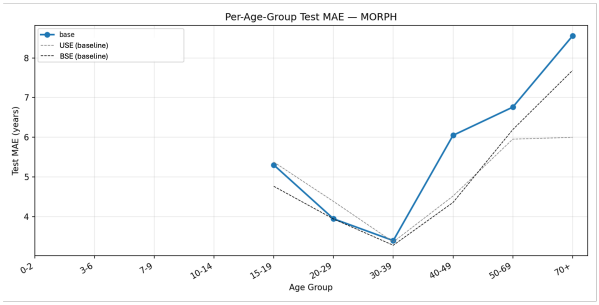
AgeMoVLE: Age Estimation with Mixture of VL Experts

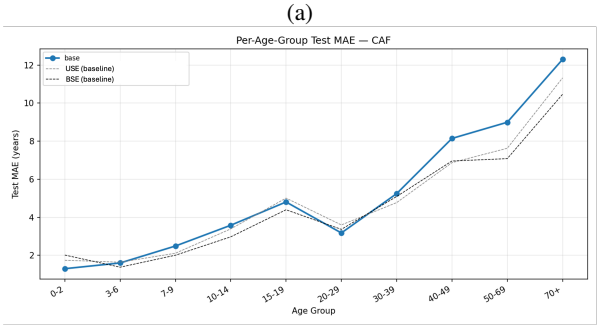
Supplementary Material

1. Cross-Dataset Per-Age-Group MAE

Fig. 1 extends the per-age-group MAE analysis of Fig. 4 in the main paper to the two held-out datasets, MORPH and CAF. On MORPH (adults only, ages 16–77), AgeMoVLE tracks the baselines closely in the well-represented range of 20–39 but diverges at the older tail (40+), where the router falls behind the Uniform Static Ensemble (USE), reflecting the distribution mismatch between the UTKFace training set and the adult-only MORPH test set. On CAF, the router yields the largest gains at the youngest ages (0–2 and 3–6) and in the 20–29 range, consistent with the UTKFace results, confirming that the routing advantage is dataset-agnostic in the groups that are structurally underrepresented in training. In both cases, the router was trained solely on UTKFace with no retraining on the target dataset.



(a)



(b)

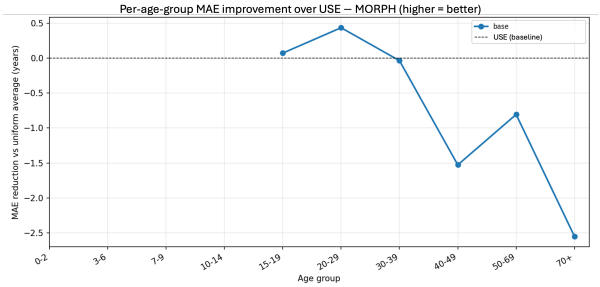
Figure 1. Per-age-group Test MAE for AgeMoVLE (`base`) on (a) MORPH and (b) CAF, compared to the USE and BSE baselines. The router was trained on UTKFace only; no retraining was performed on either held-out dataset.

2. Cross-Dataset MAE Improvement Over Baselines

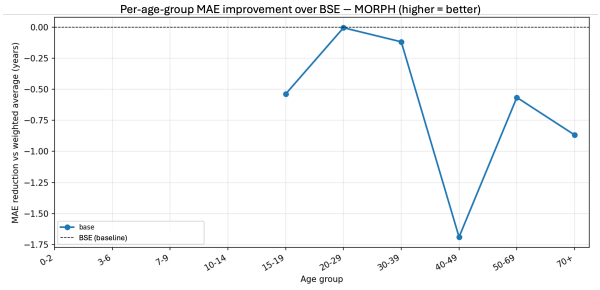
Fig. 2 and Fig. 3 report the per-age-group MAE improvement of AgeMoVLE (`base`) over the USE and Best Static Ensemble (BSE) baselines on MORPH and CAF respectively, mirroring the UTKFace analysis in Fig. 3 of the main paper.

