DEPTH-GRADED PARAMETER-EFFICIENT ADAPTATION OF FOUNDATION MODELS FOR FACE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision foundation models are an attractive starting point for training face recognition (FR) models, which often require large identity-labelled datasets. Current approaches to use foundation models for FR adapt them with low-rank adapters that use the same rank in every transformer block. We ask *where in the foundation model network needs capacity for FR adaptation*. After probing CLIP ViT-L/14 with high-rank adapters, we find that the rank of the learned adapter grows with model depth, while a uniform rank-16 adapter is saturated in deep blocks and under-used in shallow ones. Motivated by this, we propose *depth-graded rank allocation*, a search-free schedule that increases adapter rank linearly with depth. We pair it with rank-stabilised scaling and instantiate it with low-rank and weight-decomposed adaptation (called DG-LoRA and DG-DoRA, respectively). We study eight ViT-L foundation models in our experiments. We show that depth-graded ranks improve adapted ViT-L over uniform rank adapters with the same budget. Trained on WebFace4M our model with 14M trainable parameters reaches 95.1% TAR on IJB-C (FAR=10^{-4}), surpassing prior parameter-efficient adaptation of foundation models and a ViT-L trained from scratch on the same data.

1 INTRODUCTION

Face recognition (FR) models have reached high accuracy by training convolutional or transformer backbones from scratch with margin-based softmax losses on millions of labelled identities (Deng et al., 2019; Wang et al., 2018; Kim et al., 2022; Zhu et al., 2021). Vision foundation models, such as CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2023), SigLIP (Zhai et al., 2023), etc., are pretrained on large corpus of data and can offer a different starting point. Pre-trained vision foundation models encode rich visual structure that can be turned into an identity representation as well. However, as off-the-shelf models, they have poor performance for face recognition task. Chettaoui et al. (2025) showed that most of this gap can be closed with parameter-efficient fine-tuning. They used low-rank adaptation (LoRA) (Hu et al., 2021) of rank 16 in the attention projections of every block (with same rank) and trained the model with a margin loss. Their results indicated that low-rank adaptation can bring CLIP and DINOv2 close to ViTs trained from scratch.

FR is fine-grained and is a different task from the general data used in pretraining of foundation models. FR models need to separate thousands of similar identities while staying invariant to pose, age, and illumination. In addition, identity rarely appears in the image captions in pretraining data, or no self-supervised objective asks a foundation model to encode identity information during pretraining. Several studies consistently find that early layers learn generic features that transfer across tasks, whereas later layers are task-specific (Yosinski et al., 2014; Raghu et al., 2021). This suggests that adapting a foundation model to FR should adapt the deep blocks much more than the shallow ones, and that a uniform rank budget spends capacity where it is least needed. We test this hypothesis directly. We train adapters with a deliberately generous rank (256 in every block of CLIP ViT-L/14) and inspect the singular value spectra of the updates they learn (Figure 1). The effective dimensionality of the update grows almost monotonically with depth. In the first block, 19 directions carry 90% of the energy of the value-projection update, whereas the last block needs 100. Conversely, when the rank is fixed to 16, deep adapters use practically all of their rank while shallow ones use about 60% of it. Capacity is saturated exactly where FR needs it most.

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

Motivated by this observation, we propose *depth-graded rank allocation* (DG), where the adapter rank grows from a small value in the first block to a large value in the last. We also use rank-stabilised scaling (Kalajdzievski, 2023), so that a higher rank adds directions without inflating the update. We deploy our rank allocation to low-rank adaptation (Hu et al., 2021) and weight-decomposition adaptation (Liu et al., 2024a), and call the resulting methods DG-LoRA and DG-DoRA, respectively. We use our proposed adaptation methods to adapt foundation models for FR using an angular margin loss. The resulting model, called FMFace, achieve state-of-the-art performance for adapting foundation models for face recognition.

In summary, our contributions are as follows:

- We perform a **spectral rank probing analysis** of learned adapter updates, which shows that the effective rank grows with depth and also uniform adapters saturate in deep blocks.
- We propose a **depth-graded rank allocation** method and integrate it with LoRA and DoRA (called DG-LoRA and DG-DoRA). The proposed rank allocation schedule adds no search, pruning or training overhead.
- The adapted models achieve **state-of-the-art performance** for parameter-efficient adaption of foundation models for FR. Trained on WebFace4M, our adaptation of CLIP ViT-L/14 with 14M trainable parameters reach 95.1% TAR at FAR$= 10^{-4}$ on IJB-C, which is above prior foundation-model adaptation and same ViT-L trained from scratch.

2 Related Work

Deep face recognition. FR models learn a hyperspherical embedding with marginal softmax losses such as SphereFace (Liu et al., 2017), CosFace (Wang et al., 2018), ArcFace (Deng et al., 2019), CurricularFace (Huang et al., 2020), MagFace (Meng et al., 2021), and AdaFace (Kim et al., 2022). These losses are typically applied to ResNet backbones trained from scratch on MS1MV2 (Deng et al., 2019) or WebFace260M subsets (Zhu et al., 2021), using the Partial FC method (An et al., 2022) used to scale the classification layer to millions of identities. Vision transformers (ViTs) have also been trained from scratch for FR, but require large identity-labelled datasets to achieve a remarkable performance. (Zhong & Deng, 2021; Dan et al., 2023; Qin et al., 2023)

