
Learning to See Beauty: Personalized Ranking from Portrait Comparisons

Chicken Angel

Independent research project
chicken-angel.com

Abstract

A portrait comparison expresses a preference for people as they appear in particular photographs. We study whether a compact personal scorer can preserve those choices while becoming less sensitive to the photograph. Our estimator fits a linear Bradley–Terry score to two-portrait comparisons and penalizes score variation between training photographs of the same person. We derive an exact photograph-substitution interpretation and a logistic-loss bound. A prospective single-observer study uses 52 photographs, 24 personalization comparisons, 12 held-out comparisons, and four reversed repeats. Tongdak Cheonsa is the pre-specified target, excluded from personalization. With the primary frozen encoder, consistency reduces within-person test-score variance from 0.670 to 0.404, but matches only 3/11 decisive held-out choices, versus 2/11 for the matched ablation and 9/11 for an identity-only control. The observer selects the target in all four scheduled comparisons; the primary model matches none and ranks her 12th of thirteen panels. Different framing and encoder choices substantially alter the prediction. The results separate optimization of photograph consistency from successful preference transfer: the proposed penalty controls its fitted objective, but this case does not establish a reliable personal beauty ranking.

1 Introduction

What would it mean for a system to learn that one person finds a particular face beautiful? A generic image score can reproduce a broad visual prior, yet miss the individual’s preference. A personal model can fit a handful of choices, yet learn lighting, expression, hairstyle, or the event at which a photograph was taken. These are separate difficulties: agreement with an observer and stability under a change of photograph.

We study them through a concrete target case. The participant compares portrait panels of twelve actresses. A shared score is learned from these choices and then applied to different photographs, including Tongdak Cheonsa, whose photographs and preference labels are entirely excluded from personalization. The central question is whether choices among other people support a transferable preference for this target, and how strongly that answer depends on the photographs used to represent her. The target is therefore part of the prediction problem, rather than a label inserted into the training objective.

The proposed estimator pairs a small preference head with an explicit consistency prior. Each candidate is represented by the mean features of two photographs. In addition to matching observed pairwise choices, the model penalizes different scores for two training photographs of the same person. This is an established form of graph-based regularization applied to a specific portrait-learning problem. We derive its exact connection to photograph substitution, then compare it with the identical scorer with the consistency term removed.

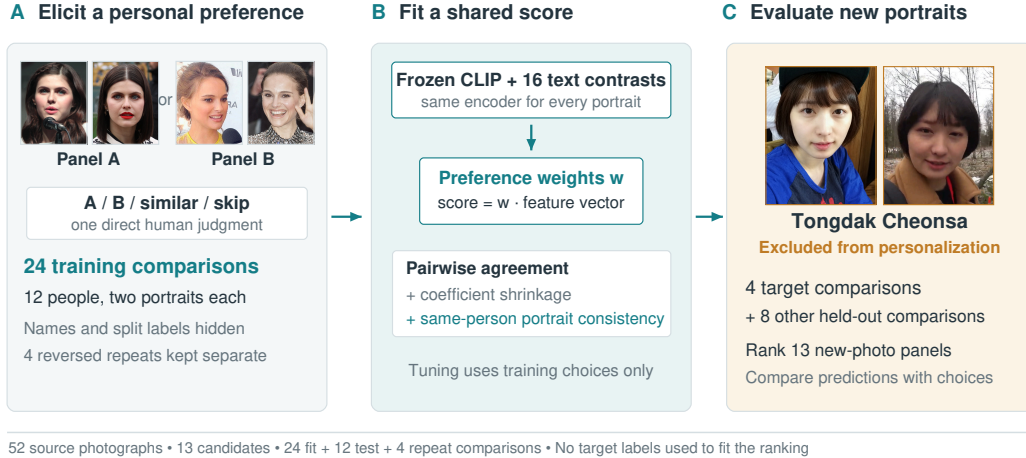


Figure 1: The study asks whether preferences learned from other people transfer to Tongdak Cheonsa. Each response compares two photographs of one person with two of another. Training includes twelve actresses; the target appears only at evaluation. The final thirteen-candidate ranking is a model prediction, checked only against the scheduled human comparisons.

