

---

# Supplementary Material for NeurIPS 2025 Submission #49 On the rankability of visual embeddings

---

Anonymous Author(s)

Affiliation

Address

email

1 **Plan.** This document aims to supplement the main paper as an extension to the appendix. In  
2 Section B.1, we describe the datasets used in our study, and report fine-grained results. In Section B.2,  
3 we compare our results against some popular SOTA methods.

## 4 B.1 Further details on datasets and dataset-specific results

5 In the main paper, we present results for each dataset aggregated over all models (Table 2 in the main  
6 paper) and rankabilities (Spearman  $\rho$ ) for each model-dataset pair (Table 3 in the main paper). While  
7 the main results convey the primary evidence towards our claim (vision embeddings are rankable),  
8 we also report more detailed results on each dataset in this section. Please also refer to Table 1 in the  
9 main paper for a condensed overview of all datasets considered.

### 10 B.1.1 Age

11 **UTKFace**, introduced in [14], is a dataset of face images with age labels ranging from 0 to 116.  
12 Following [12, 5], we use a smaller subset with ages ranging between 21 and 60. The dataset was  
13 downloaded from the official website (<https://susanqq.github.io/UTKFace/>). We report results in  
14 Table A1.

Table A1: **UTKFace (Age)**. Spearman’s rank correlation  $\rho$  across evaluation strategies. **No-train**:  
linear probe on untrained encoder. **Rankability**: linear probe on pretrained encoder. **Nonlinear**:  
MLP on frozen encoder. **Finetuned**: encoder + head trained end-to-end. Higher is better.

Model	No-train lower bound	Rankability main	Nonlinear upper bound	Finetuned upper bound
ResNet-50	0.212	0.633	0.636	0.762
ViT-B/32	0.283	0.739	0.749	0.737
ConvNeXtV2-L	0.283	0.772	0.776	0.815
DINOv2 ViT-B/14	0.054	0.770	0.770	0.812
OpenAI CLIP ResNet-50	0.130	0.820	0.830	0.790
OpenAI CLIP ViT-B/32	0.250	0.810	0.830	0.830
OpenCLIP ConvNeXt-L (D, 320px)	0.180	0.820	0.840	0.850
Mean	0.199	0.766	0.776	0.799

15 **Adience**, introduced in [2], is another age dataset. Unlike UTKFace, it contains coarse labels  
16 (8 age groups instead of exact ages). We use the “aligned” version of the images and five-fold  
17 cross-validation as in [12]. Results can be found in Table A2.

Table A2: **Adience (Age)**. Spearman’s rank correlation  $\rho$  across evaluation strategies.

Model	No-train lower bound	Rankability main	Nonlinear upper bound	Finetuned upper bound
ResNet-50	0.328	0.723	0.759	0.894
ViT-B/32	0.310	0.828	0.860	0.892
ConvNeXtV2-L	0.522	0.871	0.885	0.910
DINOv2 ViT-B/14	0.120	0.853	0.877	0.914
OpenAI CLIP ResNet-50	0.070	0.898	0.914	0.894
OpenAI CLIP ViT-B/32	0.292	0.924	0.922	0.928
OpenCLIP ConvNeXt-L (D, 320px)	0.220	0.928	0.932	0.938
<b>Mean</b>	0.266	0.861	0.878	0.910

18 **B.1.2 Crowd count**

19 **UCF-QNRF**, introduced in [1], is a large crowd counting dataset containing images from diverse  
20 parts of the world. We use the official download link at <https://www.crcv.ucf.edu/data/ucf-qnrf/> and  
21 the official train-test splits. Results can be found in Table A3.

Table A3: **UCF-QNRF (Crowd Count)**. Spearman’s rank correlation  $\rho$  across evaluation strategies.