Foundation models for faces. General-purpose vision encoders are pre-trained contrastively on image-text pairs, as in CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023), EVA-CLIP (Sun et al., 2023), DFN (Fang et al., 2024), MetaCLIP (Xu et al., 2024), SigLIP (Zhai et al., 2023) and SigLIP 2 (Tschannen et al., 2025), or with self-supervision, as in DINOv2 (Oquab et al., 2023). FaRL (Zheng et al., 2022) pre-trains on face image-text pairs for facial analysis tasks, and LAFS (Sun et al., 2024) learns landmark-based self-supervised face representations. Without adaptation, foundation models remain far behind domain-specific FR models (Sony et al., 2025; Ozturk et al., 2026; Shahreza & Marcel, 2025; 2026). The IJCB 2026 AFMFR competition (Chettaoui et al., 2026) benchmarked the adaptation of CLIP ViT-L/14 to FR with synthetic data, where the strongest entries used full fine-tuning or rank-stabilised LoRA with a uniform rank. Closest to our work, FRoundation (Chettaoui et al., 2025) fine-tuned DINOv2 and CLIP for FR with LoRA at several data scales and showed that foundation models can be adapted for FR. Using a uniform rank ($r = 16$), their adaptation outperforms training from scratch with small training data (*e.g.,* CASIA-WebFace), but with larger data (*e.g.,* WebFace4M), their adaption led to a lower performance than training from scratch. We study the capacity and required rank at each block of foundation models and propose to replace uniform rank with depth-graded rank allocation. We study 8 foundation models and adapt them for FR, and show that our model adaptation can outperform training from scratch with both small and large datasets.

Parameter-efficient fine-tuning and rank allocation. Adapters (Houlsby et al., 2019), prompt tuning (Jia et al., 2022) and low-rank adaptation (LoRA; Hu et al., 2021) adapt foundation models by training few parameters. LoRA variants change the scaling (rsLoRA; Kalajdzievski, 2023), the parameterisation (DoRA; Liu et al., 2024a), the initialisation (PiSSA; Meng et al., 2024) or the optimisation (LoRA+; Hayou et al., 2024). Other LoRA variants may replace additive updates with orthogonal fine-tuning (OFT, BOFT; Qiu et al., 2023; Liu et al., 2024b) with uniform rank for every adapted matrix. Another approach is to allocate rank non-uniformly. AdaLoRA (Zhang

et al., 2023) prunes an SVD-parameterised budget by importance, SoRA (Ding et al., 2023) and DyLoRA (Valipour et al., 2023) learn or sample ranks, ARD-LoRA (Shinwari & Usama, 2025) and LAARA (Tripathi et al., 2026) learn per-layer ranks from meta-objectives or Fisher information, SR-LoRA (Zhang et al., 2025) sets them from the stable rank of the pre-trained weights, and PARA (Kumaravelu et al., 2026) prunes ranks after training with a global singular-value threshold. PRILoRA (Benedek & Wolf, 2024) uses an increasing rank combined with importance-based pruning of LoRA matrices on GLUE. In contrast to previous studies which are mainly in NLP domain, our rank allocation is motivated by a spectral probe of FR adaptation in ViTs. We benchmark it against equal-budget, and show how the adapter scale must be chosen when rank varies with depth.

3 METHOD

3.1 PRELIMINARIES

Low-rank adaptation of a ViT. Consider a vision transformer with L blocks of width d. Following common practice (Hu et al., 2021; Chettaoui et al., 2025), we adapt the query and value projections $W_0 \in \mathbb{R}^{d\times d}$, $p \in \mathcal{P} = \{q, v\}$, of every block ℓ and keep all pre-trained weights frozen. LoRA replaces a frozen matrix by

$$W = W_0 + \gamma_r\, BA, \qquad B \in \mathbb{R}^{d\times r},\ A \in \mathbb{R}^{r\times d}, \tag{1}$$

where B is initialised to zero, A randomly, and γ_r is a scaling factor. While, standard LoRA used $\gamma_r = \alpha/r$, rank-stabilised LoRA (rsLoRA; Kalajdzievski, 2023) used $\gamma_r = \alpha/\sqrt{r}$ and showed that with this scaling the adapter's forward contribution and its gradients neither explode nor collapse as rank r grows.

Weight-decomposed adaptation. DoRA (Liu et al., 2024a) reparameterises the adapted weight into a learnable magnitude and a low-rank-updated direction,

$$W = \boldsymbol{m} \odot \frac{W_0 + \gamma_r BA}{\|W_0 + \gamma_r BA\|_{\text{row}}}, \qquad \boldsymbol{m} \in \mathbb{R}^d, \quad \boldsymbol{m}^{(0)} = \|W_0\|_{\text{row}}, \tag{2}$$

where $\|\cdot\|_{\text{row}}$ is the vector of ℓ_2 norms of the rows (one per output unit), the division is row-wise, and the norm is treated as a constant in the backward pass. At initialisation $W = W_0$. Inspired by rescaling in rsLoRA, we define rsDoRA with $\gamma_r = \alpha/\sqrt{r}$. Both Eq. 1 and Eq. 2 can be merged into W_0 after training, so the adapted encoder has exactly the inference cost of the pre-trained one.

Face recognition objective. We use vision encoder of foundation models which are based on ViT. An aligned face image $\boldsymbol{x}$ is encoded by the ViT, which is then mapped by the encoder's output projection $P \in \mathbb{R}^{d_e\times d}$ to the identity embedding $f(\boldsymbol{x}) \in \mathbb{R}^{d_e}$ (d_e=768 for CLIP ViT-L/14). We use ArcFace (Deng et al., 2019) loss for our model adaptation:

$$\mathcal{L} = -\log \frac{e^{s\cos(\theta_y+m)}}{e^{s\cos(\theta_y+m)} + \sum_{j\neq y} e^{s\cos\theta_j}}, \qquad \cos\theta_j = \left\langle \frac{f(\boldsymbol{x})}{\|f(\boldsymbol{x})\|}, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} \right\rangle, \tag{3}$$

where $\boldsymbol{w}_j$ is the centre of identity j, y the label, s the scale and m the additive angular margin.

