hat{\mathbf{x}}_j^{(l)} \mathbf{W}_{rt}^{(l)}, K)), \quad (2)$$

where $\mathbf{W}_{rt}^{(l)} \in \mathbb{R}^{D \times N_{re}}$ and $\text{Top-K}(\cdot, K)$ retains the K largest logits. The MoE layer output for each token is a weighted combination of its Top-K activated routed experts and all shared experts:

$$\text{MoE}^{(l)}(\hat{\mathbf{x}}_j^{(l)}) = \sum_{m \in \mathcal{T}_K(\hat{\mathbf{x}}_j^{(l)})} [R^{(l)}(\hat{\mathbf{x}}_j^{(l)})]_m \cdot E_m^{r,(l)}(\hat{\mathbf{x}}_j^{(l)}) + \sum_{n=1}^{N_{se}} E_n^{s,(l)}(\hat{\mathbf{x}}_j^{(l)}), \quad (3)$$

where $\mathcal{T}_K(\hat{\mathbf{x}}_j^{(l)})$ is the index set of the Top-K selected routed experts for input token, and $[R^{(l)}(\cdot)]_m$ denotes the routing weight assigned to expert $E_m^{r,(l)}$. Auxiliary loss $\mathcal{L}_{aux}$ [53] is introduced during training to prevent routing collapse. For a batch of N_t tokens in the l -th MoE block, let $f_m^{(l)} = 1/N_t \cdot \sum_{j=1}^{N_t} \mathbb{I}[m \in \mathcal{T}_K(\hat{\mathbf{x}}_j^{(l)})]$, $\bar{p}_m^{(l)} = 1/N_t \cdot \sum_{j=1}^{N_t} [R^{(l)}(\hat{\mathbf{x}}_j^{(l)})]_m$ denote the fraction of tokens dispatched to expert $E_m^{r,(l)}$ and its average routing probability, respectively. $\mathcal{L}_{aux}$ goes as:

$$\mathcal{L}_{aux} = \alpha \cdot N_{re} \cdot \sum_{l=1}^{L/2} \sum_{m=1}^{N_{re}} f_m^{(l)} \cdot \bar{p}_m^{(l)}, \quad (4)$$

where α controls the loss strength, and sum over $L/2$ layers follows the common practice of inserting MoE blocks alternately. Under a perfectly balanced distribution, $f_m^{(l)} = K/N_{re}$ and $\bar{p}_m^{(l)} = 1/N_{re}$, yielding $\mathcal{L}_{aux} = \alpha K L/2$ as the theoretical minimum.

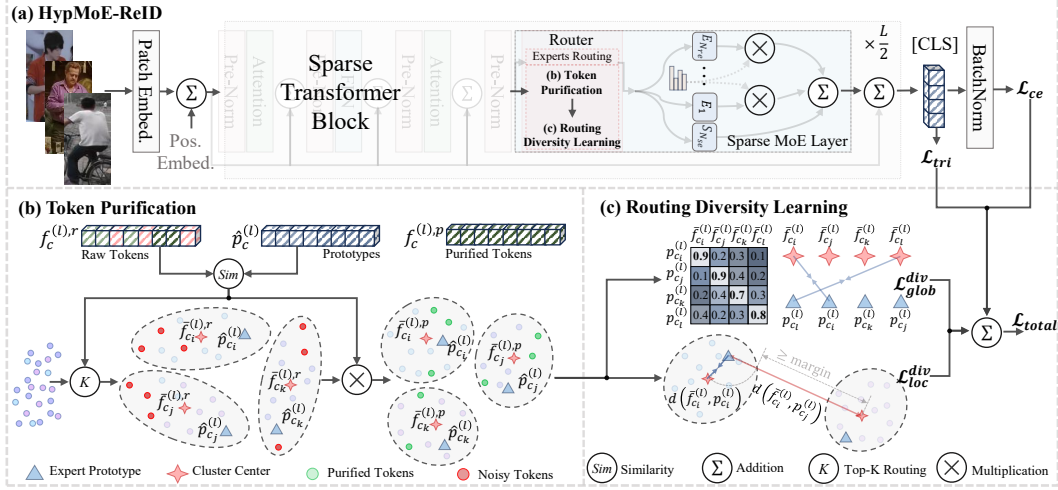


Figure 3: Overview of HypMoE-ReID. Given input image, it is processed by interleaved Dense and Sparse blocks. (a) In each MoE block, token features are ℓ_2 normalized and routed to Top-K experts via cosine similarity with learnable prototypes. (b) Token Purification reweights the routed tokens based on their affinity with the assigned prototype, suppressing noisy embeddings. (c) Routing Diversity Learning then regularizes the prototype-cluster geometry through Global Separation and Local Diversification losses, enforcing experts to specialize on distinct fine-grained patterns.

4 Proposed Methods

In this section, we elaborate on our framework shown in Figure 3 from three aspects: General Architecture, Token Purification, and Routing Diversity Learning, with geometric interpretation also illustrated within each section. Additional proof and algorithmic details are provided in Appendix AB.

4.1 Framework Overview

Following previous architectures [16; 23], we interleave dense transformer blocks with MoE blocks in the backbone, as illustrated in Figure 3(a). Let $\mathcal{M} = \{1, 3, \dots, L-1\}$ denote the index set of sparse blocks. For the l -th block, the self-attention branch remains unchanged as $\mathbf{f}_i^{(l)} = \text{MHSA}(\text{LN}(\mathbf{f}_o^{(l-1)})) + \mathbf{f}_o^{(l-1)}$, while the feed-forward process is redefined as:

$$\mathbf{f}_o^{(l)} = \Gamma(\text{LN}(\mathbf{f}_i^{(l)})) + \mathbf{f}_i^{(l)}, \text{ with } \Gamma(\cdot) = \mathbb{I}[l \in \mathcal{M}] \cdot \text{MoE}(\cdot) + \mathbb{I}[l \notin \mathcal{M}] \cdot \text{FFN}(\cdot). \quad (5)$$

For the l -th MoE layer, we define N_{se} shared experts $\{E_s^{(l)}\}_{s=1}^{N_{se}}$, N_{re} routed experts $\{E_r^{(l)}\}_{r=1}^{N_{re}}$ together with N_{re} learnable expert prototypes $\mathcal{P}^{(l)} = \{\mathbf{p}_c^{(l)} \in \mathbb{R}^D\}_{c=1}^{N_{re}}$. Given an input token $\mathbf{x}_t^{(l)}$, it is ℓ_2 -normalized with expert prototypes by $\hat{\mathbf{x}}_t^{(l)} = \mathbf{x}_t^{(l)} / \|\mathbf{x}_t^{(l)}\|_2$ and $\hat{\mathbf{p}}_c^{(l)} = \mathbf{p}_c^{(l)} / \|\mathbf{p}_c^{(l)}\|_2$. Inspired by [57], the routing score is calculated using cosine similarity. Through Top-K affinity, each token is assigned to the experts with Top-K affinity:

$$(w_t, r_t) = \underset{1 \leq c \leq N_{re}}{\text{Top-K}} \langle \hat{\mathbf{x}}_t^{(l)}, \hat{\mathbf{p}}_c^{(l)} \rangle, \quad w_t \in \mathbb{R}^K, \quad r_t \in \{1, \dots, N_{re}\}^K. \quad (6)$$

where w_t denotes the routed expert weights of token $\hat{\mathbf{x}}_t^{(l)}$, r_t denotes the index set of routed experts. Let $r_{t,1}, \dots, r_{t,K}$ be the selected expert indices and $w_{t,1}, \dots, w_{t,K}$ the corresponding routing weights obtained from Equation (6), the output of MoE layer goes as:

$$\text{MoE}^{(l)}(\hat{\mathbf{x}}_t^{(l)}) = E_s^{(l)}(\hat{\mathbf{x}}_t^{(l)}) + \sum_{k=1}^K w_{t,k} \cdot E_{r_{t,k}}^{(l)}(\hat{\mathbf{x}}_t^{(l)}). \quad (7)$$

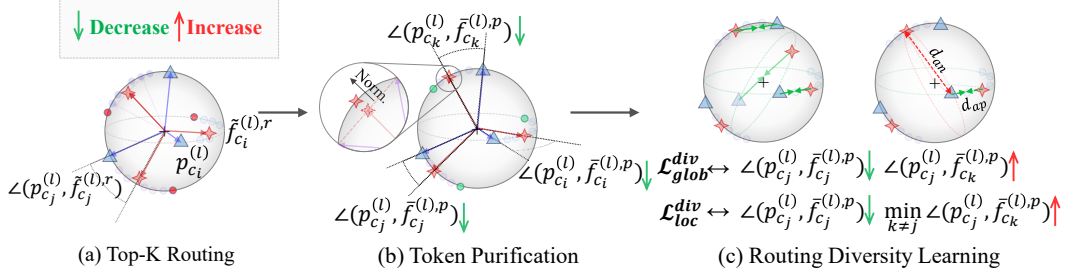


Figure 4: HypMoE-ReID on unit hypersphere $\mathbb{S}^{D-1}$. (a) Top-K Routing assigns tokens to the K nearest expert prototypes via cosine similarity. (b) Token Purification reweights tokens by their similarity to the assigned prototype. (c) Routing Diversity Learning: Global Separation Loss aligns prototypes to their own cluster centers and pushes others toward orthogonality. Local Diversification Loss maximizes the angular margin between the anchor-positive pair and the hardest negative center.

Through Equation (7), routing reduces to nearest-prototype search on unit hypersphere under cosine similarity, as illustrated in Figure 4(a). However, tokens assigned to each expert could contain noises which harm the quality of expert specialization. This motivates the Token Purification mechanism presented in next section.

4.2 Token Purification

Sparse routing enables each expert with a subset of tokens, yet the subset can be polluted by noisy contents, harmful to expert specialization. To mitigate this, we purify the token clusters before Routing Diversity Learning, as shown in Figure 3(b).

Given routed expert $E_c^{(l)}$ in layer l , let $\mathcal{T}_c = \{t \mid r_t = c\}$ be the set of token indices routed to it, where r_t is obtained from Equation (6). All tokens are ℓ_2 normalized, yielding $\hat{\mathbf{x}}_t^{(l)}$. The raw cluster center is defined as $\tilde{\mathbf{f}}_c^{(l),r} = 1/|\mathcal{T}_c| \cdot \sum_{t \in \mathcal{T}_c} \hat{\mathbf{x}}_t^{(l)}$. However, $\tilde{\mathbf{f}}_c^{(l),r}$ treats all routed tokens equally, which is sub-optimal when some routed tokens are weakly aligned with the assigned expert prototype. We instead reweight the tokens based on their similarity with each expert, based on a prior that noisy features are inherently less similar to the correct prototype. Specifically, we define the weights as:

$$a_{t,c}^{(l)} = [\langle \hat{\mathbf{x}}_t^{(l)}, \hat{\mathbf{p}}_c^{(l)} \rangle]_+, \quad [u]_+ = \max(u, 0), \quad w_{t,c}^{(l)} = \frac{a_{t,c}^{(l)}}{\sum_{u \in \mathcal{T}_c} a_{u,c}^{(l)}}. \quad (8)$$

The purified cluster is then obtained as $\tilde{\mathbf{f}}_c^{(l),p} = \sum_{t \in \mathcal{T}_c} w_{t,c}^{(l)} \cdot \hat{\mathbf{x}}_t^{(l)}$, $\tilde{\mathbf{f}}_c^{(l),p} = \tilde{\mathbf{f}}_c^{(l),p} / \|\tilde{\mathbf{f}}_c^{(l),p}\|_2$. Essentially, purification acts as a directional voting operator followed by projection back onto unit hypersphere, geometrically down-weighting tokens with large angular deviation from routed expert prototype, as illustrated in Figure 4(b).