Model	No-train lower bound	Rankability main	Nonlinear upper bound	Finetuned upper bound
ResNet-50	0.466	0.864	0.870	0.826
ViT-B/32	0.288	0.837	0.840	0.794
ConvNeXtV2-L	−0.054	0.810	0.816	0.938
DINOv2 ViT-B/14	0.219	0.788	0.842	0.961
OpenAI CLIP ResNet-50	0.240	0.870	0.880	0.820
OpenAI CLIP ViT-B/32	0.280	0.870	0.870	0.900
OpenCLIP ConvNeXt-L (D, 320px)	0.100	0.860	0.860	0.960
<b>Mean</b>	0.220	0.843	0.854	0.886

22 **ShanghaiTech**, introduced in [13], is another crowd counting dataset consisting of two parts: A and  
23 B. While part A was crawled from the Internet and features larger crowds in general, part B was  
24 taken from metropolitan areas of Shanghai and features much smaller crowds. We use the DropBox  
25 link available at <https://github.com/desenzhou/ShanghaiTechDataset> and official train-test splits for  
26 both parts. Results can be found in Table A4 and Table A5.

Table A4: **ShanghaiTech-A (Crowd Count)**. Spearman’s rank correlation  $\rho$  across evaluation strategies.

Model	No-train lower bound	Rankability main	Nonlinear upper bound	Finetuned upper bound
ResNet-50	0.359	0.799	0.802	0.529
ViT-B/32	0.148	0.700	0.623	0.558
ConvNeXtV2-L	0.061	0.695	0.753	0.834
DINOv2 ViT-B/14	−0.017	0.653	0.722	0.786
OpenAI CLIP ResNet-50	0.200	0.760	0.770	0.510
OpenAI CLIP ViT-B/32	0.010	0.750	0.770	0.700
OpenCLIP ConvNeXt-L (D, 320px)	0.080	0.780	0.800	0.910
<b>Mean</b>	0.120	0.734	0.749	0.689

Table A5: **ShanghaiTech-B (Crowd Count)**. Spearman’s rank correlation  $\rho$  across evaluation strategies.

Model	No-train lower bound	Rankability main	Nonlinear upper bound	Finetuned upper bound
ResNet-50	0.280	0.879	0.906	0.672
ViT-B/32	0.225	0.878	0.889	0.710
ConvNeXtV2-L	0.070	0.867	0.876	0.955
DINOv2 ViT-B/14	0.020	0.821	0.869	0.972
OpenAI CLIP ResNet-50	0.270	0.890	0.900	0.690
OpenAI CLIP ViT-B/32	0.040	0.860	0.860	0.900
OpenCLIP ConvNeXt-L (D, 320px)	0.040	0.890	0.910	0.980
<b>Mean</b>	0.135	0.869	0.887	0.840

### 27 B.1.3 Headpose (Euler angles)

28 The **BIWI Kinect** dataset, introduced in [3], is a collection of 24 different videos wherein the subject  
29 of the video sits about a meter away from a Kinect (<https://en.wikipedia.org/wiki/Kinect>) sensor  
30 and rotates their head to span the entire range of possible head-pose angles pitch (rotation about  
31 the x-axis), yaw (rotation about the y-axis) and roll (rotation about the z-axis). As there exists no  
32 official split that we know of, we randomly hold out 6 sequences for testing. Results can be found in  
33 Table A6 (pitch), Table A7 (yaw) and Table A8 (roll).

Table A6: **Kinect (Pitch)**. Spearman’s rank correlation  $\rho$  across evaluation strategies.

Model	No-train lower bound	Rankability main	Nonlinear upper bound	Finetuned upper bound
ResNet-50	0.359	0.663	0.615	0.973
ViT-B/32	0.401	0.673	0.505	0.951
ConvNeXtV2-L	0.548	0.909	0.882	0.984
DINOv2 ViT-B/14	0.231	0.716	0.986	0.979
OpenAI CLIP ResNet-50	0.450	0.860	0.870	0.970
OpenAI CLIP ViT-B/32	0.400	0.920	0.940	0.980
OpenCLIP ConvNeXt-L (D, 320px)	0.450	0.880	0.880	0.990
<b>Mean</b>	0.405	0.803	0.811	0.975

Table A7: **Kinect (Yaw)**. Spearman’s rank correlation  $\rho$  across evaluation strategies.