On MORPH (Fig. 2), improvements over USE are concentrated in the young-adult range (15–19), while degradation is observed at the older tail (40+), reflecting the mismatch between the UTKFace training distribution and the adult-only MORPH distribution. Compared to BSE, the router offers comparable or marginal gains across most age groups, consistent with the narrower routing advantage expected when tail groups are absent.

On CAF (Fig. 3), the router recovers substantial gains over USE at ages 0–2 and 20–29, while performance degrades relative to USE at ages 40 and above. Compared to BSE, the router outperforms at the youngest groups but falls behind in the older ranges, indicating that the BSE's static weights are better calibrated for the CAF adult distribution than for the child groups. These patterns corroborate the main-paper finding that routing gains are largest precisely where training data are most sparse.

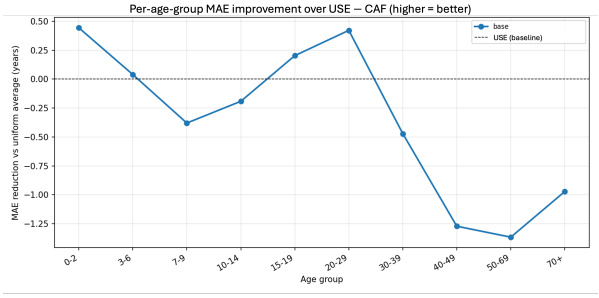


(a)

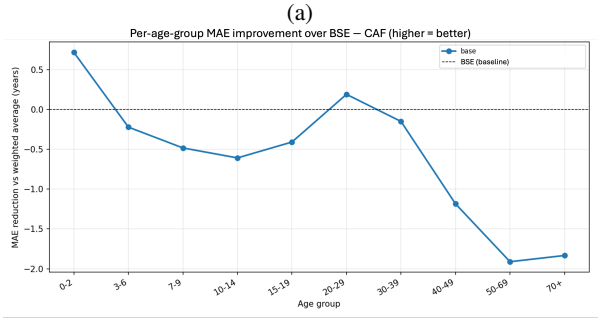


(b)

Figure 2. Per-age-group MAE improvement of AgeMoVLE (`base`) over (a) USE and (b) BSE on MORPH. Positive values indicate lower MAE than the respective baseline.



(a)



(b)

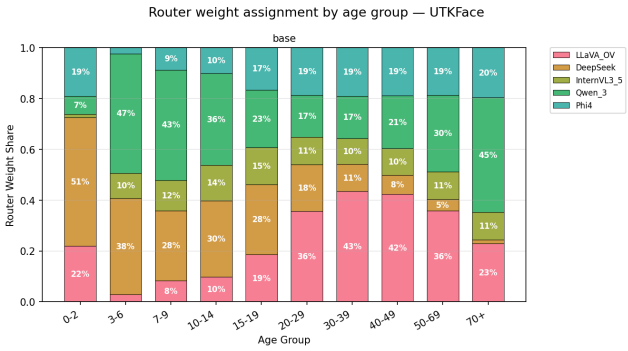
Figure 3. Per-age-group MAE improvement of AgeMoVLE (`base`) over (a) USE and (b) BSE on CAF. Positive values indicate lower MAE than the respective baseline.

3. Router Weight Assignment Across Datasets

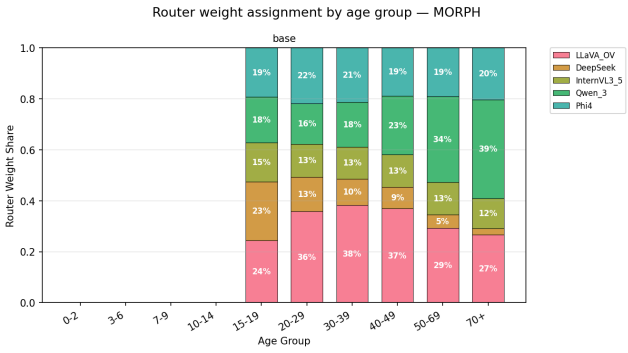
Fig. 4 shows the mean router weight assigned to each VLM expert per age group on UTKFace, MORPH, and CAF. These bar charts complement the line plot in Fig. 5 of the main paper and extend it to the two held-out datasets.