Consistency alone is insufficient: a constant score is perfectly stable and predicts no preference. Conversely, a model may gain agreement by retaining expression or styling differences that the participant actually values. The study therefore keeps preference accuracy and photograph stability as distinct outcomes. It also separates new photographs of known candidates from the more demanding excluded-target case. A complete model ranking is reported as a prediction, with only the scheduled held-out choices serving as human validation.

Our contribution is a fully specified estimator and analytical interpretation, together with a prospective individual experiment that makes the preference–stability tradeoff inspectable. The observed outcome is a transfer failure: a lower photograph-variation penalty coexists with poor prediction of the observer’s new-photo choices. The small sample supports a detailed case study, not a population ranking or a claim that twenty-four comparisons are universally sufficient.

2 Related work

Personalized aesthetic judgments. Learning an individual’s visual preferences has a substantial history. Altwaijry and Belongie [Altwaijry and Belongie, 2013] learn a personal RankSVM from an ordering of facial images and evaluate on different faces. Their problem is a direct predecessor to ours; personal facial ranking itself is not a new task. Ren et al. [Ren et al., 2017] adapt a generic aesthetic predictor through compact aesthetic and content features. These approaches motivate learning a small preference head rather than attempting to train a visual encoder from a few choices. However, personalization does not automatically improve individual predictions. Goree et al. [Goree et al., 2023] collect personal enjoyment comparisons and show that average adaptation performance can obscure substantial differences between observers. We therefore use this participant’s own held-out choices, retain ties and skips, and compare errors on exactly the same trials.

Preference representations and pairwise fitting. CLIP [Radford et al., 2021] supplies a shared visual–language representation; a text similarity is a useful coordinate, but it is not a calibrated measurement of beauty. Recent personalization systems use richer shared training and supervision. ViPer [Salehi et al., 2024] learns from comments about visual attributes, PreferThinker [Xu et al., 2026] constructs and reasons over preference profiles, and Kim et al. [Kim et al., 2026] study personalized photographic processing using pairwise preferences. Personalize at Test Time [Bai et al., 2026], a recent preprint, is particularly close in using a regularized Bradley–Terry utility over a compact shared basis. Our linear likelihood model therefore builds on established machinery [Bradley and Terry, 1952]; we do not present a new pairwise learning principle. The distinction investigated

here is whether a same-person, different-photograph constraint helps a compact scorer transfer from displayed training panels to different portraits and one excluded target identity.

Photograph variation and consistency. Different photographs of the same person can change an observer’s impressions [Todorov and Porter, 2014]. The magnitude depends on the judgment: in their first experiment, attractiveness exhibited more between-person than within-person variance, although image selection could still reverse particular rankings. This motivates measuring photograph sensitivity rather than assuming that either all variation is identity or all variation is nuisance. Our penalty is a small graph regularizer [Belkin et al., 2006]: the two training portraits of a known person are connected, and a linear scorer is discouraged from assigning different scores along that edge. The resulting objective is generalized ridge regression with a logistic preference loss. Its aesthetic assumption is testable and potentially wrong: expression, styling, or photographic presentation can be part of what the participant likes.

The present work contributes an explicit two-portrait estimator, an exact interpretation of its regularizer under photograph substitution, and a controlled individual case study. It does not establish state-of-the-art performance on existing aesthetics benchmarks. Those studies use different participants, images, supervision, and outcomes; their published accuracy figures are not comparable scores for this experiment.

3 Learning a preference that survives photograph changes

3.1 From photographs to a panel score

For candidate i , let x_{iv} be photograph v , $e(x_{iv}) \in \mathbb{R}^{512}$ its unit-normalized frozen visual embedding, and t_k^+, t_k^- a fixed positive/negative text contrast. We define

$$a_k(x_{iv}) = e(x_{iv})^\top e(t_k^+) - e(x_{iv})^\top e(t_k^-), \quad z_{iv,k} = \frac{a_k(x_{iv}) - \mu_k}{\max(\sigma_k, 10^{-6})}. \quad (1)$$

The sixteen contrasts cover expression, visible facial form, styling, pose, and photographic presentation. They are fixed text probes, not independently validated facial attributes. The mean μ_k and standard deviation σ_k use only the 24 training portraits. Neither target photographs nor other held-out photographs enter this normalization.