Model	No-train lower bound	Rankability main	Nonlinear upper bound	Finetuned upper bound
ResNet-50	0.160	0.624	0.726	0.990
ViT-B/32	0.209	0.305	0.305	0.838
ConvNeXtV2-L	0.113	0.384	0.716	0.989
DINOv2 ViT-B/14	−0.046	0.804	0.871	0.994
OpenAI CLIP ResNet-50	−0.060	0.120	0.530	0.980
OpenAI CLIP ViT-B/32	0.160	0.360	0.330	0.990
OpenCLIP ConvNeXt-L (D, 320px)	0.010	0.440	0.700	0.990
<b>Mean</b>	0.078	0.434	0.597	0.967

Table A8: **Kinect (Roll)**. Spearman’s rank correlation  $\rho$  across evaluation strategies.

Model	No-train lower bound	Rankability main	Nonlinear upper bound	Finetuned upper bound
ResNet-50	0.202	0.352	0.430	0.930
ViT-B/32	0.098	0.196	0.375	0.477
ConvNeXtV2-L	0.550	0.298	0.368	0.963
DINOv2 ViT-B/14	0.256	0.512	0.551	0.912
OpenAI CLIP ResNet-50	0.140	0.090	0.300	0.920
OpenAI CLIP ViT-B/32	−0.060	0.020	0.170	0.850
OpenCLIP ConvNeXt-L (D, 320px)	−0.130	0.060	0.090	0.960
<b>Mean</b>	0.151	0.218	0.326	0.859

#### 34 B.1.4 Aesthetics (Mean Opinion Score)

35 The **Aesthetics Visual Analysis (AVA)** dataset, introduced in [9], is a large-scale dataset includ-  
36 ing aesthetic preference scores provided by human annotators. Each image is labeled by mul-  
37 tiple annotators, each assigning a score in the range 1-10. The mean opinion score (MOS) of  
38 the image is then computed as a weighted average over the ratings where the weight of a rat-  
39 ing is provided by its frequency. We use the split provided by [12] in their official repository  
40 (<https://github.com/uynaes/RankingAwareCLIP/tree/main/examples>). Results are reported in Ta-  
41 ble A9.

Table A9: **AVA (Image Aesthetics)**. Spearman’s rank correlation  $\rho$  across evaluation strategies.

Model	No-train lower bound	Rankability main	Nonlinear upper bound	Finetuned upper bound
ResNet-50	0.237	0.589	0.628	0.672
ViT-B/32	0.157	0.609	0.666	0.672
ConvNeXtV2-L	0.158	0.644	0.685	0.728
DINOv2 ViT-B/14	0.057	0.566	0.648	0.590
OpenAI CLIP ResNet-50	0.150	0.700	0.710	0.700
OpenAI CLIP ViT-B/32	0.200	0.710	0.730	0.700
OpenCLIP ConvNeXt-L (D, 320px)	0.130	0.750	0.780	0.790
<b>Mean</b>	0.156	0.653	0.692	0.693

42 **KonIQ-10k**, introduced in [4], is another aesthetics or image quality assessment (IQA) dataset  
43 that aims to model naturally occurring image distortions with mean opinion scores ranging roughly  
44 between 1 and 100. We use the official train-test splits. Results can be found in Table A10.

Table A10: **KoniQ-10k (Image Aesthetics)**. Spearman’s rank correlation  $\rho$  across evaluation strategies.

