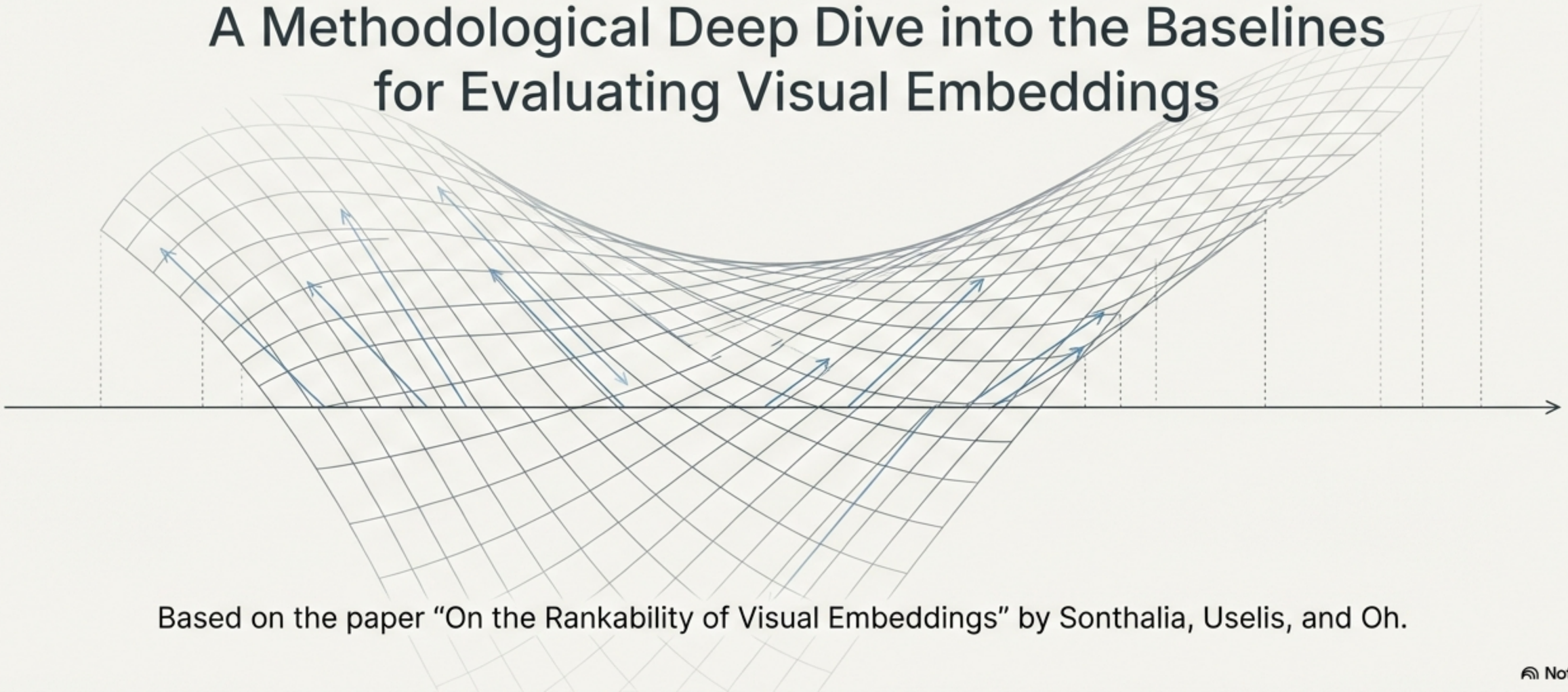
Contextualizing Rankability

A Methodological Deep Dive into the Baselines for Evaluating Visual Embeddings

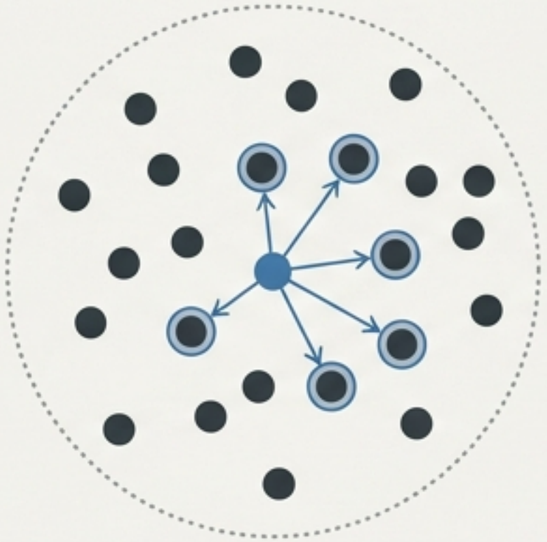


Based on the paper “On the Rankability of Visual Embeddings” by Sonthalia, Uselis, and Oh.