Lemma 1 (Geometrical Gain of Token Purification). *Let $\{\hat{\mathbf{x}}_t\}_{t=1}^n \subset \mathbb{S}^{D-1}$ be a set of ℓ_2 -normalized token features assigned to the same prototype $\hat{\mathbf{p}} \in \mathbb{S}^{D-1}$. Define $c_t := \langle \hat{\mathbf{x}}_t, \hat{\mathbf{p}} \rangle$, $a_t = [c_t]_+$, and assume $\sum_{t=1}^n a_t > 0$. Let the raw mean be $\tilde{\mathbf{f}}^r = 1/n \sum_{t=1}^n \hat{\mathbf{x}}_t$, and the purified mean be $\tilde{\mathbf{f}}^p = \sum_{t=1}^n w_t \hat{\mathbf{x}}_t$, $w_t = a_t / \sum_{j=1}^n a_j$. Denote the normalized purified center by $\tilde{\mathbf{f}}^p = \tilde{\mathbf{f}}^p / \|\tilde{\mathbf{f}}^p\|_2$. Then purification improves the prototype-direction alignment in the sense that*

$$\langle \hat{\mathbf{p}}, \tilde{\mathbf{f}}^p \rangle \geq \langle \hat{\mathbf{p}}, \tilde{\mathbf{f}}^r \rangle \geq \langle \hat{\mathbf{p}}, \tilde{\mathbf{f}}^r \rangle. \quad (9)$$

Full proof is provided in Appendix A.1.

4.3 Routing Diversity Learning

After purification, each routed expert is associated with a purified cluster center $\tilde{\mathbf{f}}_c^{(l),p}$. Inspired by principles of metric learning [58; 59], we explicitly regularize the prototype-cluster geometry by diversity-oriented supervision, consisting of a Global Separation Loss and a Local Diversification

Table 1: Comparison Results on Large Scale Person ReID.

Model	Arch.	Pub.	# Config.	# FLOPs.	# Params.(Act./Act. Exp.)	APE.		Average Acc.	
						mAP	R@1	mAP	R@1
ViT [14]	DENSE	<i>ICLR 2021</i>	-	11.35 G	100.50 M / -	64.13	75.28	64.45	75.66
Qwen 2.5 [37]	SPARSE	<i>arXiv 2025</i>	R16, S1, K2	10.45 G	93.48 M / 21.24 M	58.57	72.07	54.75	67.37
Mixtral 8x7B [38]	SPARSE	<i>arXiv 2024</i>	R8, S0, K2	18.66 G	157.14 M / 84.93 M	31.31	39.21	49.20	61.61
DBRX [41]	SPARSE	<i>Springer 2024</i>	R12, S0, K4	29.62 G	242.09 M / 169.87 M	12.14	17.79	29.40	43.06
Grok 2.5 [39]	SPARSE	<i>IA 2025</i>	R12, S0, K2	20.48 G	171.29 M / 84.93 M	16.47	24.00	28.21	41.11
LLaMA-MoE [40]	SPARSE	<i>EMNLP 2024</i>	R8, S0, K2	11.81 G	104.10 M / 10.62 M	51.01	62.61	53.10	65.18
Yuan 2.0-M32 [42]	SPARSE	<i>arXiv 2024</i>	R8, S0, K2	18.67 G	157.22 M / 84.93 M	30.29	38.60	47.62	60.68
DeepSeek-R1 [36]	SPARSE	<i>Nature 2025</i>	R12, S1, K2	37.31 G	157.20 M / 85.00 M	35.17	42.81	55.29	67.30
HypMoE-ReID	SPARSE	<i>This Paper</i>	R16, S1, K1	11.36 G	100.58 M / 28.34 M	66.21	77.17	66.60	77.61
						+2.08↑	+1.89↑	+2.15↑	+1.95↑
MSP-ReID [30]	DENSE	<i>ICASSP 2026</i>	-	11.35 G	85.73 M / -	36.35	53.51	31.16	45.87
+HypMoE-ReID	SPARSE	<i>This Paper</i>	R16, S1, K1	11.36 G	85.83 M / 28.34 M	41.81	59.23	35.89	50.84
						+5.46↑	+5.72↑	+4.73↑	+4.97↑
PromptSG [60]	DENSE	<i>CVPR 2024</i>	-	14.08 G	109.92 M / -	57.41	68.39	63.11	75.17
+HypMoE-ReID	SPARSE	<i>This Paper</i>	R16, S1, K1	14.09 G	110.00 M / 28.34 M	59.15	69.96	65.07	76.96
						+1.74↑	+1.57↑	+1.96↑	+1.79↑
CLIP-ReID [11]	DENSE	<i>AAAI 2023</i>	-	11.10 G	109.45 M / -	67.63	76.12	74.02	83.31
+HypMoE-ReID	SPARSE	<i>This Paper</i>	R16, S1, K1	11.11 G	109.53 M / 28.34 M	69.65	77.16	76.28	84.52
						+2.02↑	+1.04↑	+2.26↑	+1.21↑

R = number of routed experts, S = number of shared experts, K = Top-K routing.

Loss. For the l -th MoE block, let the valid expert set be $\mathcal{V}^{(l)} = \{c \in \{1, \dots, N_{re}\} \mid |\mathcal{T}_c^{(l)}| > 0\}$, where $\mathcal{T}_c^{(l)}$ denotes the index set of tokens routed to expert $E_c^{(l)}$. For each valid expert, we compare $\hat{\mathbf{p}}_c^{(l)}$ with all purified cluster centers $\{\bar{\mathbf{f}}_k^{(l),p}\}_{k \in \mathcal{V}^{(l)}}$.

Global Separation Loss. The global separation loss is formulated as cross-entropy over cosine similarities:

$$\mathcal{L}_{glob}^{(l),div} = -\frac{1}{|\mathcal{V}^{(l)}|} \sum_{c \in \mathcal{V}^{(l)}} \log \frac{\exp(\langle \hat{\mathbf{p}}_c^{(l)}, \bar{\mathbf{f}}_c^{(l),p} \rangle / \tau)}{\sum_{k \in \mathcal{V}^{(l)}} \exp(\langle \hat{\mathbf{p}}_c^{(l)}, \bar{\mathbf{f}}_k^{(l),p} \rangle / \tau)}, \quad (10)$$

where τ is temperature. The geometric effect is formalized in Lemma 2.

Lemma 2 (Angular Meaning of Global Separation Loss). Define $\theta_{c,k} := \angle(\hat{\mathbf{p}}_c, \bar{\mathbf{f}}_k)$. For any valid expert $c \in \mathcal{V}^{(l)}$, the gradient of $\mathcal{L}_{glob}^{div}$ w.r.t. the listed angles satisfies:

$$\forall k \neq c, \quad \frac{\partial \mathcal{L}_{glob}^{div}}{\partial \theta_{c,c}} > 0 \wedge \frac{\partial \mathcal{L}_{glob}^{div}}{\partial \theta_{c,k}} < 0, \quad (11)$$

so gradient descent persistently decreases $\theta_{c,c}$ (prototype-cluster alignment) while increasing $\theta_{c,k}$ (inter-expert separation). Full proof is provided in Appendix A.2.

Local Diversification Loss. Inspired by triplet-based metric learning [58], we further introduce a fine-grained diversification signal. Define anchor-positive distance for expert $E_c^{(l)}$ as $d_{ap,c}^{(l)} = \|\hat{\mathbf{p}}_c^{(l)} - \bar{\mathbf{f}}_c^{(l),p}\|_2$, and the hardest anchor-negative distance as $d_{an,c}^{(l)} = \min_{k \in \mathcal{V}^{(l)}, k \neq c} \|\hat{\mathbf{p}}_c^{(l)} - \bar{\mathbf{f}}_k^{(l),p}\|_2$. The local diversity loss is defined as a soft-margin triplet loss:

$$\mathcal{L}_{loc}^{(l),div} = \frac{1}{|\mathcal{V}^{(l)}|} \sum_{c \in \mathcal{V}^{(l)}} \log \left(1 + \exp(d_{ap,c}^{(l)} - d_{an,c}^{(l)}) \right). \quad (12)$$

Equation (12) explicitly sharpens the boundary between expert prototype and its hardest negative center, complementing the global functions in Equation (10). Its geometrical essence is revealed in Lemma 3, where the effects of different metrics are unified under a same framework.

Lemma 3 (Hyperspherical Equivalence). For any $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{D-1}$,

$$\|\mathbf{u} - \mathbf{v}\|_2^2 = 2 - 2\langle \mathbf{u}, \mathbf{v} \rangle. \quad (13)$$

Consequently, for a fixed anchor $\mathbf{u} \in \mathbb{S}^{D-1}$, Euclidean distance, cosine similarity, and angular distance induce identical rankings over candidates on the unit hypersphere. Full proof is provided in Appendix A.3.

Training and Inference. As shown in Figure 3, HypMoE-ReID is trained from end-to-end. Following previous works [10; 11], the model employs ReID loss [59; 58] and Equation (10)(12) for

optimization:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \alpha \times \mathcal{L}_{tri} + \sum_{l \in \mathcal{M}} (\mathcal{L}_{glob}^{(l),div} + \beta \times \mathcal{L}_{loc}^{(l),div}) \quad (14)$$

where α and β are hyper-parameters controlling the loss strengths. [CLS] token extracted from the last block is used for evaluation.

5 Experiments

5.1 Experimental Setup

Datasets and Metrics. The proposed benchmark combines 13 public datasets: Market-1501 [6], MSMT17 [7], CUHK-SYSU [61], CUHK03 [28], IUST-PersonReID [32], LaST [62], MARS [63], RPIField [64], LPW [65], iLIDS-VID [66], PRID2011 [67], DeepSportRadar [68], and PRW [69]. Video datasets are converted to image ones following [70] (see Appendix C.1). We report mAP, R@1, and their average. Overhead metrics include Active Parameter Efficiency (APE.), per-token FLOPs, Active Parameters (Act. Params.), and Active Expert Parameters (Act. Exp. Params.). Details are provided in Appendix C.3.

Compared Methods. We compare our method against (1) Raw architectures, including dense ViT [14] and sparse MoEs (DeepSeek-R1 [36], Qwen [37], Mixtral [38], etc.). Details of the sparse architectures are provided in Appendix C.2 (2) ReID methods, including MSP-ReID [30], CLIP-ReID [11], PromptSG [60]. For ReID models, we replace their backbones with our HypMoE-ReID backbone and compare the performance between *with* and *without* our framework. Results on large scale vehicle ReID are also reported in Appendix E, where we also build combined vehicle datasets to validate the generalizability of our methods.