Model	No-train lower bound	Rankability main	Nonlinear upper bound	Finetuned upper bound
ResNet-50	0.563	0.739	0.739	0.874
ViT-B/32	0.488	0.713	0.753	0.813
ConvNeXtV2-L	0.487	0.744	0.765	0.930
DINOv2 ViT-B/14	0.324	0.681	0.753	0.948
OpenAI CLIP ResNet-50	0.400	0.800	0.840	0.900
OpenAI CLIP ViT-B/32	0.460	0.790	0.830	0.890
OpenCLIP ConvNeXt-L (D, 320px)	0.320	0.860	0.870	0.950
<b>Mean</b>	0.435	0.761	0.793	0.901

### 45 B.1.5 Image recency

46 **Historical Color Images (HCI)**, introduced in [10], was designed for the task of classifying an  
 47 image by the decade during which it was taken. Therein emerges a natural ordering over the decades,  
 48 defining the ordinal attribute of image “modernness” or “recency”. We use the split provided by [12]  
 49 in their repository (<https://github.com/uynaes/RankingAwareCLIP/tree/main/examples>) and report  
 50 the results in Table A11.

Table A11: **HCI (Historical Color Images)**. Spearman’s rank correlation  $\rho$  across evaluation strategies.

Model	No-train lower bound	Rankability main	Nonlinear upper bound	Finetuned upper bound
ResNet-50	0.351	0.600	0.592	0.614
ViT-B/32	0.377	0.592	0.618	0.529
ConvNeXtV2-L	0.362	0.631	0.663	0.771
DINOv2 ViT-B/14	0.131	0.571	0.601	0.748
OpenAI CLIP ResNet-50	0.320	0.780	0.770	0.760
OpenAI CLIP ViT-B/32	0.430	0.770	0.780	0.760
OpenCLIP ConvNeXt-L (D, 320px)	0.300	0.820	0.790	0.870
<b>Mean</b>	0.324	0.680	0.688	0.722

## 51 B.2 Comparison with SOTA

52 Prior research has presented results from dedicated or general efforts to solve the datasets considered  
 53 in our study. Our main aim is to understand the rankability emerging out of the structure in visual  
 54 embedding spaces, and we contextualize our numbers using reference metrics (lower bound provided  
 55 by the no-encoder baseline and upper bound provided by nonlinear regression and finetuned encoders).  
 56 However, in Table A12, we also provide comparisons with state-of-the-art results to further contextu-  
 57 alize our results. Some comparisons suggest that simple linear regression over pretrained embeddings  
 58 often performs comparably with or even surpasses dedicated efforts. Although architectural and  
 59 training dataset differences mean that this comparison is not always fair, we emphasize the contrast  
 60 in implementational simplicity between dedicated efforts and simple linear regression over pretrained  
 61 embeddings that are often readily available and easy to use.

Table A12: **Comparing linear / nonlinear regression against recent state-of-the-art methods.** “Linear” and “Nonlinear” use regression over CLIP-ConvNeXt-L embeddings. Dashes indicate metrics unreported in prior work. We take the numbers for age, aesthetics and recency from [12], and crowd count from [8]. Under “Downstream model”, we report the components used on top of pretrained visual embeddings (CLIP or non-CLIP models); sometimes, we also report if the encoder itself was retrained. Under “Downstream data”, we report the additional data used for training the method. Finally, under “Other ingredients”, we also report miscellaneous extra components used by the corresponding method. All method interpretations are to the best of our knowledge.