A displayed panel contains two photographs of the same candidate. Its feature vector and utility are

$$m_i = \frac{1}{2}(z_{i1} + z_{i2}), \quad s_i = w^\top m_i, \quad p(i \succ j) = \text{sigmoid}(w^\top (m_i - m_j)). \quad (2)$$

The two photographs are not treated as two independent preference labels. One panel comparison produces one response. This mean defines the object being ranked: the person as represented by this particular pair of photographs. A higher score predicts this participant’s preference under the study conditions; it is not an absolute attractiveness rating.

3.2 A penalty on within-person score variation

For each of I training identities, define $\delta_i = z_{i1} - z_{i2}$ and

$$S = \frac{1}{4I} \sum_{i=1}^I \delta_i \delta_i^\top. \quad (3)$$

Given T usable training comparisons with feature differences $d_t = m_{i_t} - m_{j_t}$ and responses $y_t \in \{0, \frac{1}{2}, 1\}$, we minimize

$$\mathcal{L}_{\lambda,\gamma}(w) = \frac{1}{T} \sum_{t=1}^T [\log(1 + \exp(w^\top d_t)) - y_t w^\top d_t] + \frac{\lambda}{2} \|w\|_2^2 + \gamma w^\top S w. \quad (4)$$

A left preference has $y = 1$, a right preference $y = 0$, and a tie $y = 1/2$. A skipped response contributes no supervised term. All prespecified training photographs still define the feature normalization and S .

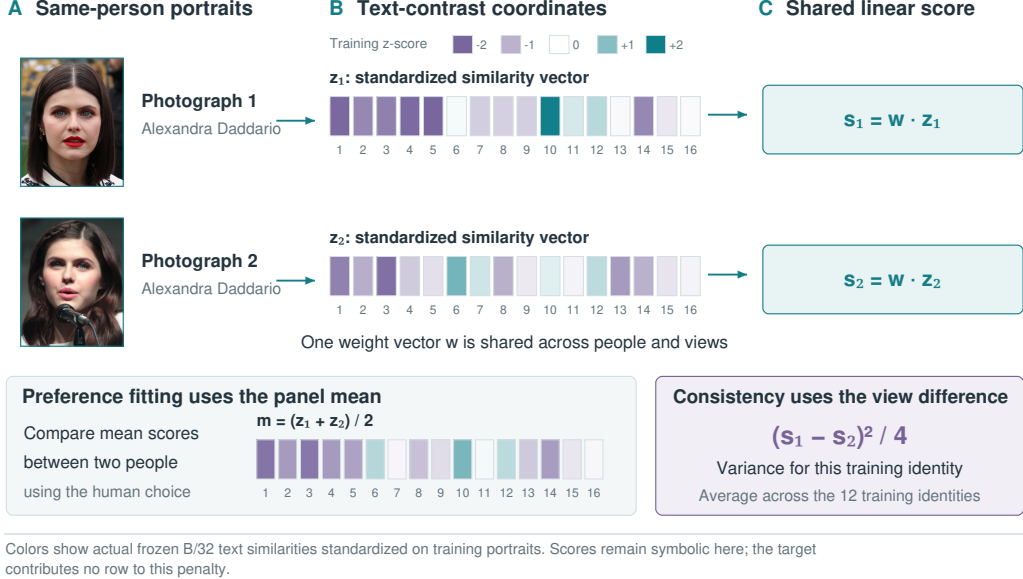


Figure 2: Two uses of the same photographs. Preference fitting compares panel means; consistency penalizes the score difference between views of each training person. The colored coordinates are actual standardized text similarities for two training portraits, with color saturation at ± 2 ; no personal weights are represented by the colors. One linear head is shared across photographs and identities. Coordinate definitions appear in Table 10.

Since S is positive semidefinite and $\lambda > 0$, the objective is strictly convex and has a unique minimizer. The penalty matrix is $\lambda I + 2\gamma S$: directions that change a person’s score substantially between training photographs receive stronger shrinkage. We solve the objective with L-BFGS and check the gradient norm. A sixteen-coordinate head is computationally small, but not fully identified by the comparisons alone: differences among twelve fixed panel means span at most eleven dimensions. Regularization and the feature representation therefore materially influence the fitted score.