3.2 WHERE DOES FACE RECOGNITION ADAPTATION NEED CAPACITY?

A uniform rank implicitly assumes that every block needs the same number of update directions. We measure this directly with a rank probing analysis. To this end, we adapt CLIP ViT-L/14 on CASIA-WebFace with rsLoRA of rank r=256 in every block, which is large enough not to constrain any block, and analyse the learned updates $\Delta W_\ell^p = \gamma_r B_\ell^p A_\ell^p$. Let $\sigma_1 \geq \sigma_2 \geq \dots$ be their singular values. We consider the energy rank $r_{90}(\Delta W)$ as the number of singular directions carrying 90% of the energy of the learned update ΔW:

$$r_{90}(\Delta W) = \min\Big\{k : \textstyle\sum_{i\leq k} \sigma_i^2 \geq 0.9 \sum_i \sigma_i^2\Big\}. \tag{4}$$

In addition, we calculate the effective rank $\text{erank}(\Delta W)$ as defined in (Roy & Vetterli, 2007):

$$\text{erank}(\Delta W) = \exp\big(-\sum_i \bar{\sigma}_i \log \bar{\sigma}_i\big), \quad \bar{\sigma}_i = \sigma_i / \sum_j \sigma_j. \tag{5}$$

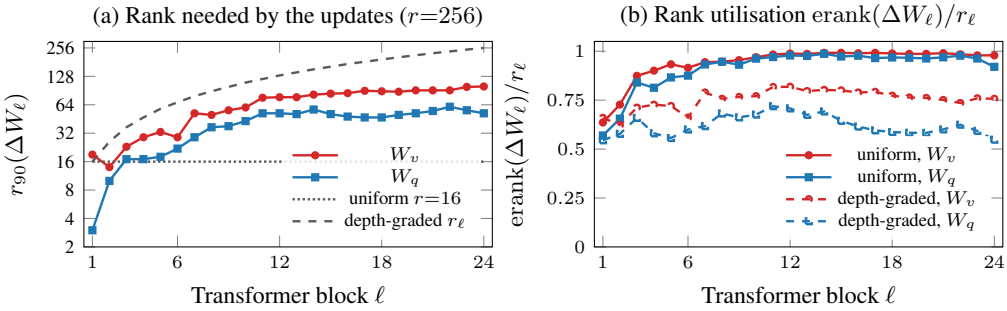

Figure 1: **Face recognition adaptation needs more rank in deeper blocks.** CLIP ViT-L/14 adapted on CASIA-WebFace. (a) With a large rank probe ($r = 256$ in every block), the number of singular directions carrying 90% of the energy of the learned update ΔW_ℓ grows with depths, far above the uniform rank of 16 used in prior work. (b) With uniform rank 16, deep adapters use almost all of their rank (effective rank $\geq 0.92\,r$ from block 10 onwards), a sign of saturation, while the first block uses about $0.6\,r$. In contrast, depth-graded ranks leave head-room at every depth.

Figure 1 illustrates the rank probing of the rsLoRA adapted CLIP ViT-L/14 at different model depths. Figure 1.a shows the energy rank (r_{90}) for each layer. As can be seen in Figure 1.a, the required rank increases in depth of the model. For W_v, r_{90} grows from 19 in the first block to 100 in the last, for W_q it grows from 3 to 52. Averaged over the first six blocks, only 17% (in W_v) and 9% (in W_q) of the update energy lies outside the top 16 directions. Therefore, a rank-16 adapter can represent the updates that shallow blocks learn, but it has a fraction of required capacity for deep blocks. Figure 1.b confirms this from the opposite point of view. Figure 1.b illustrates effective rank (erank) for different layers of the model when the rank is fixed to 16. In this case, the adapters of all blocks from the tenth onwards have an effective rank of at least $0.92\,r$, which means that all available directions are in use, whereas the first block uses about $0.6\,r$. In contrast, with a depth-graded allocation (dashed), utilisation stays well below saturation throughout and leave a head-room at every depth. We have similar observation for DoRA adapted CLIP ViT-L/14 as well as adaptaion of other vision encoders as reported in Appendix.

3.3 Depth-graded rank allocation

Based on our observation in rank probing analysis, we propose *depth-graded rank allocation* which assigns block $\ell \in \{1, \ldots, L\}$ the rank

$$r_\ell = \mathrm{round}\Big(r_1 + (r_L - r_1)\,\frac{\ell - 1}{L - 1}\Big), \tag{6}$$

shared by all adapted projections of that block. The schedule has two hyper-parameters (r_1 and r_L) and requires no importance estimation, pruning or rank search, so training cost is identical to that of uniform adapters with the same total rank. The number of trainable adapter parameters is $|\mathcal{P}| \cdot 2d \cdot \sum_\ell r_\ell$ (for LoRA), plus $|\mathcal{P}| \cdot L \cdot d$ magnitude parameters for DoRA.

When rank varies across blocks, the scaling factor also determines how strongly each block is updated. Standard scaling $\gamma_\ell = \alpha/r_\ell$ shrinks the deepest updates, which is shown (Kalajdzievski, 2023) to result in the gradient-collapse for large ranks (*i.e.,* our deepest blocks). In contrast, with rsLoRA (Kalajdzievski, 2023) scaling $\gamma_\ell = \alpha/\sqrt{r_\ell}$ decreases along depth. Applying this to our setup keeps the magnitude of each block's update comparable while deeper blocks gain directions, since the schedule changes capacity (*i.e.* rank), not step size. Hence, we use $\gamma_\ell = \alpha/\sqrt{r_\ell}$ for rank stablastion for LoRA (as in rsLoRA) and DoRA (so called rsDoRA) adaptation. Combining depth-graded rank allocation in Eq. 6 with rsLoRA in Eq. 1 gives **DG-LoRA**, and combining it with rsDoRA in Eq. 2 gives **DG-DoRA**.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Foundation models. We use 8 publicly available ViT-L image encoders, including CLIP ViT-L/14 (Radford et al., 2021), EVA02-CLIP-L/14 (Sun et al., 2023), DINOv2-L/14 (Oquab et al., 2023), OpenCLIP ViT-L/14 trained on LAION-2B (Cherti et al., 2023; Schuhmann et al., 2022), MetaCLIP ViT-L/14 (400M) (Xu et al., 2024), DFN-CLIP ViT-L/14 (DFN-2B) (Fang et al., 2024), SigLIP-L/16 and SigLIP 2-L/16 at 256^2 pixels (Zhai et al., 2023; Tschannen et al., 2025). For the scaling experiments, we additionally use CLIP ViT-B/16 and ViT-B/32 (Radford et al., 2021).