Beyond Retrieval: Are Visual Embeddings Rankable?

Traditional Retrieval (Local Similarity)

Finds images that are “nearby” a query in the embedding space. Relies on local similarity search.



A New Demand (Global Ordering)

Sorts an entire collection along a specific, continuous attribute (e.g., age, formality, aesthetics). This requires a global ordering.

What is Rankability?

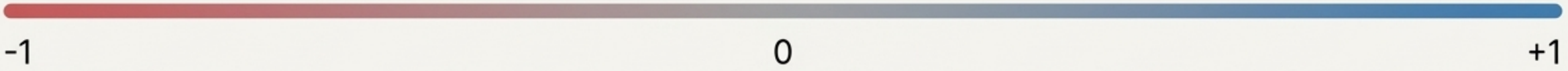
A representation f is rankable for an ordinal attribute A if there exists a **rank axis $\boldsymbol{v}_A$** such that for any images $\boldsymbol{x}_1, \boldsymbol{x}_2$ with $A(\boldsymbol{x}_1) \geq A(\boldsymbol{x}_2)$, it follows that the projection $\boldsymbol{v}^\top_A \boldsymbol{f}(\boldsymbol{x}_1) \geq \boldsymbol{v}^\top_A \boldsymbol{f}(\boldsymbol{x}_2)$.

The central question is not just *if* embeddings are rankable, but *how much*.

We measure rankability using **Spearman's Rank Correlation Coefficient (SRCC)**, denoted as ρ.

SRCC calculates the correlation between the predicted ranking (from projecting embeddings onto a learned rank axis) and the true attribute ranking.

The score ρ ranges from -1 (perfectly opposite order) to +1 (perfectly matched order), with 0 indicating no correlation.



> We trained a simple linear regressor and got an SRCC score of **0.86** for age on the Adience dataset. Is that good?
>
> A single score is meaningless in isolation. To understand its significance, we need context. We need reference points.

To Interpret the Score, We Build a Framework of Baselines

We establish three critical reference points to understand the true performance of our main linear regressor.

3. THE SKY: Finetuning Upper Bound

Question: What is the absolute best this model *architecture* could possibly achieve on this task?

Analogy: The theoretical maximum for this model family.

2. THE ROOM'S CEILING: Nonlinear Upper Bound

Question: What is the maximum information *already contained* within the existing embedding space?

Analogy: The limit of what's possible with the current embeddings.

1. THE FLOOR: No-Train Lower Bound

Question: What is the performance from a model with no prior knowledge?

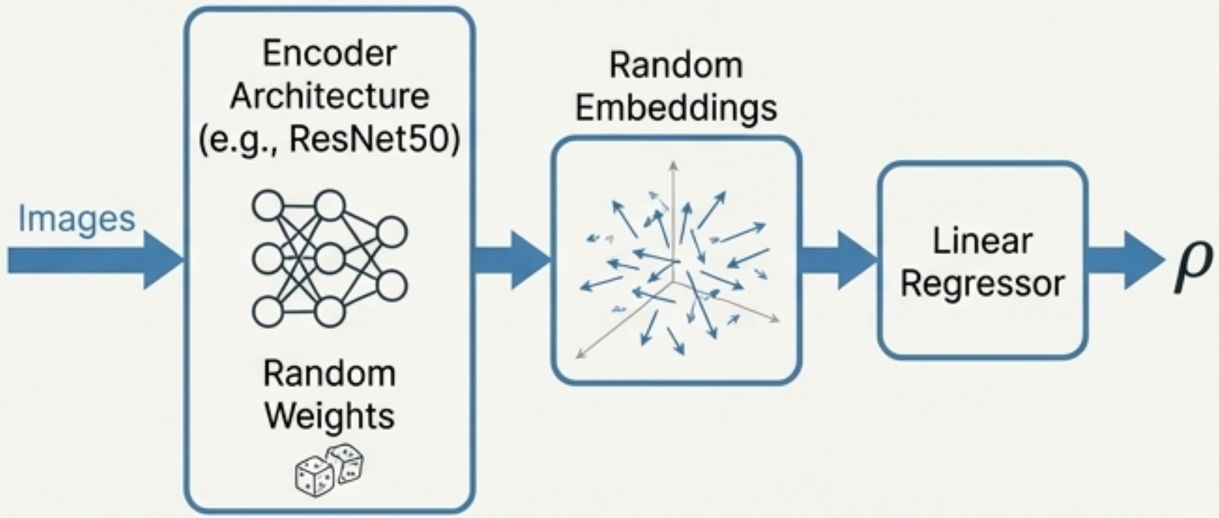
Analogy: The absolute minimum, the starting point.