Exact substitution interpretation. Write $u_{iv} = w^\top z_{iv}$ and $V_i = (u_{i1} - u_{i2})^2 / 4$. Then

$$W \equiv w^\top S w = \frac{1}{I} \sum_i V_i. \quad (5)$$

If a training identity is chosen uniformly and two of its photographs are drawn independently with replacement, the expected squared change in score is $2W$. Thus the consistency term is exactly a penalty on a specified photograph-substitution sensitivity, rather than a generic preference for small coefficients.

The same quantity bounds a related logistic-loss perturbation. Replace each two-photograph panel by one of its two photographs, independently and uniformly. For a fixed comparison i, j , the new margin is $s_i - s_j + \varepsilon_{ij}$, where $\mathbb{E}\varepsilon_{ij} = 0$ and $\mathbb{E}\varepsilon_{ij}^2 = V_i + V_j$. Because logistic cross entropy has second derivative at most $1/4$,

$$0 \leq \mathbb{E}\ell(s_i - s_j + \varepsilon_{ij}, y) - \ell(s_i - s_j, y) \leq \frac{V_i + V_j}{8}. \quad (6)$$

For the complete 24-comparison training schedule, where every identity appears four times, the average upper bound is $W/4$. Appendix A gives the proof and the unequal-degree version applicable after skips. This bound keeps the original panel label fixed. It does not assert that a person would give the same response after a photograph change, nor does it guarantee performance on unseen photographs or the excluded target.

3.3 Selection and matched comparisons

We select $\lambda \in \{0.1, 1, 10\}$ and $\gamma \in \{0.1, 1, 10\}$ using leave-one-training-comparison-out cross entropy. This estimates prediction of another edge among the known training identities; it is not an unseen-identity validation scheme. Ties in mean validation loss are resolved deterministically by stronger ridge, then weaker consistency. The primary encoder and grids are fixed before collection.

The decisive ablation retains the same features, choices, and selected λ and sets $\gamma = 0$. A second semantic baseline tunes its own ridge without consistency. Together, they distinguish the effect of the added penalty at a matched ridge from comparison with an independently tuned alternative. Additional controls include a generic beauty-text contrast, a signed pair-difference centroid, identity-only Bradley-Terry, and a visual-embedding Bradley-Terry head. Their definitions and transfer assumptions appear in Table 2.

4 An individual portrait-comparison study

Candidates and target. We assemble four independently sourced photographs for each of twelve adult actresses and four user-supplied photographs of Tongdak Cheonsa, for 52 images and thirteen identities. The actress pool was chosen at the participant’s request before any responses were collected. An earlier Korean-celebrity setup, containing no human responses, was archived and replaced; neither a preference result nor model performance prompted this revision. The final image files, source groups, prompts, split, trial order, and model settings were then locally timestamped and hash-frozen. This is a local prospective specification, not an external preregistration.

Tongdak Cheonsa is the prespecified target case. She contributes no photograph to feature normalization or the consistency penalty and no comparison to preference fitting or hyperparameter selection. Her two primary photographs are used for four held-out comparisons; the remaining two are reserved for a source-substitution diagnostic. This exclusion concerns personalization. We cannot establish that her identity or any actress was absent from the encoder’s original pretraining data.

Photographs and viewing conditions. Source metadata records the photographer or credited author, Commons revision, license, date, source group, and file hash. Every actress has four distinct source groups. Source separation within a person does not by itself imply that all public events are globally separated between splits. The complete manifest permits inspection of that distinction. Each original is EXIF-oriented, converted to RGB, and cropped to a face-centered 4:5 rectangle using a fixed landmark-padding rule. The display copy is 512×640 pixels. We apply no retouching, generative editing, or face warping; known editing already present in a source is retained in its provenance record. The participant sees these canonical portraits; the encoder’s official processor resizes them to 224×280 and takes a 224×224 center crop. Thus the human and model receive the same underlying source, with a narrower field of view for the model. Appendix H illustrates the exact sequence.