Training data. We train on CASIA-WebFace (Yi et al., 2014) (0.49M images, 10.6K identities) for all controlled comparisons, and on WebFace4M (Zhu et al., 2021) (4.2M images of 206K identities) for the scaling experiments. We use the standard 112×112 aligned crops, resized to the native resolution of each encoder with bicubic interpolation and normalised with the encoder's statistics.

Optimisation. We optimise the adapters, the projection (when unfrozen) and the classifier head with AdamW (Loshchilov & Hutter, 2019) at learning rate 10^{-4}, weight decay 0.05 on the backbone parameters, gradient clipping at norm 5, and a cosine schedule with warm restarts (Loshchilov & Hutter, 2017) (T_0=5, T_{mult}=2) after five epochs at constant learning rate, for 20 epochs. For ArcFace hyperparamters, we use s=64 and m=0.3. For training on WebFace4M, we implement the identity classifier with Partial FC (An et al., 2022), which updates the positive centers and a random 10% of negative centers per step. The models are trained with a cluster equipped with NVIDIA H100 GPUs.

Evaluation. We report face recognition accuracy on LFW (Huang et al., 2008), CFP-FP (Sengupta et al., 2016), AgeDB-30 (Moschoglou et al., 2017), CALFW (Zheng et al., 2017) and CPLFW (Zheng & Deng, 2018) following their standard 10-fold protocols, and their average (Avg.). We also report the true acceptance rate (TAR) at low false acceptance rates (FAR) on the template-based IJB-B (Whitelam et al., 2017) and IJB-C (Maze et al., 2018) datasets.

4.2 ANALYSIS

Which foundation model should be adapted for FR? Table 1 compares eight ViT-L foundation models with similar backbones adapted with rsLoRA ($r = 16$) on CASIA-WebFace. This table shows that CLIP and EVA02-CLIP are the best performing foundation models for face recognition. Among all models CLIP leads on the benchmark average (94.33) and on IJB-B (77.31), and EVA02-CLIP on IJB-C (82.63). For all models we also finetuned the output projection, except DINOv2, which does not have a projection. DINOv2 has competitive performance on the average (93.62) but $5.53/3.35$ points behind CLIP on IJB-B/C at FAR$= 10^{-4}$. Based on Table 1, the size of the pre-training dataset does not predict how well a model transfers to face recognition. DFN-CLIP is only 2.24 points below CLIP on the average but 38.18 points below on IJB-C. We use CLIP ViT-L/14 in

Table 1: Adaptation of different vision foundation models for face recognition. Every ViT-L encoder is adapted with rsLoRA with r=16 on W_q, W_v of every block with ArcFace head, trained on CASIA-WebFace, with the output projection trained as well. Accuracy on five verification benchmarks and their average; TAR on IJB-B and IJB-C. Bold: best per column. DINOv2 has no output projection and is adapted with adapters only.

		Verification accuracy (%)						TAR (%) @ FAR$= 10^{-4}$	
Backbone	Pre-training data	LFW	CFP-FP	AgeDB	CALFW	CPLFW	Avg.	IJB-B	IJB-C
CLIP ViT-L/14	WIT-400M	**99.47**	**96.14**	91.55	**92.82**	**91.67**	**94.33**	**77.31**	81.95
EVA02-CLIP ViT-L/14	Merged-2B	99.43	96.11	90.02	92.02	91.40	93.80	76.80	**82.63**
DINOv2 ViT-L/14	LVD-142M	99.12	95.44	**92.17**	91.50	89.85	93.62	71.78	78.60
OpenCLIP ViT-L/14	LAION-2B	99.13	94.71	89.53	91.45	89.23	92.81	36.24	45.88
MetaCLIP ViT-L/14	MetaCLIP-400M	99.18	93.90	88.00	91.38	88.57	92.21	41.92	47.73
DFN-CLIP ViT-L/14	DFN-2B	99.07	93.06	88.93	91.35	87.48	91.98	28.41	30.85
SigLIP ViT-L/16	WebLI	98.58	92.79	86.85	90.35	87.20	91.15	71.11	76.27
SigLIP 2 ViT-L/16	WebLI	98.68	92.90	85.10	90.73	87.80	91.04	69.27	75.22

Table 2: Rank budget. CLIP ViT-L/14 on CASIA-WebFace with the output projection trained with a Partial FC head. The first block of the table keeps the total rank, and so the number of trainable parameters (Params), fixed and changes only how it is distributed over depth; the last row spends almost twice that budget on a uniform rank. Bold: best per column.

				Verification accuracy (%)						TAR (%) @ FAR= 10^{-4}	
Allocation	rank	Adapter	Params	LFW	CFP-FP	AgeDB	CALFW	CPLFW	Avg.	IJB-B	IJB-C
Same budget ($\sum_\ell r_\ell = 3264$, 14.2M parameters)											
Uniform	136	rsDoRA	14.20M	99.38	**95.97**	89.00	**91.67**	**91.08**	93.42	76.03	80.80
Depth-Graded	$16 \to 256$	rsDoRA	14.20M	**99.40**	95.79	**89.53**	91.57	90.85	**93.43**	76.76	80.83
Larger budget											
Uniform	256	rsDoRA	26.00M	99.27	95.57	89.47	91.38	90.97	93.33	**77.32**	**81.38**

Table 3: Comparison of our adaption of CLIP ViT-L/14 with DG-LoRA ($r_\ell : 16 \to 256$) and DG-DoRA ($r_\ell : 16 \to 256$) on WebFace4M with training same ViT from scratch and FRoundation (Chettaoui et al., 2025). Bold: best per column.