Baseline 1: The Floor (No-Train Lower Bound)

Purpose: To establish a 'no-information' baseline and measure the performance attributable purely to random chance and architectural priors.

How It's Generated

1. **Start with the Encoder Architecture**: Take the same visual encoder (e.g., ResNet50, ViT-B/32).
2. **Random Initialization**: Instead of using pre-trained weights, initialize the encoder with random weights. It has not been trained on any images.
3. **Generate 'Random' Embeddings**: Pass the dataset images through this untrained encoder. The output vectors are effectively random.
4. **Train Linear Regressor**: Train the same simple linear regressor on these random embeddings to find the optimal rank axis v_A.



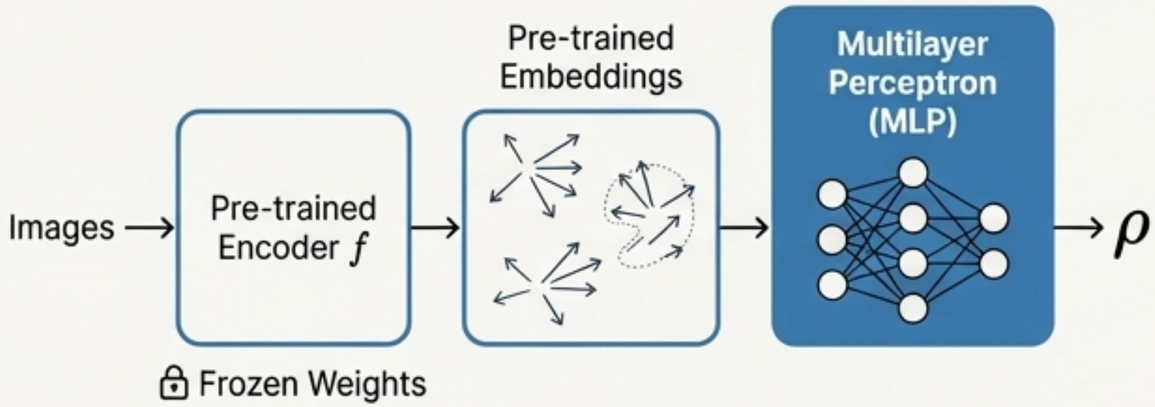
Key Insight: The resulting SRCC score is not zero. Even randomly initialized encoders produce a small positive correlation. This non-zero value serves as the true "floor" for our comparison.

Baseline 2: The Room's Ceiling (Nonlinear Upper Bound)

Purpose: To estimate the *total amount of ordinal information* present in the pre-trained embedding space, including complex, non-linear relationships.

How It's Generated

1. **Start with Pre-trained Embeddings:** Use the embeddings generated by the standard, fully-trained visual encoder f.
2. **Replace the Linear Regressor:** Instead of a simple linear regressor, use a more powerful model: a **two-layer Multilayer Perceptron (MLP)**.
3. **Train the MLP:** Train the MLP to predict the attribute values from the frozen, pre-trained embeddings.



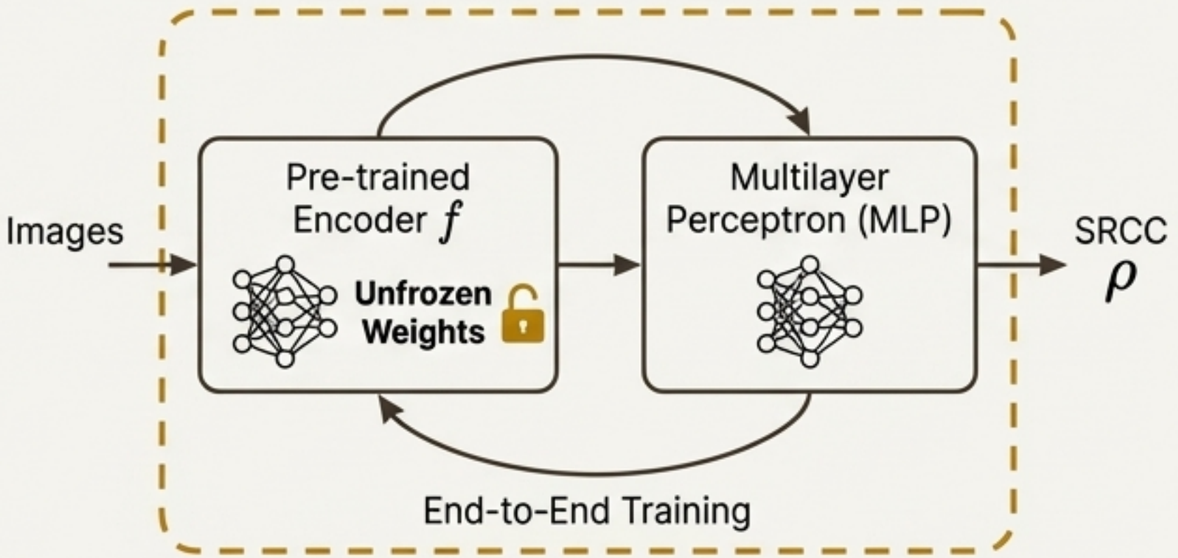
Key Insight: A two-layer MLP is a known **universal approximator**. It can, in theory, extract all available ranking information from the embeddings. This score tells us the maximum performance possible *without changing the embeddings themselves*.

