HAMoBE: Hierarchical and Adaptive Mixture of Biometric Experts for Video-based Person ReID
Supplementary Materials

Yiyang Su[1*] Yunping Shi[2*] Feng Liu[2] Xiaoming Liu[1]
[1] Department of Computer Science and Engineering, Michigan State University
[2] Department of Computer Science, Drexel University
{suyiyan1, liuxm}@msu.edu, {ys839, fl397}@drexel.edu

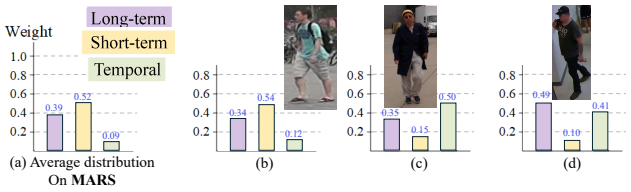

Figure 1. (a) Average $\mathbf{W}^{*2}$ distribution on the MARS dataset. One example from (b) MARS, (c) CCVID, and (d) MEVID.

A. Additional Experiments

Number of First-Layer Experts. As suggested, we ablate the number of first-layer experts (4, 8, 16). Rank-1 accuracy under the different-clothes setting on CCVID improves from 82.6% (4 experts) to **85.8%** (8 experts), but drops to 84.5% with 16, due to redundancy and reduced gating selectivity. This validates 8 as the most effective choice.

Ablaiton on Backbone. We train a model using a ViT-B backbone to align with TF-CLIP. While this reduces overall capacity, it still outperforms TF-CLIP in Rank-1 accuracy under the *consistent-clothing setting*, achieving **90.5%** (-0.6%) vs. 89.4% on MARS and **84.8%** (-0.4%) vs. 83.8% on LS-VID. This highlights that our gains primarily stem from the framework design rather than the backbone size.

Temporal Aggregation Strategy. We compare 4 temporal aggregation methods on MEVID under the different-clothes setting: mean pooling (9.8%), LSTM (12.6%), Transformer encoder (15.9%), and our Transformer decoder (**20.1%**). The decoder performs best, as its learnable query enables selective, frame-aware aggregation, offering better robustness to appearance changes.

Additional Visualizations. Fig.1(a) shows expert weight distributions in consistent-clothing scenarios. The increased contribution of ***short-term*** experts aligns with our design, which emphasizes appearance cues when clothing is stable.

If we compare Fig. **??** and Fig. 1, expert weights adapt meaningfully to context: long-term ***dominates*** on MEVID, short-term on MARS (*consistent-clothing*), and temporal on CCVID. These shifts reflect the model's ability to prioritize identity-relevant cues under varying conditions. Concrete examples in Fig. 1 further illustrate distinct expert dominance across cases, ruling out fixed bias and supporting effective disentanglement.

B. Limitations and Societal Impacts

Despite advancements, HAMoBE has limitations that require further exploration. Currently, HAMoBE does not incorporate background information, potentially reducing its effectiveness in cluttered environments. Its performance depends heavily on the quality and variability of input video data, making it less effective in low-resolution or occluded conditions. While HAMoBE aims to enhance person ReID for public safety, it also raises ethical concerns regarding privacy and surveillance, underscoring the need for stringent safeguards to prevent misuse.

*Equal contribution