		Verification accuracy (%)						TAR (%) @ FAR= 10^{-4}	
Method	Adaptation	LFW	CFP-FP	AgeDB	CALFW	CPLFW	Avg.	IJB-B	IJB-C
Train from scratch	Full training	**99.68**	96.47	94.60	94.85	92.63	95.65	92.92	94.79
FRoundation (Chettaoui et al., 2025)	rsLoRA, r=16	99.65	96.50	93.72	94.37	93.73	95.59	90.72	93.40
FMFace (ours)	DG-LoRA, $16 \to 256$	99.65	97.19	95.72	**95.30**	**94.82**	**96.53**	**93.10**	**95.12**
FMFace (ours)	DG-DoRA, $16 \to 256$	99.67	**97.33**	**95.77**	95.17	94.50	96.49	93.00	95.09

the remaining experiments, which also allows a direct comparison with previous FR adaptation in the literature. Table 9 in the Appendix reports the performance of models without finetuning output projection.

Uniform versus depth-graded rank with same budget. To evaluate the performance of adaptation, we consider the output projection frozen and evaluate different adaptation with same budget (same number of adaptation parameters). As can be seen in Table 2, rsDoRA adaptation with uniform rank of $r = 136$ has the same number of parameters as depth-graded ($16 \to 256$) adaptation, but depth-graded adaptation achieves better performance on IJB-B, IJB-C and on average. Comparing with larger budget $r = 256$, depth-graded adaptation has a competitive performance with 1.8 times fewer number of trainable parameters. It performs 1.1 points better on average benchmarks, but is inferior for $0.56/0.55$ TAR points on IJB-B/C at FAR= 10^{-4}.

Comparison with prior work. Table 3 compares our models with FRoundation on WebFace4M for CLIP ViT-L/14. On WebFace4M, our model with DG-LoRA reaches $93.10/95.12$ TAR at FAR= 10^{-4} on IJB-B/C and a benchmark average of 96.53. That is $2.38/1.71$ points TAR at FAR= 10^{-4} on IJB-B/C and 0.94 of average above FRoundation's uniform-LoRA CLIP on the same data. This is also higher than a ViT-L trained from scratch at FAR= 10^{-4} on IJB-B/C ($92.92/94.79$), which DG-LoRA exceeds by 0.88 points on the average while training 4.7% as many parameters. Our model with DG-DoRA also has very close performance to our model trained with DG-LoRA, and achieve superior performance than FRoundation and training from scratch.

4.3 Ablations

Rank-scale coupling. Table 4 evaluates the rank stabilisation under depth grading. With the standard scale α/r_ℓ, the scale of the deepest block is smaller than that of the rank-stabilised $\alpha/\sqrt{r_\ell}$, which can effect training with higher rank in deep blocks. As the results in Table 4 show, the rank-stabilised $\alpha/\sqrt{r_\ell}$ improves TAR at FAR= 10^{-4} by $1.72/1.48$ points on IJB-B/C, at a cost of 0.26 points of benchmark average.

Attention projection In our experiments, we only adapt attention projections of W_q and W_v. Table 5 shows that adapting all four attention projections instead of W_q and W_v nearly doubles the trainable parameters and changes TAR by $-0.06/+0.45$ points on IJB-B/C, so the query and value projections are sufficient.

Table 4: Rank stabilisation under depth grading. DG-DoRA with $r_\ell : 16 \rightarrow 256$ on CASIA-WebFace (CLIP ViT-L/14, adapters only, fully connected ArcFace head), scaling the update of block ℓ by α/r_ℓ or by the rank-stabilised $\alpha/\sqrt{r_\ell}$, with α=16. Bold: better of the two scaling methods.

	Verification accuracy (%)						TAR (%) @ FAR= 10^{-4}	
Scale	LFW	CFP-FP	AgeDB	CALFW	CPLFW	Avg.	IJB-B	IJB-C
α/r_ℓ (standard)†	**99.53**	**95.99**	**91.27**	**92.82**	91.60	**94.24**	77.85	82.42
$\alpha/\sqrt{r_\ell}$ (rank-stabilised)†	99.48	95.81	90.58	92.07	**91.93**	93.98	**79.57**	**83.90**

Table 5: Which attention projections to adapt. DG-DoRA ($r_\ell : 16 \rightarrow 256$) on CASIA-WebFace with CLIP ViT-L/14, the output projection trained and a Partial FC head, adapting the query and value projections of every block or all four attention projections. Bold: better of the two.

		Verification accuracy (%)						TAR (%) @ FAR= 10^{-4}	
Adapted	Params	LFW	CFP-FP	AgeDB	CALFW	CPLFW	Avg.	IJB-B	IJB-C
W_q, W_v	14.20M	**99.40**	**95.79**	**89.53**	91.57	**90.85**	**93.43**	**76.76**	80.83
W_q, W_k, W_v, W_o	27.62M	99.27	95.47	88.73	**91.82**	90.28	93.11	76.70	**81.28**

Encoder size. Table 6 compares two encoder sizes (ViT-B/16 and ViT-L/14) adapted with DG-DoRA on the WebFace4M dataset. With ViT-B/16, whose image encoder has 86.2M parameters, our adaptation reaches $88.98/92.17\%$ TAR at FAR= 10^{-4} on IJB-B/C, where ViT-L/14, with 3.5 times more model parameters, adds $4.02/2.92$ points to TAR at FAR= 10^{-4} on IJB-B/C and 1.11 points to benchmarks average.

Classification head. While training complexity on large datasets such as WebFace4M increases, Partial FC keeps the classifier cheap for the 206K identities of WebFace4M. It is shown that Partial FC may slightly affect the performance (An et al., 2022). Our results in Table 7 also indicate that on CASIA-WebFace, where a fully connected head is affordable, Partial FC costs $1.52/1.54$ TAR points at FAR= 10^{-4} on IJB-B/C and 0.24 points on benchmark average.