Implementation Details. HypMoE-ReID is upcycled from ViT-B/16 [71] with 16 routed experts and 1 shared expert per sparse layer. Expert hidden dimension is set to 1536, and Top-1 routing is adopted. We apply common data augmentations, including random flip, erase, crop. We train the model for 60 epochs using the AdamW optimizer. The temperature in Global Separation Loss is set to 0.07, and the loss weights α and β are set to 1.0 and 0.5, respectively. All experiments are conducted on one NVIDIA RTX 4090 GPU.

5.2 Comparison Results

In this section, we compare the proposed HypMoE-ReID with state-of-the-art dense model-based ReID methods and MoE-based approaches on the proposed large-scale ReID benchmark. Detailed comparisons of the configuration setting (# Config.), FLOPs, number of parameters (# Params.), APE, and Average Acc are reported in Tab. 1.

Against Sparse MoEs. As shown in Table 1, our method substantially outperforms all sparse architectures. Compared to the strongest sparse competitor Qwen 2.5 [37], our method outperforms it by **7.64%/5.10%** mAP/R@1 on APE, and by **11.85%/10.24%** mAP/R@1 on the average accuracy, despite sharing the same expert configuration. These results highlight the importance of top-1 routing over top-2 routing in ReID, where activating a single specialized expert per token avoids the interference between heterogeneous expert outputs and encourages stronger specialization. Architectures with heavier routing budgets, such as Mixtral [38] and DeepSeek-R1 [36], all perform significantly worse despite substantially larger active parameters and FLOPs, indicating that directly transplanting LLM-style MoE architecture into vision backbones is suboptimal. Models without a shared expert consistently underperform those with one, validating the role of shared experts in preserving generic representations alongside routed specialization.

Against Dense SOTAs. Our sparse model consistently outperforms its dense counterparts. Compared to the ViT [14], our method yields gains of **2.08%/1.89%** mAP/R@1 on the APE and **2.15%/1.95%** mAP/R@1 on average accuracy, becoming the only sparse architecture to outperform dense baseline on the benchmark. We further validate the effectiveness by replacing the backbones of three established ReID frameworks with our HypMoE-ReID. CLIP-ReID+HypMoE-ReID achieves the overall best accuracy of **76.28%/84.52%** on Average Accuracies, setting a new state-of-the-art on the large-scale benchmark.

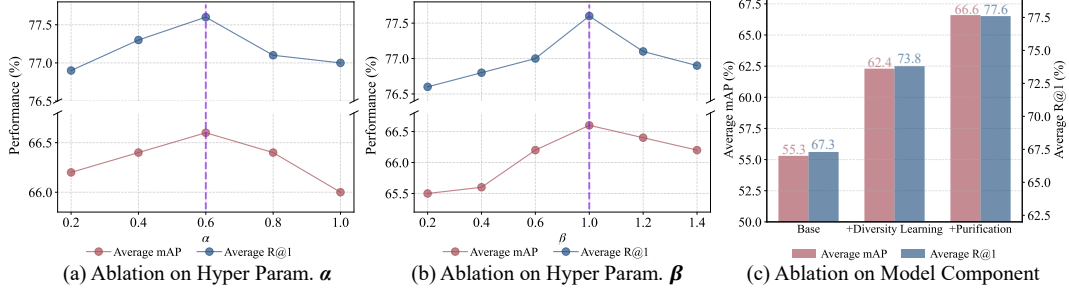


Figure 5: Ablations on hyper-parameters and model components.

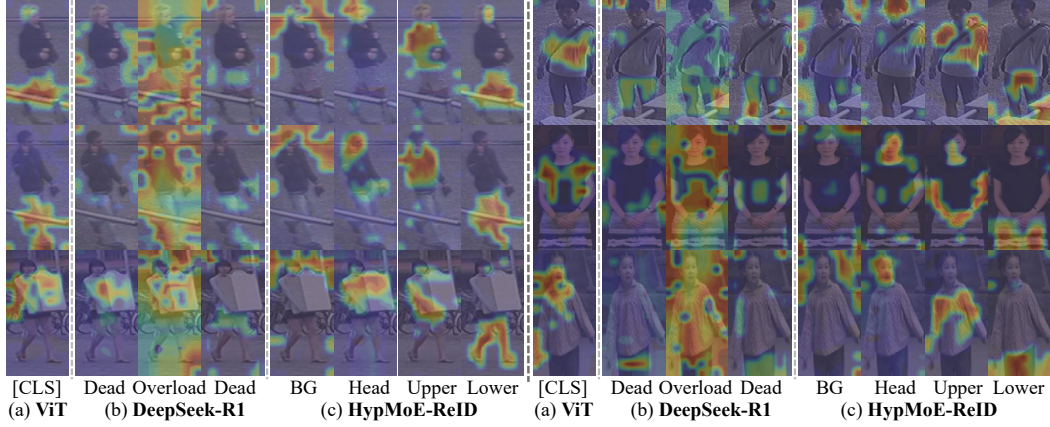


Figure 6: Attention visualizations on dense ViT, sparse DeepSeek-R1, and HypMoE-ReID.

Overhead Analysis. Our method introduces negligible overhead compared to the dense baseline. Notably, FLOPs are nearly identical since top-1 routing in HypMoE-ReID ensures that only one expert is executed per token per MoE layer. Additionally, the active parameters of HypMoE-ReID increase by only 0.08M compared to ViT due to the additional routing weights, while significantly increasing APE compared to all sparse methods. In contrast, architectures such as DeepSeek-R1 activate far more parameters while delivering lower accuracy. These results collectively demonstrate that the proposed Routing Diversity Learning mechanism achieves the optimal trade-off between model capacity and efficiency for large scale ReID.

5.3 Ablation Studies and Additional Analysis

Ablations on Hyper-parameters. Figure 5(a)(b) analyzes the effects of α and β . It is found that small β provides insufficient diversity pressure, leaving experts partially collapsed, while too large β over-regularizes the routing and prevents experts from specializing on discriminative regions. Similarly, a large α over-suppress discriminative tokens and degrades performance.

Ablations on Model Components. Using DeepSeek-R1 as the base architecture, Figure 5(c) shows that incorporating the Routing Diversity Learning mechanism yields substantial gains of **7.1%/6.5%** in mAP/R@1, indicating it effectively prevents routing collapse on the extreme homogeneous ReID data. Adding Token Purification module further contributes **4.2%/3.8%** mAP/R@1, confirming that suppressing noisy tokens is beneficial for expert specialization.

Visualization. As shown in Figure 6, for dense ViT, shared weights across all tokens make it harder for attention to capture discriminative fine-grained features (e.g., heads). For DeepSeek-R1, the homogeneous ReID data induces severe routing collapse, where most tokens are routed to the same expert. Although GShard [53] loss penalizes such imbalance, the forced redistribution compels overloaded experts to process tokens they are not specialized for, and ultimately under-perform the dense ViT.

In contrast, our method exhibits semantically meaningful expert specialization, including Background Experts (BG for short), Head Experts, etc. This demonstrates that the proposed diversity learning mechanism successfully prevents routing collapse and drives the MoE backbone toward a human-interpretable, part-aware representation. A similar phenomenon is also observed on vehicle datasets, as shown in the attention visualizations for different architectures in Figure 9.

6 Conclusion

In this study, we identify the fundamental limitations of existing widely used dense model-based ReID methods and MoE architectures in large-scale ReID scenarios, which exhibit limited capacity for learning diverse representations and suffer from routing collapse, respectively. To address these issues, we propose a novel framework, **HypMoE-ReID**, which incorporates a Token Purification mechanism to enhance expert specialization and a Routing Diversity Learning scheme to mitigate routing collapse based on theoretical analysis. Furthermore, we construct a large-scale benchmark by unifying 13 public datasets, reflecting realistic distribution shifts and domain gaps. Extensive experiments demonstrate that our approach achieves state-of-the-art performance in large-scale ReID, consistently outperforming both dense model-based methods and existing MoE architectures.

Limitation Discussion. HypMoE-ReID adds a small router-induced increase in active parameters. By appropriately tuning expert hidden dimensions, active parameters can remain close to dense counterparts, substantially alleviating representation rigidity with no extra per-token cost.

Acknowledgement

This work was a research achievement of National Engineering Research Center of New Electronic Publishing Technologies, and was supported by the National Natural Science Foundation of China (62376011) and the National Key R&D Program of China (2024YFA1410000).

References

- [1] Y. Gong, L. Huang, and L. Chen, “Person re-identification method based on color attack and joint defence,” in *CVPR*, 2022, pp. 4313–4322.
- [2] Z. Cui, J. Zhou, and Y. Peng, “Dma: Dual modality-aware alignment for visible-infrared person re-identification,” *IEEE TIFS*, vol. 19, pp. 2696–2708, 2024.
- [3] J. Shi, Y. Zhang, X. Yin, Y. Xie, Z. Zhang, J. Fan, Z. Shi, and Y. Qu, “Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification,” in *ICCV*. IEEE, 2023, pp. 11 218–11 228.
- [4] J. Zhou, K. Xu, F. Zhuo, X. Zou, and Y. Peng, “Distribution-aware knowledge aligning and prototyping for non-exemplar lifelong person re-identification,” *IEEE TPAMI*, 2025.
- [5] K. Xu, Z. Liu, X. Zou, Y. Peng, and J. Zhou, “Long short-term knowledge decomposition and consolidation for lifelong person re-identification,” *IEEE TPAMI*, vol. 47, no. 9, pp. 7796–7811, 2025.
- [6] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.
- [7] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer gan to bridge domain gap for person re-identification,” in *CVPR*, 2018, pp. 79–88.
- [8] K. Xu, X. Zou, and J. Zhou, “Lstkc: Long short-term knowledge consolidation for lifelong person re-identification,” in *AAAI*, vol. 38, no. 14, 2024, pp. 16 202–16 210.
- [9] K. Xu, X. Zou, Y. Peng, and J. Zhou, “Distribution-aware knowledge prototyping for non-exemplar lifelong person re-identification,” in *CVPR*. IEEE, 2024, pp. 16 604–16 613.