Attribute	Dataset	Method	Spearman $\rho$	MAE	Downstream model	Downstream data	Other ingredients
Age	UTKFace	Yu et al. [12]	–	3.83	Cross-attn encoder, two ranking heads, learnable text prompt tokens	Images with age labels	Text encoder
		MiVOLO [5]	–	4.23	Regression heads	Body images, face patches, age labels	Feature enhancer module for fused joint representations
		Linear	–	4.25	Linear regressor	Images with age labels	None
		Nonlinear	–	4.10	2-layer MLP regressor	Images with age labels	None
Age	Adience	Yu et al. [12]	–	0.36 (0.03)	Cross-attn encoder, two ranking heads, learnable text prompt tokens	Images with age labels	Text encoder
		OrdinalCLIP [6]	–	0.47 (0.06)	(Retrain image encoder for the task)	Images with age labels	Text encoder; learn “continuous” rank prototype (text) embeddings for each rank
		Linear regressor	–	0.48 (0.02)	Linear	Images with age labels	None
		Nonlinear	–	0.45 (0.02)	2-layer MLP regressor	Images with age labels	None
Crowd Count	UCF-QNRF	CLIP-EBC [8]	–	80.3	Blockwise classification module	Images with count labels	Text encoder
		CrowdCLIP [7]	–	283.3	(Retrain image encoder)	Crowd images	Text encoder, three-stage progressive filtering during inference
		Linear	–	246.4	Linear	Images with count labels	None
		Nonlinear	–	248.0	2-layer MLP	Images with count labels	None
Crowd Count	ST-A	CLIP-EBC [8]	–	52.5	Blockwise classification module	Images with count labels	Text encoder
		CrowdCLIP [7]	–	146.1	(Retrain image encoder)	Crowd images	Text encoder, three-stage progressive filtering during inference
		Linear	–	167.1	Linear	Images with count labels	None
		Nonlinear	–	151.7	2-layer MLP	Images with count labels	None
Crowd Count	ST-B	CLIP-EBC [8]	–	6.6	Blockwise classification module	Images with count labels	Text encoder
		CrowdCLIP [7]	–	69.3	(Retrain image encoder)	Crowd images	Text encoder, three-stage progressive filtering during inference

*Continued on next page*

Table A12 – Continued from previous page

Attribute	Dataset	Method	Spearman $\rho$	MAE	Downstream model	Downstream data	Other ingredients
		Linear	–	34.7	Linear	Images with count labels	None
		Nonlinear	–	29.7	2-layer MLP	Images with count labels	None
Aesthetics	AVA*	Yu et al. [12]	0.747	–	Cross-attn encoder, two ranking heads, learnable text prompt tokens	Images with MOS labels (1–10)	Text encoder
		CLIP-IQA [11]	0.415	–	Softmax over two similarity scores	None	Text encoder, prompt engineering, remove position embedding
		Linear	0.749	–	Linear	Images with MOS labels(1–10)	None
		Nonlinear	0.775	–	2-layer MLP	Images with MOS labels (1–10)	None
Aesthetics	KonIQ-10k	Yu et al. [12]	0.911	–	Cross-attn encoder, two ranking heads, learnable text prompt tokens	Images with MOS labels (1–100)	Text encoder
		CLIP-IQA [11]	0.727	–	Softmax over two similarity scores	Images with MOS labels (1–100)	Text encoder, prompt engineering, remove position embedding
		Linear	0.860	–	Linear	Images with MOS labelss (1–100)	None
		Nonlinear	0.870	–	2-layer MLP	Images with MOS labels (1–100)	None
Recency	HCI*	Yu et al. [12]	–	0.32 (0.03)	Cross-attn encoder, two ranking heads, learnable text prompt tokens	Images with decade labels	Text encoder
		OrdinalCLIP [6]	–	0.67 (0.03)	(Retrain image encoder for the task)	Images with decade labels	Text encoder; learn “continuous” rank prototype (text) embeddings for each rank
		Linear	–	0.64	Linear	Images with decade labels	None
		Nonlinear	–	0.60	2-layer MLP	Images with decade labels	None