Depth grading at scale. Across all our comparisons, the adapter type matters less than the allocation. The results in Table 8 indicate that, when trained on CASIA-WebFace, DG-LoRA leads DG-DoRA by $0.75/1.39$ TAR points at FAR= 10^{-4} on IJB-B/C and 0.24 points on benchmark average. When trained on WebFace4M, difference between DG-LoRA and DG-DoRA reduced to $0.10/0.03$ TAR points at FAR= 10^{-4} on IJB-B/C and 0.04 points on benchmark average.

5 Conclusion

We asked where a foundation model needs capacity when it is adapted to face recognition. A spectral ranking probe of learned adapter updates shows that the rank of the update grows with depth, and uniform rank adapters saturate in exactly the blocks that matter the most. We proposed depth-graded rank allocation which turns this into a schedule, which achieves better performance with same adaptation budget (trainable parameters). We also used rank stabilization to improve adaptation in deeper layers where we allocated higher ranks. On large scale training with WebFace4M,

Table 6: Effect of encoder size. CLIP image encoders of two sizes adapted with DG-DoRA ($r_\ell : 16 \rightarrow 256$) trained on WebFace4M. Params: parameters of the image encoder, excluding the adapters and the text encoder. Bold: better between the two model sizes.

		Verification accuracy (%)						TAR (%) @ FAR= 10^{-4}	
Encoder	Params	LFW	CFP-FP	AgeDB	CALFW	CPLFW	Avg.	IJB-B	IJB-C
ViT-B/16	86.2M	99.48	96.57	93.25	94.48	93.13	95.38	88.98	92.17
ViT-L/14	304.0M	**99.67**	**97.33**	**95.77**	**95.17**	**94.50**	**96.49**	**93.00**	**95.09**

Table 7: Effect of the classification head. DG-DoRA ($r_\ell : 16 \rightarrow 256$) on CASIA-WebFace with CLIP ViT-L/14 and the output projection trained, with a fully connected ArcFace head or with Partial FC (An et al., 2022), which updates the positive class centres and a random 10% of the negative ones at every step. Results use the final checkpoint of each run. Bold: better of the two.

	Verification accuracy (%)						TAR (%) @ FAR= 10^{-4}	
Head	LFW	CFP-FP	AgeDB	CALFW	CPLFW	Avg.	IJB-B	IJB-C
Fully connected	**99.42**	**95.79**	**89.98**	**92.03**	**91.12**	**93.67**	**78.28**	**82.37**
Partial FC, sampling rate 0.1	99.40	**95.79**	89.53	91.57	90.85	93.43	76.76	80.83

Table 8: Depth-graded LoRA against depth-graded DoRA ($r_\ell : 16 \rightarrow 256$). CLIP ViT-L/14 trained on CASIA-WebFace or WebFace4M with Partial FC. Results use the final checkpoint of each run. Bold: Better between DG-LoRA and DG-DoRA on each dataset.

		Verification accuracy (%)						TAR (%) @ FAR= 10^{-4}	
Training data	Method	LFW	CFP-FP	AgeDB	CALFW	CPLFW	Avg.	IJB-B	IJB-C
CASIA-WebFace	DG-LoRA	**99.40**	95.74	89.38	**91.82**	**90.87**	**93.44**	**77.51**	**82.22**
	DG-DoRA	**99.40**	**95.79**	**89.53**	91.57	90.85	93.43	76.76	80.83
WebFace4M	DG-LoRA	99.65	97.19	95.72	**95.30**	**94.82**	**96.53**	**93.10**	**95.12**
	DG-DoRA	**99.67**	**97.33**	**95.77**	95.17	94.50	96.49	93.00	95.09

models adapted with our method outperformed previous adaptation for face recognition and also training from scratch on the same data.

AI USE STATEMENT

In this work, we used AI assistant for writing assistance and for revising the paper, such as finding typography. We stress that the AI assistant did not design the method, choose the experiments, or run the training. All numbers in the paper are produced by our own training and evaluation code on our cluster. The authors verified all claim in the text and reviewed all assistant-written code. We take responsibility for the final content of this work, including all text, claims and artifacts.

ETHICS STATEMENT

In this study, we used only established research datasets (CASIA-WebFace, WebFace4M, and the standard evaluation benchmarks) under their research licenses and did not collect new data. These datasets were crawled from the web and their subjects did not give explicit consent for this use. Adapted foundation models may inherit demographic biases from training data in both the pre-training stage and also our adaptation. We evaluated the biases across different demographic groups in Appendix B.

REPRODUCIBILITY STATEMENT

We release the source code of our experiments upon acceptance of the paper.

REFERENCES

Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, and Tongliang Liu. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4032–4041. IEEE, 2022.

Nadav Benedek and Lior Wolf. Prilora: Pruned and rank-increasing low-rank adaptation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 252–263, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for

contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

Tahar Chettaoui, Naser Damer, and Fadi Boutros. Froundation: Are foundation models ready for face recognition? *Image and Vision Computing*, 156:105453, 2025.

Tahar Chettaoui, Guray Ozgur, Eduarda Caldeira, Arturas Nakvosas, Hatef Otroshi Shahreza, Sébastien Marcel, Rishabh Shukla, Aditya Takkar, Rushil Khullar, Lalak Yadav, et al. Ijcb-afmfr 2026: Competition on adapting foundation models for face recognition using synthetic training data. *arXiv preprint arXiv:2607.24422*, 2026.

Jun Dan, Yang Liu, Haoyu Xie, Jiankang Deng, Haoran Xie, Xuansong Xie, and Baigui Sun. Transface: Calibrating transformer training for face recognition from a data-centric perspective. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20585–20596. IEEE, 2023.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. Sparse low-rank adaptation of pre-trained language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 4133–4145, 2023.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations*, volume 2024, pp. 36221–36237, 2024.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5900–5909. IEEE, 2020.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023.

Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18729–18738. IEEE, 2022.