- [10] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, “Transreid: Transformer-based object re-identification,” in *ICCV*, 2021, pp. 15 013–15 022.
- [11] S. Li, L. Sun, and Q. Li, “Clip-reid: exploiting vision-language model for image re-identification without concrete text labels,” in *AAAI*, vol. 37, no. 1, 2023, pp. 1405–1413.
- [12] K. Xu, J. Li, Y. Guo, Y. Peng, and J. Zhou, “Mitigate catastrophic remembering via continual self-paced dual-knowledge purification for noisy lifelong person re-identification,” *IEEE TPAMI*, 2026.
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *NeurIPS*, vol. 30, 2017.
- [14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [15] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive mixtures of local experts,” *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.
- [16] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby, “Scaling vision with sparse mixture of experts,” *NeurIPS*, vol. 34, pp. 8583–8595, 2021.
- [17] R. Collobert, S. Bengio, and Y. Bengio, “A parallel mixture of svms for very large scale problems,” *NeurIPS*, vol. 14, 2001.
- [18] K. Xu, Y. Zhang, W. Qiang, and J. Zhou, “Hyper-llava: Hyperbolic uncertainty-aware modality-balanced routing for multimodal continual instruction tuning,” in *ICML*, 2026.
- [19] Z. Liu, K. Xu, B. Su, X. Zou, Y. Peng, and J. Zhou, “Stop: Integrated spatial-temporal dynamic prompting for video understanding,” in *CVPR*. IEEE, 2025, pp. 13 776–13 786.
- [20] B. Cao, Y. Sun, P. Zhu, and Q. Hu, “Multi-modal gated mixture of local-to-global experts for dynamic image fusion,” in *ICCV*, 2023, pp. 23 555–23 564.
- [21] Y. Li, S. Jiang, B. Hu, L. Wang, W. Zhong, W. Luo, L. Ma, and M. Zhang, “Uni-moe: Scaling unified multimodal llms with mixture of experts,” *IEEE TPAMI*, 2025.
- [22] M. Videau, A. Leite, M. Schoenauer, and O. Teytaud, “Mixture of experts in image classification: What’s the sweet spot?” *TMLR*, 2025.
- [23] X. Han, L. Wei, Z. Dou, Y. Sun, Z. Han, and Q. Tian, “Vimoe: An empirical study of designing vision mixture-of-experts,” *IEEE Transactions on Image Processing*, vol. 34, pp. 7209–7221, 2025.
- [24] Z. Sun, L. Tan, Y. Shen, C. Cai, X. Sun, P. Dai, L. Cao, and R. Ji, “Flexireid: Adaptive mixture of expert for multi-modal person re-identification,” in *ICML*, 2025.
- [25] Y. Dai, X. Li, J. Liu, Z. Tong, and L.-Y. Duan, “Generalizable person re-identification with relevance-aware mixture of experts,” in *CVPR*, 2021, pp. 16 145–16 154.
- [26] K. Ren, C. Hu, H. Xi, Y. Li, J. Fan, and L. Liu, “Mosce-reid: Mixture of semantic clustering experts for person re-identification,” *Neurocomputing*, vol. 626, p. 129587, 2025.
- [27] Y. Wang, Y. Zhang, X. Ren, and Y. Deng, “Moda: Mixture of domain adapters for parameter-efficient generalizable person re-identification,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 5, pp. 1–19, 2025.
- [28] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *CVPR*, 2014, pp. 152–159.
- [29] Q. Li, K. Xu, Y. Peng, and J. Zhou, “Exemplar-free lifelong person re-identification via prompt-guided adaptive knowledge consolidation,” *IJCV*, vol. 132, no. 11, pp. 4850–4865, 2024.

- [30] X. He and L. Wan, “Msp-reid: Hairstyle-robust cloth-changing person re-identification,” in *ICASSP*. IEEE, 2026, pp. 9807–9811.
- [31] H. Zheng, X. Zhong, W. Huang, K. Jiang, W. Liu, and Z. Wang, “Visible-infrared person re-identification: A comprehensive survey and a new setting,” *Electronics*, vol. 11, no. 3, p. 454, 2022.
- [32] M. R. Mohammadi, “Iust_personreid: A new domain in person re-identification datasets,” *Available at SSRN 5107561*.
- [33] S. Yıldız and A. N. Kasım, “Entire-id: An extensive and diverse dataset for person re-identification,” in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2024, pp. 1–5.
- [34] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” in *ICLR*, 2017.
- [35] S. Mu and S. Lin, “A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications,” *arXiv preprint arXiv:2503.07137*, 2025.
- [36] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi *et al.*, “Deepseek-r1 incentivizes reasoning in llms through reinforcement learning,” *Nature*, vol. 645, no. 8081, pp. 633–638, 2025.
- [37] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu *et al.*, “Qwen2.5-coder technical report,” *arXiv preprint arXiv:2409.12186*, 2024.
- [38] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. I. Casas, E. B. Hanna, F. Bressand *et al.*, “Mixtral of experts,” *arXiv preprint arXiv:2401.04088*, 2024.
- [39] M. E. de Carvalho Souza and L. Weigang, “Grok, gemini, chatgpt and deepseek: Comparison and applications in conversational artificial intelligence,” *Inteligencia Artificial*, vol. 2, no. 1, 2025.
- [40] T. Zhu, X. Qu, D. Dong, J. Ruan, J. Tong, C. He, and Y. Cheng, “Llama-moe: Building mixture-of-experts from llama with continual pre-training,” in *EMNLP*, 2024, pp. 15 913–15 923.
- [41] N. Gupta and J. Yip, “Dbrx: Creating an llm from scratch using databricks,” in *Databricks Data Intelligence Platform: Unlocking the GenAI Revolution*. Springer, 2024, pp. 311–330.
- [42] S. Wu, J. Luo, X. Chen, L. Li, X. Zhao, T. Yu, C. Wang, Y. Wang, F. Wang, W. Qiao *et al.*, “Yuan 2.0-m32: Mixture of experts with attention router,” *arXiv preprint arXiv:2405.17976*, 2024.
- [43] Z. Fan, R. Sarkar, Z. Jiang, T. Chen, K. Zou, Y. Cheng, C. Hao, Z. Wang *et al.*, “M³vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design,” *NeurIPS*, vol. 35, pp. 28 441–28 457, 2022.
- [44] Z. Fei, M. Fan, C. Yu, D. Li, and J. Huang, “Scaling diffusion transformers to 16 billion parameters,” *arXiv preprint arXiv:2407.11633*, 2024.
- [45] S. Yang, J. Wen, and B. Fang, “Fg-moe: Heterogeneous mixture of experts model for fine-grained visual classification,” *Pattern Recognition*, p. 113050, 2026.
- [46] L. Zhang, S. Huang, and W. Liu, “Enhancing mixture-of-experts by leveraging attention for fine-grained recognition,” *TMM*, vol. 24, pp. 4409–4421, 2021.
- [47] R. Ha, S. Jiang, B. Li, B. Pan, Y. Zhu, J. Zhang, X. Zhu, S. Gong, and J. Wang, “Multi-modal multi-platform person re-identification: Benchmark and method,” in *ICCV*, 2025, pp. 10 251–10 261.
- [48] S. Li, C. Li, A. Zheng, A. Lu, J. Tang, and J. Ma, “Next: Multi-grained mixture of experts via text-modulation for multi-modal object re-id,” *arXiv e-prints*, pp. arXiv–2505, 2025.

- [49] B. Xu, J. Liang, L. He, and Z. Sun, “Mimic embedding via adaptive aggregation: Learning generalizable person re-identification,” in *ECCV*. Springer, 2022, pp. 372–388.
- [50] Y. Su, Y. Shi, F. Liu, and X. Liu, “Hamobe: Hierarchical and adaptive mixture of biometric experts for video-based person reid,” in *ICCV*, 2025, pp. 11 525–11 536.
- [51] Y. Wang, Y. Liu, A. Zheng, and P. Zhang, “Decoupled feature-based mixture of experts for multi-modal object re-identification,” in *AAAI*, vol. 39, no. 8, 2025, pp. 8141–8149.
- [52] Y. Feng, J. Li, C. Xie, L. Tan, and J. Ji, “Multi-modal object re-identification via sparse mixture-of-experts,” in *ICML*, 2025.
- [53] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, “Gshard: Scaling giant models with conditional computation and automatic sharding,” in *ICLR*, 2020.
- [54] W. Fedus, B. Zoph, and N. Shazeer, “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity,” *JMLR*, vol. 23, no. 120, pp. 1–39, 2022.
- [55] T. Li and K. He, “Back to basics: Let denoising generative models denoise,” in *CVPR*, 2026, pp. 36 115–36 125.
- [56] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *ICML*. PMLR, 2021, pp. 10 347–10 357.
- [57] Y. Wei, S. Zhang, H. Yuan, Y. Han, Z. Chen, J. Wang, D. Zou, X. Liu, Y. Liu, H. Shan *et al.*, “Routing matters in moe: Scaling diffusion transformers with explicit routing guidance,” in *ICLR*, vol. 2026, 2026, pp. 23 381–23 405.
- [58] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *CVPR*, 2015, pp. 815–823.
- [59] D. R. Cox, “The regression analysis of binary sequences,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215–232, 1958.
- [60] Z. Yang, D. Wu, C. Wu, Z. Lin, J. Gu, and W. Wang, “A pedestrian is worth one prompt: Towards language guidance person re-identification,” in *CVPR*, 2024, pp. 17 343–17 353.
- [61] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, “End-to-end deep learning for person search,” *arXiv preprint arXiv:1604.01850*, vol. 2, no. 2, p. 4, 2016.
- [62] X. Shu, X. Wang, X. Zang, S. Zhang, Y. Chen, G. Li, and Q. Tian, “Large-scale spatio-temporal person re-identification: Algorithms and benchmark,” *CSVT*, vol. 32, no. 7, pp. 4390–4403, 2021.
- [63] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, “Mars: A video benchmark for large-scale person re-identification,” in *ECCV*. Springer, 2016, pp. 868–884.
- [64] M. Zheng, S. Karanam, and R. J. Radke, “Rpipfield: A new dataset for temporally evaluating person re-identification,” in *CVPRW*, 2018, pp. 1893–1895.
- [65] K. Xu, F. Zhuo, J. Li, X. Zou, and J. Zhou, “Self-reinforcing prototype evolution with dual-knowledge cooperation for semi-supervised lifelong person re-identification,” in *ICCV*, 2025, pp. 3564–3574.
- [66] T. Wang, S. Gong, X. Zhu, and S. Wang, “Person re-identification by discriminative selection in video ranking,” *IEEE TPAMI*, vol. 38, no. 12, pp. 2501–2514, 2016.
- [67] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, “Person re-identification by descriptive and discriminative classification,” in *Scandinavian conference on Image analysis*. Springer, 2011, pp. 91–102.

- [68] G. Van Zandycke, V. Somers, M. Istasse, C. D. Don, and D. Zambrano, “Deepsporadar-v1: Computer vision dataset for sports understanding with high quality annotations,” in *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, 2022, pp. 1–8.
- [69] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, “Person re-identification in the wild,” in *CVPR*, 2017, pp. 1367–1376.
- [70] K. Xu, C. Jiang, P. Xiong, Y. Peng, and J. Zhou, “Dask: Distribution rehearsing via adaptive style kernel learning for exemplar-free lifelong person re-identification,” in *AAAI*, vol. 39, no. 9, 2025, pp. 8915–8923.
- [71] A. Komatsuzaki, J. Puigcerver, J. Lee-Thorp, C. R. Ruiz, B. Mustafa, J. Ainslie, Y. Tay, M. Dehghani, and N. Houlsby, “Sparse upcycling: Training mixture-of-experts from dense checkpoints,” in *ICLR*, 2022.
- [72] N. Shazeer, “Glu variants improve transformer,” *arXiv preprint arXiv:2002.05202*, 2020.
- [73] X. Liu, W. Liu, H. Ma, and H. Fu, “Large-scale vehicle re-identification in urban surveillance videos,” in *ICME*. IEEE, 2016, pp. 1–6.
- [74] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, “Veri-wild: A large dataset and a new method for vehicle re-identification in the wild,” in *CVPR*, 2019, pp. 3235–3243.
- [75] P. Wang, B. Jiao, L. Yang, Y. Yang, S. Zhang, W. Wei, and Y. Zhang, “Vehicle re-identification in aerial imagery: Dataset and approach,” in *ICCV*, 2019, pp. 460–469.
- [76] A. Kanacı, X. Zhu, and S. Gong, “Vehicle re-identification in context,” in *German Conference on Pattern Recognition*. Springer, 2018, pp. 377–390.
- [77] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, “Deep relative distance learning: Tell the difference between similar vehicles,” in *CVPR*, 2016, pp. 2167–2175.