## References

- [1] “Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds”. In: Haroon Idrees et al. *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2018, pp. 544–559. ISBN: 978-3-030-01215-1 978-3-030-01216-8. DOI: 10.1007/978-3-030-01216-8\_33. URL: [https://link.springer.com/10.1007/978-3-030-01216-8\\_33](https://link.springer.com/10.1007/978-3-030-01216-8_33).
- [2] Eran Eidinger, Roeen Enbar, and Tal Hassner. “Age and Gender Estimation of Unfiltered Faces”. In: *IEEE Transactions on Information Forensics and Security* 9.12 (Dec. 2014), pp. 2170–2179. ISSN: 1556-6021. DOI: 10.1109/TIFS.2014.2359646. URL: <https://ieeexplore.ieee.org/document/6906255>.
- [3] Gabriele Fanelli et al. “Real Time Head Pose Estimation from Consumer Depth Cameras”. In: *Pattern Recognition*. Ed. by Rudolf Mester and Michael Felsberg. Red. by David Hutchison et al. Vol. 6835. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 101–110. ISBN: 978-3-642-23122-3 978-3-642-23123-0. DOI: 10.1007/978-3-642-23123-0\_11. URL: [http://link.springer.com/10.1007/978-3-642-23123-0\\_11](http://link.springer.com/10.1007/978-3-642-23123-0_11).
- [4] Vlad Hosu et al. “KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 4041–4056. ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2020.2967829. arXiv: 1910.06180 [cs]. URL: <http://arxiv.org/abs/1910.06180>.
- [5] Maksim Kuprashevich and Irina Tolstykh. *MiVOLO: Multi-input Transformer for Age and Gender Estimation*. Sept. 22, 2023. DOI: 10.48550/arXiv.2307.04616. arXiv: 2307.04616 [cs]. URL: <http://arxiv.org/abs/2307.04616>. Pre-published.
- [6] Wanhua Li et al. *OrdinalCLIP: Learning Rank Prompts for Language-Guided Ordinal Regression*. Oct. 1, 2022. DOI: 10.48550/arXiv.2206.02338. arXiv: 2206.02338 [cs]. URL: <http://arxiv.org/abs/2206.02338>. Pre-published.
- [7] Dingkan Liang et al. *CrowdCLIP: Unsupervised Crowd Counting via Vision-Language Model*. Apr. 9, 2023. DOI: 10.48550/arXiv.2304.04231. arXiv: 2304.04231 [cs]. URL: <http://arxiv.org/abs/2304.04231>. Pre-published.
- [8] Yiming Ma, Victor Sanchez, and Tanaya Guha. *CLIP-EBC: CLIP Can Count Accurately through Enhanced Blockwise Classification*. Version 3. Mar. 25, 2025. DOI: 10.48550/arXiv.2403.09281. arXiv: 2403.09281 [cs]. URL: <http://arxiv.org/abs/2403.09281>. Pre-published.
- [9] N. Murray, L. Marchesotti, and F. Perronnin. “AVA: A Large-Scale Database for Aesthetic Visual Analysis”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI: IEEE, June 2012. DOI: 10.1109/cvpr.2012.6247954. URL: <http://ieeexplore.ieee.org/document/6247954/>.
- [10] Frank Palermo, James Hays, and Alexei A. Efros. “Dating Historical Color Images”. In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon et al. Red. by David Hutchison et al. Vol. 7577. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 499–512. ISBN: 978-3-642-33782-6 978-3-642-33783-3. DOI: 10.1007/978-3-642-33783-3\_36. URL: [http://link.springer.com/10.1007/978-3-642-33783-3\\_36](http://link.springer.com/10.1007/978-3-642-33783-3_36).
- [11] Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. “Exploring CLIP for Assessing the Look and Feel of Images”. In: *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. Vol. 37. AAAI’23/IAAI’23/EAAI’23. AAAI Press, Feb. 7, 2023, pp. 2555–2563. ISBN: 978-1-57735-880-0. DOI: 10.1609/aaai.v37i2.25353. URL: <https://doi.org/10.1609/aaai.v37i2.25353>.
- [12] Wei-Hsiang Yu et al. “RANKING-AWARE ADAPTER FOR TEXT-DRIVEN IMAGE ORDERING WITH CLIP”. In: *ICLR* (2025).
- [13] Yingying Zhang et al. “Single-Image Crowd Counting via Multi-Column Convolutional Neural Network”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, June 2016, pp. 589–597. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.70. URL: <http://ieeexplore.ieee.org/document/7780439/>.



118 [14] Zhifei Zhang, Yang Song, and Hairong Qi. *Age Progression/Regression by Conditional*  
119 *Adversarial Autoencoder*. Mar. 28, 2017. DOI: 10.48550/arXiv.1702.08423. arXiv:  
120 1702.08423 [cs]. URL: <http://arxiv.org/abs/1702.08423>. Pre-published.