Vishnuprasadh Kumaravelu, Sunil Gupta, and PK Srijith. Post-optimization adaptive rank allocation for lora. *arXiv preprint arXiv:2604.27796*, 2026.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024a.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 6738–6746. IEEE, 2017.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, et al. Parameter-efficient orthogonal finetuning via butterfly factorization. In *International Conference on Learning Representations*, volume 2024, pp. 38317–38350, 2024b.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pp. 158–165. IEEE, 2018.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 2024.

Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14220–14229. IEEE, 2021.

Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 1997–2005. IEEE, 2017.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Unsal Ozturk, Hatef Otroshi Shahreza, and Sébastien Marcel. Demographic fairness in multimodal llms: A benchmark of gender and ethnicity bias in face verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1184–1193, 2026.

Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2223–2234, 2023.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in neural information processing systems*, 36:79320–79362, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? volume 34, pp. 12116–12128, 2021.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. volume 35, pp. 25278–25294, 2022.

Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–9. IEEE, 2016.

Hatef Otroshi Shahreza and Sébastien Marcel. Foundation models and biometrics: A survey and outlook. *IEEE Transactions on Information Forensics and Security*, 2025.

Hatef Otroshi Shahreza and Sébastien Marcel. Benchmarking multimodal large language models for face recognition. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13557–13561. IEEE, 2026.

Haseeb Ullah Khan Shinwari and Muhammad Usama. Ard-lora: Dynamic rank allocation for parameter-efficient fine-tuning of foundation models with heterogeneous adaptation needs. *IEEE Transactions on Artificial Intelligence*, 2025.

Redwan Sony, Parisa Farmanifard, Arun Ross, and Anil K Jain. Foundation versus domain-specific models: Performance comparison, fusion, and explainability in face recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3715–3725. IEEE, 2025.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Zhonglin Sun, Chen Feng, Ioannis Patras, and Georgios Tzimiropoulos. Lafs: Landmark-based facial self-supervised learning for face recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1639–1649. IEEE, 2024.

Ashutosh Tripathi, Surya Deep Singh, Pranab Sahoo, and Sriparna Saha. Laara: Layer-aware adaptive rank allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2607.19391*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. Dylora: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3274–3287, 2023.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 5265–5274. IEEE, 2018.

Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 692–702. IEEE, 2019.

Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 592–600. IEEE, 2017.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *International Conference on Learning Representations*, volume 2024, pp. 47812–47831, 2024.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? volume 27, 2014.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Chuyan Zhang, Kefan Wang, and Yun Gu. Beyond low-rank tuning: Model prior-guided rank allocation for effective transfer in low-data and large-gap regimes. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 7, 2018.

Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18676–18688. IEEE, 2022.

Yaoyao Zhong and Weihong Deng. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*, 2021.

Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10487–10497. IEEE, 2021.

A Additional Ablations

Output projection. Table 9 compares the performance of adapted foundation models, with and without finetuning output projection. With the projection frozen, OpenCLIP and DFN-CLIP, both trained on 2B image–text pairs, reach 71.23 and 38.30 TAR on IJB-B against 77.59 for CLIP trained on 400M. Training the output projection barely changes CLIP and EVA02-CLIP, by at most 0.34 points at FAR= 10^{-4}, and helps both SigLIP models by 2.8 to 5.5 points. Note that compared to other models, for SigLIP the trainable projection is a full attention-pooling head rather than a single linear map. OpenCLIP, MetaCLIP and DFN-CLIP with the projection trained, lose 9.9 to 35.0 points on IJB-B.

Table 9: Effect of training the output projection. Each foundation model is adapted with rsLoRA with r=16 on W_q, W_v of every block with ArcFace head, trained on CASIA-WebFace. Fine-tuned column indicate the parameters trained, either the adapters alone or the adapters together with the output projection. DINOv2 has no projection and is omitted. Bold: better row of each pair.

		Verification accuracy (%)						TAR (%) @ FAR= 10^{-4}	
Backbone	Fine-tuned	LFW	CFP-FP	AgeDB	CALFW	CPLFW	Avg.	IJB-B	IJB-C
CLIP ViT-L/14	Adapters	99.45	95.59	91.42	92.30	**91.68**	94.09	**77.59**	81.82
	Adapters + projection	**99.47**	**96.14**	**91.55**	**92.82**	91.67	**94.33**	77.31	**81.95**
EVA02-CLIP ViT-L/14	Adapters	**99.60**	**96.30**	**90.73**	**92.48**	**91.72**	**94.17**	**77.06**	82.29
	Adapters + projection	99.43	96.11	90.02	92.02	91.40	93.80	76.80	**82.63**
OpenCLIP ViT-L/14	Adapters	99.07	**94.91**	89.28	**91.95**	**89.27**	**92.90**	**71.23**	**76.31**
	Adapters + projection	**99.13**	94.71	**89.53**	91.45	89.23	92.81	36.24	45.88
MetaCLIP ViT-L/14	Adapters	99.02	**93.91**	**88.82**	91.37	**89.62**	**92.55**	**70.96**	**76.63**
	Adapters + projection	**99.18**	93.90	88.00	**91.38**	88.57	92.21	41.92	47.73
DFN-CLIP ViT-L/14	Adapters	99.03	**93.10**	**89.47**	90.57	87.07	91.85	**38.30**	**43.64**
	Adapters + projection	**99.07**	93.06	88.93	**91.35**	**87.48**	**91.98**	28.41	30.85
SigLIP ViT-L/16	Adapters	**99.05**	**93.66**	**88.97**	**91.57**	**88.32**	**92.31**	66.85	73.44
	Adapters + projection	98.58	92.79	86.85	90.35	87.20	91.15	**71.11**	**76.27**
SigLIP 2 ViT-L/16	Adapters	**99.12**	**94.70**	**90.78**	**91.87**	**89.02**	**93.10**	63.80	70.49
	Adapters + projection	98.68	92.90	85.10	90.73	87.80	91.04	**69.27**	**75.22**

B Demographic bias

Table 10 reports the performance of models trained with WebFace4M (as in Table 3) on RFW (Wang et al., 2019), which contains 6,000 verification pairs for each of four demographic groups, including African, Asian, Caucasian, and Indian. We evaluate the accuracy on each group and their average, the standard deviation across groups (STD) and the skewed error ratio (SER). Our models adapted with DG-LoRA and DG-DoRA are the most accurate on average (94.71 and 94.70, respectively), with 0.9 points above the ViT-L trained from scratch and 3.1 points above FRoundation, and they improve both baselines on every demographic group. They also have the lowest STD (1.67 and 1.61). However, the SER (2.24 and 2.14) is higher than that of both baselines (2.11 and 1.84).