A Proof

A.1 Proof of Lemma 1–Geometrical Gain of Token Purification

Variable definitions follow Lemma 1 in Section 4.2, where $c_t := \langle \hat{\mathbf{x}}_t, \hat{\mathbf{p}} \rangle$, $a_t = [c_t]_+$. We introduce the following intermediate features:

$$\tilde{\mathbf{f}}^r = \frac{1}{n} \sum_{t=1}^n \hat{\mathbf{x}}_t, \quad \tilde{\mathbf{f}}^p = \sum_{t=1}^n w_t \cdot \hat{\mathbf{x}}_t, \quad \bar{\mathbf{f}}^p = \frac{\tilde{\mathbf{f}}^p}{\|\tilde{\mathbf{f}}^p\|_2} \quad (15)$$

where $\bar{\mathbf{f}}^p \in \mathbb{S}^{D-1}$, n is the number of tokens within current cluster, and $\bar{\mathbf{f}}^p$ stands for the purified token cluster center located on hypersphere. By the linear properties of inner product where

$$\langle \alpha \mathbf{v} + \beta \mathbf{w}, \mathbf{u} \rangle = \alpha \langle \mathbf{v}, \mathbf{u} \rangle + \beta \langle \mathbf{w}, \mathbf{u} \rangle \quad (16)$$

Consequently we have,

$$\langle \hat{\mathbf{p}}, \tilde{\mathbf{f}}^p \rangle = \langle \hat{\mathbf{p}}, \sum_{t=1}^n w_t \cdot \hat{\mathbf{x}}_t \rangle = \sum_{t=1}^n w_t \cdot \langle \hat{\mathbf{p}}, \hat{\mathbf{x}}_t \rangle = \sum_{t=1}^n w_t \cdot c_t \quad (17)$$

By applying Equation (8),

$$\langle \hat{\mathbf{p}}, \tilde{\mathbf{f}}^p \rangle = \frac{\sum_{t=1}^n a_t \cdot c_t}{\sum_{t=1}^n a_t} = \frac{\sum_{t \in \mathcal{P}} c_t^2}{\sum_{t \in \mathcal{P}} c_t} \quad (18)$$

where $\mathcal{P} := \{t | c_t > 0\}$, $m := |\mathcal{P}|$. According to Cauchy-Schwarz inequality,

$$m \cdot \sum_{t \in \mathcal{P}} c_t^2 \geq \left(\sum_{t \in \mathcal{P}} c_t \right)^2 \quad (19)$$

Given $n \geq m$, we have:

$$\frac{\sum_{t \in \mathcal{P}} c_t^2}{\sum_{t \in \mathcal{P}} c_t} \geq \frac{1}{m} \cdot \sum_{t \in \mathcal{P}} c_t \geq \frac{1}{n} \cdot \sum_{t \in \mathcal{P}} c_t \quad (20)$$

Define $\bar{\mathcal{P}} = \{t | c_t \leq 0\}$, we have:

$$\sum_{t=1}^n c_t = \sum_{t \in \mathcal{P}} c_t + \sum_{t \in \bar{\mathcal{P}}} c_t \leq \sum_{t \in \mathcal{P}} c_t \quad (21)$$

multiply Equation (21) by $1/n$:

$$\frac{1}{n} \sum_{t \in \mathcal{P}} c_t \geq \frac{1}{n} \sum_{t=1}^n c_t \quad (22)$$

By incorporating Equation (18) (21) (22), we have:

$$\langle \hat{\mathbf{p}}, \tilde{\mathbf{f}}^p \rangle = \frac{\sum_{t \in \mathcal{P}} c_t^2}{\sum_{t \in \mathcal{P}} c_t} \geq \frac{1}{n} \sum_{t \in \mathcal{P}} c_t \geq \frac{1}{n} \sum_{t=1}^n c_t \quad (23)$$

Note that $c_t = \langle \hat{\mathbf{p}}, \hat{\mathbf{x}}_t \rangle$. By using Equation (16), we have:

$$\langle \hat{\mathbf{p}}, \tilde{\mathbf{f}}^p \rangle \geq \frac{1}{n} \sum_{t=1}^n \langle \hat{\mathbf{p}}, \hat{\mathbf{x}}_t \rangle = \langle \hat{\mathbf{p}}, \frac{1}{n} \sum_{t=1}^n \hat{\mathbf{x}}_t \rangle = \langle \hat{\mathbf{p}}, \tilde{\mathbf{f}}^r \rangle \quad (24)$$

It is obvious that

$$\langle \hat{\mathbf{p}}, \bar{\mathbf{f}}^p \rangle = \frac{\langle \hat{\mathbf{p}}, \tilde{\mathbf{f}}^p \rangle}{\|\tilde{\mathbf{f}}^p\|_2} \quad (25)$$

Since $\mathcal{P} \neq \emptyset$ (guaranteed by $\sum_t a_t > 0$), Equation (18) shows $\langle \hat{\mathbf{p}}, \tilde{\mathbf{f}}^p \rangle = \sum_{t \in \mathcal{P}} c_t^2 / \sum_{t \in \mathcal{P}} c_t > 0$. From Triangle Inequality and Equation (8):

$$\|\tilde{\mathbf{f}}^p\|_2 = \left\| \sum_{t=1}^n w_t \cdot \hat{\mathbf{x}}_t \right\|_2 \leq \sum_{t=1}^n w_t \cdot \|\hat{\mathbf{x}}_t\|_2 = 1 \quad (26)$$

Consequently,

$$\langle \hat{\mathbf{p}}, \bar{\mathbf{f}}^p \rangle \geq \langle \hat{\mathbf{p}}, \tilde{\mathbf{f}}^p \rangle \geq \langle \hat{\mathbf{p}}, \tilde{\mathbf{f}}^r \rangle \quad (27)$$

□

A.2 Proof of Lemma 2–Angular Meaning of Global Separation Loss

Variable definitions follow Lemma 2 in Section 4.3 Apply identity transformation to Equation (10):

$$\mathcal{L}_{glob}^{div} = \frac{1}{|\mathcal{V}|} \sum_{c \in \mathcal{V}} \log \left(1 + \sum_{k \neq c} \exp \left(\frac{s_{c,k} - s_{c,c}}{\tau} \right) \right) \quad (28)$$

where τ is the temperature and $s_{c,k} := \langle \hat{\mathbf{p}}_c, \bar{\mathbf{f}}_k^p \rangle$. Define $\Delta_{c,k} = s_{c,k} - s_{c,c}$, so that:

$$\mathcal{L}_{glob}^{div} = \frac{1}{|\mathcal{V}|} \sum_{c \in \mathcal{V}} \log \left(1 + \sum_{k \neq c} \exp \left(\frac{\Delta_{c,k}}{\tau} \right) \right) \quad (29)$$

Fix an anchor c and denote the per-anchor term ℓ_c :

$$\ell_c = \log \left(1 + \sum_{k \neq c} \exp \left(\frac{\Delta_{c,k}}{\tau} \right) \right) \quad (30)$$

Take the partial derivate to $\Delta_{c,k}$:

$$\frac{\partial \ell_c}{\partial \Delta_{c,k}} = \frac{1}{\tau} \cdot \frac{\exp(\frac{\Delta_{c,k}}{\tau})}{1 + \sum_{j \neq c} \exp(\frac{\Delta_{c,j}}{\tau})} \quad (31)$$

Since $\partial \ell_c / \partial \Delta_{c,k} > 0$, gradient descent persistently drives every $\Delta_{c,k}$ downward. By definition of $s_{c,k}$ and the relation between cosine similarity and angle:

$$\Delta_{c,k} = \cos(\angle(\hat{\mathbf{p}}_c, \bar{\mathbf{f}}_k^p)) - \cos(\angle(\hat{\mathbf{p}}_c, \bar{\mathbf{f}}_c^p)) = \cos \theta_{c,k} - \cos \theta_{c,c}. \quad (32)$$

Since $\partial \ell_c / \partial \Delta_{c,k} > 0$ (Equation (31)):

$$\frac{\partial \ell_c}{\partial \theta_{c,c}} = \sum_{k \neq c} \frac{\partial \ell_c}{\partial \Delta_{c,k}} \cdot \frac{\partial (\cos \theta_{c,k} - \cos \theta_{c,c})}{\partial \theta_{c,c}} = \sum_{k \neq c} \frac{\partial \ell_c}{\partial \Delta_{c,k}} \cdot \sin \theta_{c,c} > 0, \quad (33)$$

$$\frac{\partial \ell_c}{\partial \theta_{c,k}} = \frac{\partial \ell_c}{\partial \Delta_{c,k}} \cdot \frac{\partial (\cos \theta_{c,k} - \cos \theta_{c,c})}{\partial \theta_{c,k}} = \frac{\partial \ell_c}{\partial \Delta_{c,k}} \cdot (-\sin \theta_{c,k}) < 0. \quad (34)$$

Because gradient descent updates each parameter against its gradient, Equation (33)(34) show that $\theta_{c,c}$ is persistently driven downward (prototype aligns with its own cluster center) while $\theta_{c,k}$ is driven upward (cluster centers of other experts are pushed away), which is precisely Equation (11). $\square$

A.3 Proof of Lemma 3–Hyperspherical Equivalence

Variable definitions follow Lemma 3 (Section 4.3). We prove that Euclidean distance, cosine similarity, and angular distance induce identical rankings on the unit hypersphere $\mathbb{S}^{D-1}$.

Let $\mathbf{u} \in \mathbb{S}^{D-1}$ be a fixed anchor, and let $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{S}^{D-1}$ be two candidates. Define the following three ordering propositions:

$$\begin{aligned} \mathcal{A} : & \|\mathbf{u} - \mathbf{v}_1\|_2 \leq \|\mathbf{u} - \mathbf{v}_2\|_2, \\ \mathcal{B} : & \langle \mathbf{u}, \mathbf{v}_1 \rangle \geq \langle \mathbf{u}, \mathbf{v}_2 \rangle, \\ \mathcal{C} : & \angle(\mathbf{u}, \mathbf{v}_1) \leq \angle(\mathbf{u}, \mathbf{v}_2). \end{aligned}$$

By proving $\mathcal{A} \Leftrightarrow \mathcal{B} \Leftrightarrow \mathcal{C}$, their equivalence is established through the transitivity.

For any $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{D-1}$, expanding the squared Euclidean distance gives

$$\|\mathbf{u} - \mathbf{v}\|_2^2 = \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 - 2\langle \mathbf{u}, \mathbf{v} \rangle = 2 - 2\langle \mathbf{u}, \mathbf{v} \rangle, \quad (35)$$

where the last step uses $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$. Thus,

$$\|\mathbf{u} - \mathbf{v}_1\|_2 \leq \|\mathbf{u} - \mathbf{v}_2\|_2 \Leftrightarrow \langle \mathbf{u}, \mathbf{v}_1 \rangle \geq \langle \mathbf{u}, \mathbf{v}_2 \rangle \quad (36)$$

Table 2: Comparison of Person ReID Datasets.