Baseline 3: The Sky (Finetuning Upper Bound)

Inter : To measure the broader upper bound of the encoder *architecture's capacity* for the specific attribute, assuming we can modify the embeddings.

How It's Generated

1. **Unfreeze the Encoder**: Start with the pre-trained visual encoder and the two-layer MLP.
2. **End-to-End Training**: Train the entire system jointly. Both the visual encoder's weights and the MLP's weights are updated during training.
3. **Adapt the Model**: The embedding model f is fine-tuned to become better at representing the specific attribute, pushing performance to its architectural limit.



Key Insight: This represents the "best effort" a given model architecture can make on the task. It conceptually envelops the nonlinear regression upper bound and serves as the ultimate performance ceiling.

The Results: Visual Embeddings are Highly Rankable

Attribute (Dataset)	No-train lower bound	**Rankability (main)**	Nonlinear upper bound	Finetuned upper bound
Age (UTKFace)	0.199	**0.766**	0.776	0.799
Age (Adience)	0.266	**0.861**	0.878	0.910
Crowd (UCF-QNRF)	0.220	**0.843**	0.854	0.886
Aesthetics (AVA)	0.156	**0.653**	0.692	0.693
Recency (HCI)	0.324	**0.680**	0.688	0.722

For most attributes, the rankability of a simple linear probe is much closer to the upper bounds than to the lower bound.

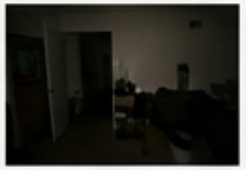
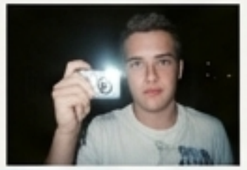
Deeper Dive: CLIP-Based Embeddings are Generally More Rankable

Attribute	RN50	ViTB32	...	CLIP-RN50	CLIP-ViTB32	CLIP-CNX
Age (Adience)	0.723	0.828	...	**0.898**	**0.924**	**0.928**
Aesthetics (AVA)	0.589	0.609	...	**0.700**	**0.710**	**0.750**
Recency (HCI)	0.600	0.592	...	**0.780**	**0.770**	**0.820**

- On attributes like age, aesthetics, and recency, CLIP encoders consistently outperform their non-CLIP counterparts.
- On some tasks like crowd count, performance is comparable.
- There are interesting exceptions: DINOv2 significantly outperforms CLIP on yaw and roll head pose angles, suggesting architectural differences matter.

> Vision-language pre-training appears to encourage the formation of more linearly separable, rankable structures for many semantic attributes.

What Rankability Looks Like in Practice

Age (UTKFace)					
	(young) 0th	25th	50th	75th	100th (old)
Crowd Count (UCF-QNRF)					
	(not crowded) 0th	25th	50th	75th	100th (crowded)
Aesthetics (KonIQ-10k)					
	(bad) 0th	25th	50th	75th	100th (good)
Head Pose (Kinect)					
	(looking down) 0th	25th	50th	75th	100th (looking up)

Images sorted by projecting their CLIP-ViT-B/32 embeddings onto a rank axis found via simple linear regression. The visual results confirm the high quantitative scores.

The Big Picture: Rankable Structure is an Unexpected and Useful Property

Modern visual embeddings are highly rankable. **Most of the ordinal information is linearly encoded**, as shown by the small gap between linear and nonlinear probes.

Practical Utility

Enables efficient ranking and sorting in vector databases using simple, fast operations. For example, a photo app could sort selfies by "apparent age" using a rank axis defined by just two examples (a child and an elderly person).

Interpretability

The existence of linear rank axes suggests we can interpret the embedding space as a collection of latent ordinal subspaces. This is a step towards making these high-dimensional "latent spaces" more understandable.

The embedding space is not just a clustered collection of points for retrieval; it possesses a rich, ordered structure that we can now rigorously measure and exploit.