Table 10: Demographic bias on RFW (Wang et al., 2019) of the WebFace4M models of Table 3: verification accuracy (%) on each of the four demographic groups and their mean, the standard deviation across groups (STD) and the skewed error ratio (SER). Lower STD and SER indicate less bias. Bold: best per column (highest accuracy, lowest STD and SER).

		Verification accuracy (%)					Bias	
Method	Adaptation	African	Asian	Caucasian	Indian	Avg.	STD	SER
Train from scratch	Full training	93.37	92.25	96.33	93.27	93.80	1.75	2.11
FRoundation (Chettaoui et al., 2025)	rsLoRA, r=16	90.40	90.28	94.73	90.90	91.57	2.11	**1.84**
FMFace (ours)	DG-LoRA, 16 → 256	**94.25**	93.70	**97.18**	93.70	**94.71**	1.67	2.24
FMFace (ours)	DG-DoRA, 16 → 256	94.03	**93.80**	97.10	**93.85**	94.70	**1.61**	2.14

C ADDITIONAL PROBE ANALYSIS

Probing DoRA. For DoRA, the effective update $\boldsymbol{m} \odot V/\|V\|_{\text{row}} - W_0$ contains a full-rank component from rescaling the rows of W_0, so its spectrum does not reflect the rank of the low-rank directional update. The spectral probe in Section 3.2 is based on LoRA, for which the update is $\gamma_r BA$. For DoRA, instead we measure the directional update $\Delta V_\ell = \gamma_r B_\ell A_\ell$, which is low-rank. Figure 2 shows a similar analysis of rank over depth as Figure 1 for rsDoRA. For W_v, r_{90} grows from 1 in the first block to 99 in the last, and for W_q from 2 to 36. With a uniform rank of 16, the adapters use at least $0.92\,r$ (W_v) and $0.85\,r$ (W_q) from block 7 onwards, while the first block uses $0.16\,r$ and $0.33\,r$; depth-graded ranks keep utilization below $0.8\,r$ in every block.

Probing other foundation models. Figures 3 illustrates the spectral rank probing analysis with rsLoRA on DINOv2 and EVA02-CLIP. For DINOv2, r_{90} grows from 9 in the first block to 103 in the last for W_v and from 11 to 48 for W_q. For EVA02-CLIP, it grows from 2 to 80 and from 7 to 65, but less monotonically: it rises over the first ten blocks, levels off (W_v) or falls (W_q) between blocks 10 and 20, and rises again towards the last blocks.

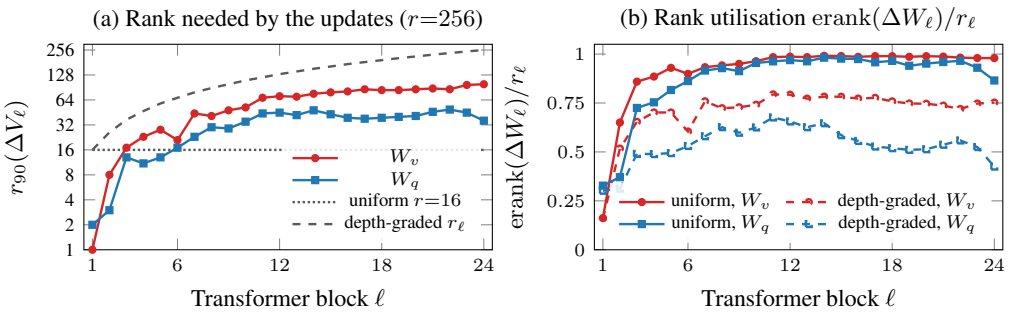

Figure 2: **Rank probe for DoRA.** The spectral rank probing analysis for rsDoRA on CLIP ViT-L/14 and CASIA-WebFace, applied to the directional update $\Delta V_\ell = \gamma_r B_\ell A_\ell$ of Eq. 2. (a) Rank probe with r=256 in every block. (b) Rank utilisation for uniform rank 16 and depth-graded ranks.

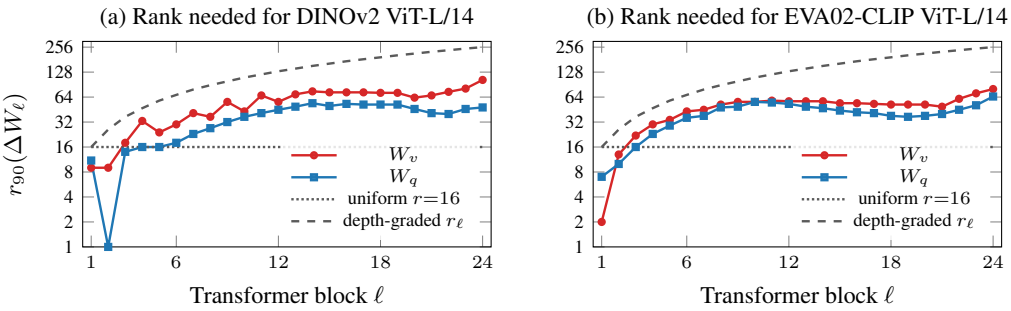

Figure 3: **Rank probe for other backbones.** The spectral rank probing analysis of rsLoRA on CASIA-WebFace with different backbones, probing with r=256 in every block: (a) DINOv2 ViT-L/14 and (b) EVA02-CLIP ViT-L/14.