Type	Dataset	Year	Scenes	# Cameras	# IDs	# Images
Image	CUHK03 [28]	2014	Campus	10	1,360	13,164
	Market1501 [6]	2015	Campus	6	1,501	32,668
	CUHK-SYSU [61]	2017	Street	–	8,432	18,184
	PRW [69]	2017	Campus	6	932	11,816
	MSMT17 [7]	2018	Indoor & Outdoor	15	4,101	126,441
	DeepSportRadar [68]	2022	Basketball Court	–	468	18,703
	IUST-PresonReID [32]	2025	Market, Campus, Mosque	–	1,847	117,455
Video	PRID [67]	2011	Street	2	385	14,558
	iLIDS-VID [66]	2014	Airport Hall	2	555	16,172
	MARS [63]	2016	Campus	6	1,246	78,282
	RPIfield [64]	2016	Sports Field	12	111	30,246
	LPW [65]	2018	Various	4	3,495	105,371
	LaST [62]	2021	Various	1,325	10,808	206,777
Image	Ours	2026	Combined	1,517	35,462	809,701

$\mathcal{A} \Leftrightarrow \mathcal{B}$ is proved. Similarly, for unit vectors, the inner product equals the cosine of the included angle: $\langle \mathbf{u}, \mathbf{v} \rangle = \cos(\angle(\mathbf{u}, \mathbf{v}))$. Since angles between unit vectors lie in $[0, \pi]$ and cosine is strictly decreasing on $[0, \pi]$:

$$\langle \mathbf{u}, \mathbf{v}_1 \rangle \geq \langle \mathbf{u}, \mathbf{v}_2 \rangle \Leftrightarrow \cos(\angle(\mathbf{u}, \mathbf{v}_1)) \geq \cos(\angle(\mathbf{u}, \mathbf{v}_2)) \Leftrightarrow \angle(\mathbf{u}, \mathbf{v}_1) \leq \angle(\mathbf{u}, \mathbf{v}_2) \quad (37)$$

$\mathcal{B} \Leftrightarrow \mathcal{C}$ is therefore proved. Therefore, the geometrical interpretation of Equation (10) (12) (8) and the effects of aforementioned Lemmas essentially describe a similar meaning: **All operations in HypMoE-ReID (Routing, Token Purification, Diversity Learning) reduce to angular alignment and separation on unit hypersphere $\mathbb{S}^{D-1}$, where smaller angular distance indicates higher semantic similarity, and the orthogonal directions enforce expert diversity.**

B Algorithms of Proposed Methods

The process of proposed HypMoE-ReID is illustrated in Algorithm 1

C Details of Datasets, MoE Architectures, and Metrics

C.1 Dataset Components

Our dataset consists of 13 publicly available ReID datasets, as summarized in Table 2. The image-based datasets span diverse scenes: *CUHK03* [28] is collected on the CUHK campus with both indoor and outdoor cameras. *Market-1501* [6] and *PRW* [69] are captured on campus under outdoor surveillance. *CUHK-SYSU* [61] covers street snapshots. *MSMT17* [7] is the largest single-source image ReID benchmark, encompassing 15 cameras across multiple indoor and outdoor scenes at varying times of day. *DeepSportRadar* [68] is collected from professional basketball games. *IUST-PersonReID* [32] includes scenes from markets, campuses, and mosques in Iran.

The video-based datasets cover equally varied environments: *PRID* [67] is recorded by two fixed street cameras. *iLIDS-VID* [66] is captured in a busy airport arrival hall under heavy occlusion. *MARS* [63] is captured on campus. *RPIfield* [64] is shot on an outdoor sports field with 12 cameras. *LPW* [65] collects pedestrians across three crowded scenes. *LaST* [62] is a large scale dataset with high extreme diversity across numerous scenes.

For the video-based datasets, we follow the protocol of [70] and treat each tracklet as a bag of frames, sampling individual frames to construct image-level training and evaluation sets. This allows all datasets to be incorporated into a unified image ReID framework.

Our own dataset (*Ours*) aggregates all of the above sources together, resulting in 35,462 identities and 809,701 images across 1,517 cameras, making it a large-scale ReID benchmark fully reflecting real-world distribution.

Algorithm 1 HypMoE-ReID

Input: Training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$; pre-trained ViT weights $\mathbf{W}_0$; epochs E ; N_{re} routed experts, N_{se} shared experts, Top- K routing; temperature τ .

Output: Trained HypMoE-ReID model $\mathcal{M}$.

MoE Initialization

Load $\mathbf{W}_0$; for each MoE block $l \in \mathcal{M}$, copy pre-trained FFN weights to every routed expert $\{E_r^{(l)}\}_{r=1}^{N_{re}}$ with small noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Training

for $e = 1$ **to** E **do**

for each mini-batch $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$ **do**

$\mathbf{z}^{(0)} \leftarrow \text{PatchEmbed}(\mathbf{x}) + \mathbf{p}_{pos}$

for $l = 1$ **to** L **do**

$\mathbf{z}_i^{(l)} \leftarrow \mathbf{z}^{(l-1)} + \text{MHSA}(\text{LN}(\mathbf{z}^{(l-1)}))$

if $l \in \mathcal{M}$ (*Is MoE Layer*) **then**

$\hat{\mathbf{X}} \leftarrow \ell_2\text{-normalize}(\text{LN}(\mathbf{z}_i^{(l)}).\text{reshape}(BN_t, D)); \hat{\mathbf{P}} \leftarrow \ell_2\text{-normalize}(\mathcal{P}^{(l)})$

$\mathbf{S} \leftarrow \hat{\mathbf{X}} \hat{\mathbf{P}}^\top; (\mathbf{w}_t, \mathbf{r}_t) \leftarrow \text{Top-K}(\mathbf{S}, K)$

$\mathbf{z}_{mlp}^{(l)} \leftarrow \mathbf{0}_{BN_t \times D}$

for $c = 1$ **to** N_{re} **do**

$\mathcal{T}_c \leftarrow \{t \mid r_t = c\}$

if $|\mathcal{T}_c| > 0$ **then**

$\mathbf{z}_{mlp}^{(l)}[\mathcal{T}_c] += w_c \cdot E_c^{(l)}(\mathbf{x}[\mathcal{T}_c])$

end if

end for

$\mathbf{z}_{mlp}^{(l)} \leftarrow \mathbf{z}_{mlp}^{(l)} + E_s^{(l)}(\mathbf{x}); \text{reshape to } (B, N_t, D)$

Routing Purification

$\mathcal{V}^{(l)} \leftarrow \{c \mid |\mathcal{T}_c| > 0\}$

for $c \in \mathcal{V}^{(l)}$ **do**

$w_{t,c} \leftarrow [\mathbf{S}_{t,c}]_+ / \sum_{u \in \mathcal{T}_c} [\mathbf{S}_{u,c}]_+ \quad \text{for } t \in \mathcal{T}_c$

$\bar{\mathbf{f}}_c^p \leftarrow \ell_2\text{-normalize}\left(\sum_{t \in \mathcal{T}_c} w_{t,c} \cdot \hat{\mathbf{x}}_t\right)$

end for

Global Soft Separation Loss

$\mathbf{Q} \leftarrow [\hat{\mathbf{p}}_c]_{c \in \mathcal{V}^{(l)}}, \mathbf{F} \leftarrow [\bar{\mathbf{f}}_c^p]_{c \in \mathcal{V}^{(l)}}$

$\mathcal{L}_{glob}^{div, (l)} \leftarrow \text{CrossEntropy}(\mathbf{Q} \mathbf{F}^\top / \tau, [0, \dots, |\mathcal{V}^{(l)}| - 1])$

Local Hard Diversification Loss

$d_{ap,c} \leftarrow \|\hat{\mathbf{p}}_c - \bar{\mathbf{f}}_c^p\|_2; d_{an,c} \leftarrow \min_{k \in \mathcal{V}^{(l)}, k \neq c} \|\hat{\mathbf{p}}_c - \bar{\mathbf{f}}_k^p\|_2$

$\mathcal{L}_{loc}^{div, (l)} \leftarrow \text{SoftMarginLoss}(d_{an} - d_{ap}, \mathbf{1})$

$\mathcal{L}_{div}^{(l)} \leftarrow \mathcal{L}_{glob}^{div, (l)} + \beta \mathcal{L}_{loc}^{div, (l)}$

$\mathbf{z}^{(l)} \leftarrow \mathbf{z}_i^{(l)} + \mathbf{z}_{mlp}^{(l)}$

else

$\mathbf{z}^{(l)} \leftarrow \mathbf{z}_i^{(l)} + \text{FFN}(\text{LN}(\mathbf{z}_i^{(l)})); \mathcal{L}_{div}^{(l)} \leftarrow 0$

end if

end for

$\mathbf{f}_g \leftarrow \text{LN}(\mathbf{z}^{(L)})_{[:,0]}$

ReID Supervision

$\mathcal{L}_{ce} \leftarrow \text{CrossEntropy}(\text{Classifier}(\text{BN}(\mathbf{f}_g)), \mathbf{y}); \mathcal{L}_{tri} \leftarrow \text{TripletLoss}(\mathbf{f}_g, \mathbf{y})$

$\mathcal{L}_{reid} \leftarrow \mathcal{L}_{ce} + \alpha \mathcal{L}_{tri}$

Total Loss

$\mathcal{L}_{total} \leftarrow \mathcal{L}_{reid} + \sum_{l \in \mathcal{M}} \mathcal{L}_{div}^{(l)}; \text{Update } \mathcal{M} \text{ via } \nabla \mathcal{L}_{total}.$

end for

end for

return $\mathcal{M}$

C.2 MoE Architectures

All compared sparse MoE architectures are transplanted under ViT-B/16 backbone via the upcycling paradigm [71], where FFNs at even-indexed blocks are replaced by MoE layer whose experts are initialized from pretrained dense FFN with small perturbations. Unless otherwise noted, each routed expert E_m is a two-layer MLP and the gating distribution $R(\cdot)$ is computed over the routed expert pool.

Mixtral [38]. Each expert adopts the SwiGLU activation [72], introducing a learned gating path inside the expert:

$$E_m(\mathbf{x}) = \mathbf{W}_{d,m}(\text{SiLU}(\mathbf{W}_{g,m}\mathbf{x}) \odot \mathbf{W}_{u,m}\mathbf{x}), \quad \mathbf{W}_{g,m}, \mathbf{W}_{u,m} \in \mathbb{R}^{D \times d_e}, \quad \mathbf{W}_{d,m} \in \mathbb{R}^{d_e \times D}, \quad (38)$$

where d_e is the expert intermediate dimension and $\odot$ denotes element-wise product. The router is a standard linear-softmax gate without noise injection: $R(\mathbf{x}) = \text{Softmax}(\mathbf{x}\mathbf{W}_r) \in \mathbb{R}^{N_{re}}$, and the top- K entries are retained. No shared expert is used. The MoE output is:

$$\text{MoE}(\mathbf{x}) = \sum_{m \in \mathcal{K}} R_m(\mathbf{x}) E_m(\mathbf{x}), \quad (39)$$

where $\mathcal{K}$ is the set of selected top- K expert indices. GShard-style load-balancing loss [53] is applied.