Direct human responses. One participant completes forty browser comparisons in Korean. The instruction asks which side feels more attractive when considering both photographs together, focusing on the faces and personal judgment. Each side contains two photographs of one person. The interface hides names and train/test membership, randomizes left/right placement and photo order, and offers left, right, similar, and skip. Names being hidden does not make famous faces unrecognizable. The participant is the requester of the study and has an existing interest in the target, so this is a deliberate personal-preference case rather than a blinded population sample.

The schedule contains 24 training comparisons, eight comparisons of the same twelve identities in different photographs, four comparisons involving the excluded target, and four reversed repeats. The training graph is connected and four-regular. Non-target test pairs are different identity pairs from the training edges, and target comparisons are balanced two per screen side. Repeats reproduce earlier training panels with left and right reversed and are never added to fitting. Figure 13 shows the complete scheduled graph.

Frozen predictions before evaluation. Feature extraction uses two pinned CLIP checkpoints, ViT-B/32 as primary and ViT-B/16 as a prespecified representation check. Training reads only the 24 training-phase response fields. It saves every model’s scores, selected penalties, tuning losses, di-

Table 1: Choice prediction with the primary ViT-B/32 encoder. Each accuracy entry is the number correct over decisive, non-skipped human choices. Fixes/regressions are paired against Semantic BT at the matched ridge.

Method	Train	Other	Target	All test	Test CE	Fix/reg.
Generic text	12/22	1/7	0/4	1/11	–	1/2
Pair centroid	18/22	4/7	2/4	6/11	–	5/1
Identity BT	18/22	6/7	3/4	9/11	0.600	8/1
Visual BT	19/22	4/7	2/4	6/11	0.696	5/1
Semantic BT, tuned	18/22	2/7	0/4	2/11	0.784	1/1
Consistency scorer	18/22	3/7	0/4	3/11	1.074	1/0
Semantic BT, matched	18/22	2/7	0/4	2/11	1.283	0/0

Human ties contribute target 1/2 to cross entropy (CE) but not to accuracy. The uncalibrated generic-text and centroid likelihoods are omitted from this comparison; their raw values remain in the records. A tied model prediction counts as incorrect for a decisive human choice.

agnostic penalty path, and bootstrap fits, then hashes these files before the evaluation script opens held-out answers. This separation reduces the opportunity for test-driven changes [Kapoor and Narayanan, 2023]; it does not turn a small purposive candidate pool into a representative sample.

Outcomes and sensitivity analyses. The primary outcome is the count of correct predictions among non-tied, non-skipped held-out choices. We report all twelve scheduled trials, the eight non-target trials, and the four target trials separately. Exact zero model margins predict a tie and count as incorrect for a decisive human response. Cross entropy includes human ties with target 1/2; skips are omitted. We report paired fixes and regressions against the matched semantic baseline rather than relying only on aggregate percentages. Generic text and centroid scores have uncalibrated scales, so their raw likelihood values are not treated as comparable calibrated probabilities.

The model-induced ranking averages the two test-portrait scores for each of thirteen candidates. It is only partially checked by the twelve held-out choices; we do not possess a complete human ranking of thirteen people. We enumerate all $2^{13} = 8192$ single-photograph substitutions, replace the target’s primary panel with its two alternate photographs, and compare portrait crops with original framing. Framing is studied both by refitting the complete pipeline and by transferring the portrait-fitted head and normalization unchanged. Human reference labels remain those for the canonical panels in either diagnostic. Finally, 1000 bootstrap fits resample the observed training comparisons while fixing the features, photographs, and selected penalties. Their rank ranges describe conditional fitting sensitivity. They are not population confidence intervals, new human judgments, or uncertainty over candidate selection. Test trials share identities and all responses come from one person; we therefore report finite-set counts without an independent-binomial significance claim.

5 Empirical results

Response accounting and the primary comparison. All forty responses were completed: training contains 22 decisive choices and two ties; held-out testing contains eleven decisive choices and one tie; no trial was skipped. All four reversed repeats select the same identity as their original trial. These four agreements describe the repeated questions, not a general reliability estimate. Predictions for every method, framing diagnostic, and bootstrap fit were frozen before held-out response fields were opened.