Several patterns are consistent across all three datasets. First, DeepSeek-VL receives the highest weight in the youngest age groups (0–9), while LLaVA-OV becomes increasingly dominant in the middle-age range (20–49). Second, Qwen3-VL carries a relatively small share in most groups on UTKFace and CAF, yet its weight is more evenly distributed on MORPH. Third, Phi-4 and InternVL3.5 exhibit stable, moderate contributions across all age groups and datasets, acting as consistent secondary experts rather than age-specific specialists.

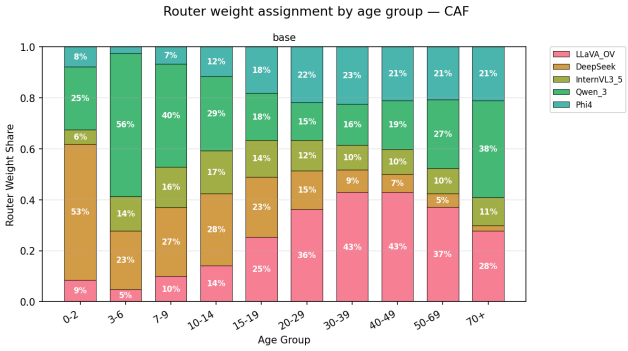
Notably, the routing pattern on MORPH is more uniform across experts than on UTKFace or CAF, consistent with MORPH covering only the adult range where no single expert dominates and all experts produce competitive predictions. On CAF, the weight assignment closely mirrors the UTKFace pattern, despite the router never being retrained on CAF, providing evidence that the learned routing strategy generalizes across datasets with different demographic compositions.



(a)



(b)



(c)

Figure 4. Mean router weight assignment per expert and age group for AgeMoVLE (`base`) on (a) UTKFace, (b) MORPH, and (c) CAF. The dominant expert shifts systematically with age group across all three datasets, and the routing pattern on CAF closely mirrors UTKFace despite no retraining.

4. Training Dynamics

Fig. 5 shows the fold-averaged training and validation MAE curves for AgeMoVLE (`base`) over 60 epochs. The training MAE decreases steadily throughout, while the validation MAE converges and plateaus around epoch 30 (marked by a star indicating the best-checkpoint epoch), after which it remains stable with no sign of overfitting. The shaded regions represent ± 1 standard deviation across the five folds, confirming low variance and stable training across folds.

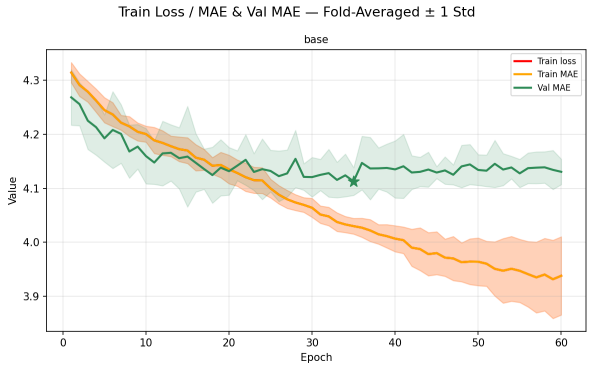


Figure 5. Fold-averaged training loss/MAE and validation MAE for AgeMoVLE (`base`) over 60 epochs (± 1 std. across folds shown as shaded bands). The star marks the best-checkpoint epoch averaged across folds.

The early plateau of validation MAE relative to training MAE is consistent with the router's limited capacity (fewer than 100K trainable parameters): the model reaches its expressive limit early, and further training reduces training error without harming generalization.