LLaMA-MoE [40]. The expert structure is identical to Mixtral. The key distinction lies in MoE hidden dimension, where LLaMA-MoE constructs experts by splitting the hidden dimension of a pretrained dense FFN, so that the union of all expert parameters exactly reconstructs the original FFN.

$$\mathbf{W}_u = [\mathbf{W}_{u,1}^\top \cdots \mathbf{W}_{u,N_{re}}^\top]^\top, \quad \mathbf{W}_{u,m} \in \mathbb{R}^{D \times (d_e/N_{re})}. \quad (40)$$

No shared experts is adopted. Token-level balance loss follows Switch Transformer [54].

Yuan 2.0-M32 [42]. Yuan replaces the standard linear router with an attention-based router. For input token $\mathbf{x}$ and a learned routing query $\mathbf{Q}_r \in \mathbb{R}^D$ and expert key matrix $\mathbf{K}_r \in \mathbb{R}^{N_{re} \times D}$, the routing score is:

$$s_m = \frac{\exp(\mathbf{Q}_r(\mathbf{x}) \cdot \mathbf{k}_{r,m}/\sqrt{D})}{\sum_{j=1}^{N_{re}} \exp(\mathbf{Q}_r(\mathbf{x}) \cdot \mathbf{k}_{r,j}/\sqrt{D})}. \quad (41)$$

providing a bilinear, scale-invariant routing score rather than a simple dot product. Expert FFNs use standard GELU activations. No shared experts.

Grok 2.5 [39]. Grok introduces two architectural features absent in all other compared methods. First, at the attention level, raw attention logits are soft-clamped before the softmax via a tanh-based cap with threshold c :

$$a_{ij} \leftarrow c \cdot \tanh\left(\frac{a_{ij}}{c}\right), \quad (42)$$

which smoothly bounds logits to $(-c, c)$ (with $c=30$ in our implementation) and prevents attention distributions from collapsing to near-one-hot spikes. Second, Grok replaces the standard single pre-norm with a dual-norm scheme. Each sub-layer (attention and FFN) is wrapped by both a pre-layer normalization and a post-layer normalization:

$$\mathbf{x} \leftarrow \mathbf{x} + \text{PostNorm}(\text{SubLayer}(\text{PreNorm}(\mathbf{x}))). \quad (43)$$

This differs from all other compared architectures, which use a single pre-norm. The MoE routing itself is a standard linear-softmax gate with SwiGLU experts and no shared expert.

DBRX [41]. DBRX uses a GLU-based expert identical to Mixtral in form, but increases the Top-K value to $K = 4$, the highest among all compared methods. With $N_{re} = 12$ and $K = 4$, each token activates one third of all routed experts simultaneously. The router is a standard linear-softmax gate. The elevated K value grants each token a richer mixture of expert outputs but substantially increases the active parameter count per forward pass, as reflected in the FLOPs and APE columns of Table 1. No shared experts.

DeepSeek-R1 [36]. DeepSeek introduces two key designs absent in the above architectures. First, the routing uses sigmoid normalization on per-expert scalar scores rather than a global softmax, so each expert score is computed independently.

$$s_m = \sigma(\mathbf{x}^\top \mathbf{w}_{r,m} + b_{r,m}), \quad \mathbf{w}_{r,m} \in \mathbb{R}^D, \quad (44)$$

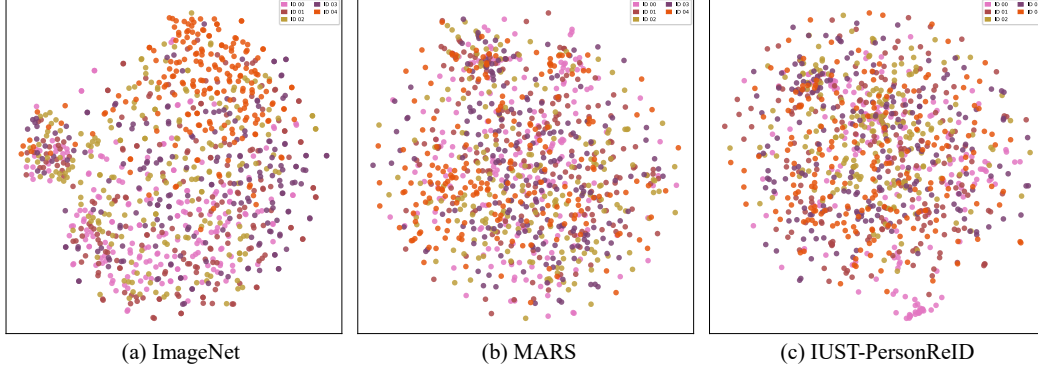


Figure 7: TSNE visualization of ImageNet patch tokens and ReID patch tokens. Features are extracted by pretrained ViT-B/16. ReID data (MARS, IUST-PersonReID) features extreme homogeneity compared to classic vision data.

and the top- K selected weights are then ℓ_1 -normalized: $R_m = s_m / \sum_{j \in \mathcal{K}} s_j$. This decouples experts from competing against each other during gating, allowing each expert’s activation probability to be calibrated independently. Second, shared experts $\{E_s\}_{s=1}^{N_{se}}$ always receive every token, providing a stable component that absorbs common representations:

$$\text{MoE}(\mathbf{x}) = \sum_{s=1}^{N_{se}} E_s(\mathbf{x}) + \sum_{m \in \mathcal{K}} R_m E_m(\mathbf{x}). \quad (45)$$

Expert FFNs use SwiGLU. An auxiliary load-balancing loss [53] is applied at the sequence level.

Qwen 2.5 [37]. Qwen 2.5 retains the shared-expert design from DeepSeek but reverts to a global softmax router over all routed experts, identical to Mixtral: $R(\mathbf{x}) = \text{Softmax}(\mathbf{x}\mathbf{W}_r)$, selecting the top- K entries. This makes routing scores mutually competing, in contrast to DeepSeek’s per-expert sigmoid. Expert FFNs are SwiGLU. The combination of global softmax routing and shared experts makes Qwen 2.5 the closest architecture to ours in terms of overall structure. The primary differences are the routing score type and the absence of the proposed diversity learning mechanism.

C.3 Metrics

mAP and R@1 are standard performance metrics for ReID [30; 11], while per token FLOPs, parameters reflect the overhead. APE jointly measures accuracy and parameter efficiency, defined as:

$$\text{APE} = \frac{\text{Average Acc.}}{\text{Act. Params. (M)}} \times 100. \quad (46)$$

where Average Acc. denotes the average mAP or R@1 across all constituent test sets, and Act. Params. is the number of active parameters in millions per forward pass. A higher APE indicates better accuracy per unit of model capacity, where we validate that HypMoE-ReID features the best trade-off between capacity, efficiency and performance.

D Additional Analysis and Visualizations

D.1 On the Essence of Routing Collapse in Person ReID

Person ReID essentially falls within the scope of fine-grained classification. Different identities share nearly identical semantic contents, differing only in subtle low-level cues such as texture, color pattern, and local appearance. As visualized in Figure 7, ViT features extracted from ReID data collapse into a highly homogeneous distribution in latent space, whereas features from ImageNet, which contains semantically diverse categories, retain mild inter-class separation. Such homogeneity is precisely what renders standard MoE routing collapse in ReID. Additionally, even on ImageNet

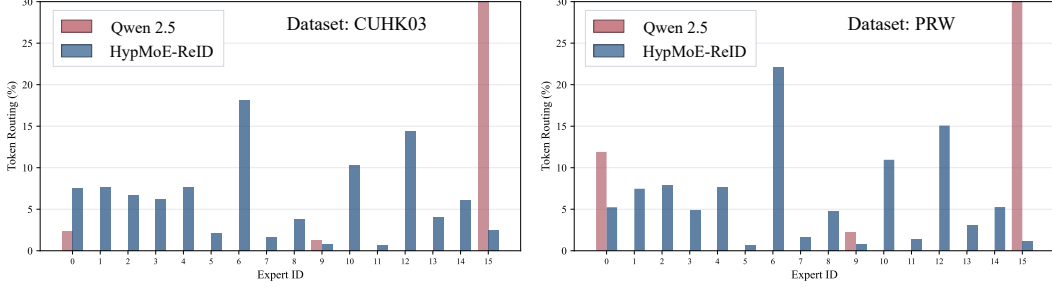


Figure 8: Comparison of MoE routing pattern on ReID datasets.

containing semantically diverse categories, sparse architectures have been reported to under-perform their dense counterparts [22]. In the much more homogeneous latent space of person ReID where features are nearly collapsed across identities, standard MoE routing therefore suffers from even more severe collapse. Our proposed method directly addresses this problem by introducing token purification and routing diversity learning mechanisms.

D.2 Routing Pattern of Sparse Architectures

Figure 8 compares the routing patterns of different sparse architectures on two representative ReID datasets. While Qwen 2.5 [37] exhibits severe routing collapse where most tokens are dispatched to few experts, our HypMoE-ReID yields a balanced and semantically diverse expert assignment. This phenomenon confirms that the proposed Routing Diversity Learning mechanism could effectively prevent routing collapse and drive the router toward meaningful specializations.

D.3 Visualization on Person ReID Datasets

We provide additional qualitative results to further examine the behavior of different architectures. As shown in Figure 10, the proposed HypMoE-ReID consistently specializes individual experts onto semantically distinct body regions. This emergent specialization arises naturally from our routing mechanism and enables fine-grained, part-aware feature extraction. In contrast, the DeepSeek-R1 MoE [36] exhibits routing collapse where most tokens are routed to few experts regardless of their spatial location. The dense ViT is susceptible to noisy, background-dominated patches due to its universally shared parameters. Its attention frequently drifts toward irrelevant regions such as ground or cluttered backgrounds, degrading discriminative representations. These observations further validate our quantitative results and confirm that fine-grained part specialization is both achievable and beneficial under our proposed routing mechanism.

E Performance on Large Scale Vehicle Re-Identification

E.1 Large Scale Vehicle ReID Datasets

To further validate the generalizability of our framework, we build a large scale vehicle ReID benchmark by unifying five publicly available datasets: VeRi-776 [73], VERI-Wild [74], VRAI [75], VRIC [76], and VehicleID [77], as summarized in Table 3. We follow the same unification protocol as for the person benchmark. Compared with existing single-source vehicle benchmarks, our combined set exhibits much larger intra-class variation and cross-domain shift, thus providing a more realistic proxy for real-world vehicle ReID.

E.2 Performance Analysis

Table 4 compares HypMoE-ReID with dense ViT and representative sparse architectures on the large scale vehicle benchmark. Our method achieves **69.74% / 79.72%** mAP/R@1 on APE, substantially outperforming all competitors. Notably, HypMoE-ReID surpasses the best sparse baseline Qwen 2.5 by **7.92% / 2.91%** on APE mAP/R@1 and by **12.35% / 8.38%** on Average mAP/R@1. Additionally,

Table 3: Comparison of Vehicle ReID Datasets.

Type	Dataset	Year	Scenes	# Cameras	# IDs	# Images
Image	VeRi-776 [73]	2016	Urban Road	39	776	51,035
	VERI-Wild [74]	2019	Large scale Urban CCTV	–	40,671	416,314
	VRAI [75]	2019	Aerial	4	6,302	66,113
	VRIC [76]	2018	Highway	–	2,792	54,808
	VehicleID [77]	2016	Urban Road	–	26,328	221,567
Image	Ours	2026	Combined	45	76,888	809,837

Table 4: Comparison Results on Large Scale Vehicle ReID.

Model	Arch.	Pub.	# Config.	# FLOPs.	# Params.(Act./Act. Exp.)	APE.		Average Acc.	
						mAP	R@1	mAP	R@1
ViT [14]	DENSE	<i>ICLR 2021</i>	-	11.35 G	100.50 M / -	58.33	70.08	58.62	70.43
Qwen 2.5 [37]	SPARSE	<i>arXiv 2025</i>	R16, S1, K2	10.45 G	93.48 M / 21.24 M	61.82	76.81	57.79	71.80
LLaMA-MoE [40]	SPARSE	<i>EMNLP 2025</i>	R8, S0, K2	11.81 G	104.10 M / 10.62 M	45.72	63.28	47.60	65.87
Mixtral 8x7B [38]	SPARSE	<i>arXiv 2024</i>	R8, S0, K2	18.66 G	157.14 M / 84.93 M	8.29	18.31	13.03	28.78
DeepSeek-R1 [36]	SPARSE	<i>Nature 2025</i>	R12, S1, K2	37.31 G	157.20 M / 85.00 M	31.30	41.81	49.21	65.72
HypMoE-ReID	SPARSE	<i>This Paper</i>	R16, S1, K1	11.36 G	100.58 M / 28.34 M	69.74	79.72	70.14	80.18
						+7.92↑	+2.91↑	+11.52↑	+9.75↑

R = number of routed experts, S = number of shared experts, K = Top-K routing.

it improves over the dense ViT baseline by **11.52%/9.75%** on Average mAP/R@1, highlighting that the rigidity of dense representations is even more pronounced in the vehicle domain, where subtle design cues (e.g., front lights) must be distinguished across drastically different scene contexts. Other sparse architectures such as Mixtral and DeepSeek-R1 suffer from severe performance degradation, confirming that standard MoE routing collapses on fine-grained, homogeneous data. These results demonstrate that HypMoE-ReID generalizes well beyond person ReID and can serve as a strong sparse backbone for large scale vehicle retrieval.

E.3 Visualizations

Figure 9 visualizes the attention maps of the dense ViT, the sparse Qwen 2.5, and our HypMoE-ReID. Dense ViT often focuses its attention over large background areas such as road surfaces, failing to isolate the discriminative parts of the target vehicle. Qwen 2.5, despite employing a mixture-of-experts, exhibits a severe routing collapse. In contrast, HypMoE-ReID successfully learns specialized experts, allowing different experts to consistently focus on structurally stable vehicle components. Such semantically aware expert specialization not only leads to better performance but also offers well-established interpretability, which further validates the effectiveness of proposed HypMoE-ReID.

F Broader Impacts

Our work studies person ReID at scale using sparse MoE architectures. The potential positive and negative societal impacts are as follows.

The potential positive societal impacts of this research include:

Public Safety and Surveillance Efficiency. Accurate large scale ReID can assist law enforcement and public safety agencies in locating missing persons or tracking suspicious individuals across camera networks, potentially reducing response times in critical scenarios.

Efficiency via Sparse Computation. By achieving performance gains without increasing active FLOPs, our framework demonstrates a path toward high-capacity visual models that do not proportionally increase energy consumption, which is beneficial for sustainable AI deployment.

The potential negative societal impacts of this research include:

Potential for Bias. The mixed dataset inherits demographic and geographic biases from its constituent sources, which may cause uneven recognition performance across different populations, raising fairness concerns in real-world deployment.

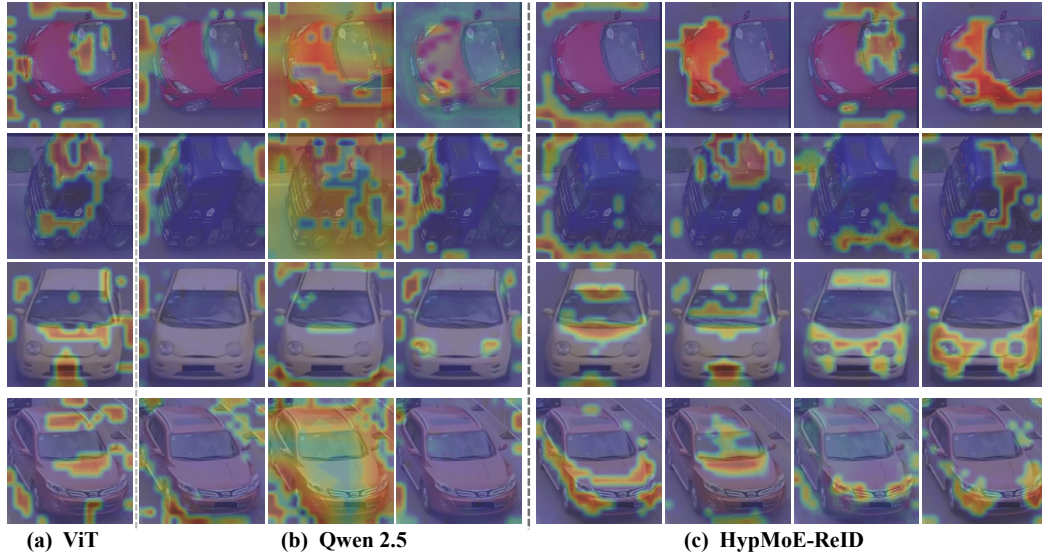
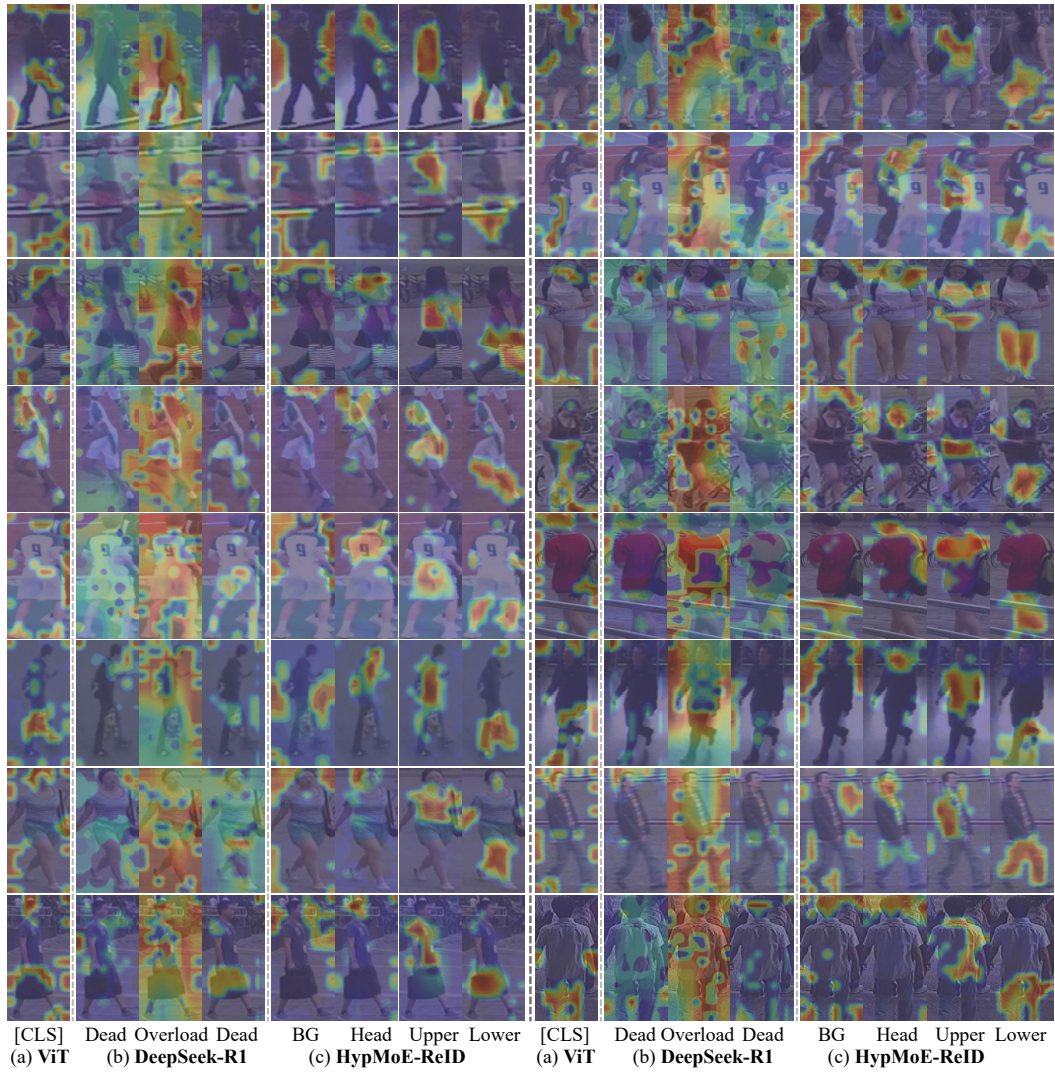


Figure 9: Visualization vehicle ReID between dense ViT, sparse Qwen 2.5 and HypMoE-ReID.



23
Figure 10: Attention heatmap visualizations on different architectures.

NeurIPS Paper Checklist

1. Claims

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

Answer: [\[Yes\]](#)

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

Guidelines:

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

2. Limitations

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

Answer: [\[Yes\]](#)

Justification: We have analyzed the limitations of proposed method over computational overhead.

Guidelines:

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

3. Theory assumptions and proofs

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

Answer: [\[Yes\]](#)

Justification: We have provided a theoretical analysis of the geometrical essence of our framework with a complete theoretical derivation proof.

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [\[Yes\]](#)

Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper.

Guidelines:

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

5. Open access to data and code

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

Answer: [Yes]

Justification: We have attached our source code and data access links in the supplementary materials, with sufficient instructions to faithfully reproduce the main experimental results.

Guidelines:

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

6. Experimental setting/details

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

Answer: [Yes]

Justification: We have specified all the training and test details necessary to understand the results.

Guidelines:

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

7. Experiment statistical significance

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

Answer: [No]

Justification: Error bars are not reported due to computational constraints. However, all experiments are conducted with consistent benchmark configurations and fixed random seeds, ensuring reproducibility. Improvements on multiple metrics substantiate the effectiveness of our method.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. Experiments compute resources

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

Answer: [Yes]

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

Guidelines:

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

9. Code of ethics

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

Answer: [Yes]

Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

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

10. Broader impacts

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

Answer: [Yes]

Justification: The Broader impacts has been discussed in the Appendix.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.

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

11. Safeguards

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

Answer: [N/A]

Justification: The paper poses no such risks.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [Yes]

Justification: They are properly credited and respected.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [N/A]

Justification: The paper does not release new assets.

Guidelines:

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

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

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

Answer: [Yes]

Justification: All the data used is publicly available and widely used in academia. Therefore, our data is usable and no personal information is required.

Guidelines:

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

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

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

Answer: [N/A]

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

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

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

Guidelines:

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