Training-pair validation selects $\lambda = 0.1$, $\gamma = 10$ for the primary scorer, with mean validation cross entropy 0.5426. The matched ablation retains $\lambda = 0.1$ and sets $\gamma = 0$; the independently tuned semantic baseline selects $\lambda = 1$. Table 1 reports the resulting predictions. Consistency matches 3/11 decisive held-out choices, compared with 2/11 for either semantic baseline, 6/11 for the visual head, and 9/11 for the identity-only control. The improvement over the matched ablation is one fix and zero regressions: the held-out preference for Scarlett Johansson over Ana de Armas. Figure 3 exposes all twelve trials, including the tie.

Every held-out choice, compared on the same photographs

	Generic text	Pair centroid	Identity BT	Visual BT	Semantic BT (tuned ridge)	Semantic BT (matched ridge)	Consistency scorer
Elizabeth Olsen > Lily Collins	0	0	1	0	0	1	1
Keira Knightley > Ana de Armas	0	0	1	0	0	0	0
Gal Gadot > Emma Watson	0	1	0	1	1	1	1
Emma Watson > Natalie Portman	1	1	1	1	0	0	0
Keira Knightley > Natalie Portman	0	0	1	0	0	0	0
Emma Stone > Anne Hathaway	0	1	1	1	0	0	0
Scarlett Johansson > Ana de Armas	0	1	1	1	1	0	1
Alexandra Daddario = Margot Robbie	=	=	=	=	=	=	=
Tongdak Cheonsa > Lily Collins	0	1	1	0	0	0	0
Tongdak Cheonsa > Elizabeth Olsen	0	1	1	1	0	0	0
Tongdak Cheonsa > Ana de Armas	0	0	0	0	0	0	0
Tongdak Cheonsa > Alexandra Daddario	0	0	1	1	0	0	0
	1/11	6/11	9/11	6/11	2/11	2/11	3/11

1 agrees with decisive choice
 0 disagrees
 = human tie
 - skipped

Figure 3: Every held-out response, using the primary encoder. Row labels give the actual human preference; cells show agreement with a decisive choice, while the purple row records a human tie excluded from the accuracy denominator. The consistency scorer repairs one matched-baseline error and still misses all four target choices. Identity BT’s target predictions use a fixed neutral utility, not the target’s visual appearance.

The consistency head’s test cross entropy is 1.074, below the matched ablation’s 1.283 but above the independently tuned semantic head’s 0.784. Thus a single corrected sign does not establish better probability prediction than the separately tuned baseline. Identity BT correctly predicts 6/7 decisive non-target choices and 3/4 target choices. Its latter result arises by comparing the target’s fixed zero utility with learned opponent utilities; it cannot be interpreted as successful visual recognition of the target’s personal appeal.

Consistency is controlled most strongly where it is fitted. For the matched and consistency heads respectively, W_{train} falls from 0.3254 to 0.000472, whereas W_{test} falls from 0.6704 to 0.4036. The latter is a 39.8% reduction in raw score variance, measured across all thirteen test panels. The weight norm changes only from 0.988 to 0.942. Individual test-photo gaps nevertheless increase for several candidates (Figure 6). The regularizer therefore changes the direction of the fitted head and tightly controls observed training views; its much weaker and nonuniform transfer does not imply an improvement in human agreement. Exact substitution-loss diagnostics in Table 4 satisfy the analytical bound.

The target case: observed preference and predicted ranking diverge. The observer prefers Tongdak Cheonsa to each of the four scheduled opponents: Lily Collins, Elizabeth Olsen, Ana de Armas, and Alexandra Daddario. The primary scorer predicts the opposite in all four cases (Figure 4). It ranks the target 12th among thirteen test panels, with nearly identical primary-photo scores, -1.310 and -1.355 . Replacing her panel by alternate sources B and C leaves the rank at 12; enumerating all 8192 single-photo configurations yields ranks 12–13. Stability here preserves a prediction that conflicts with the observed choices.

This four-opponent result does not establish that the observer ranks the target first among all thirteen candidates. The complete model ordering, including its highest-scored candidate Scarlett Johansson, is shown with every portrait score in Figure 5. A decomposition of the unasked target–Scarlett margin is available in Figure 8; it explains the computed score, not a collected human comparison or a causal reason for attraction.

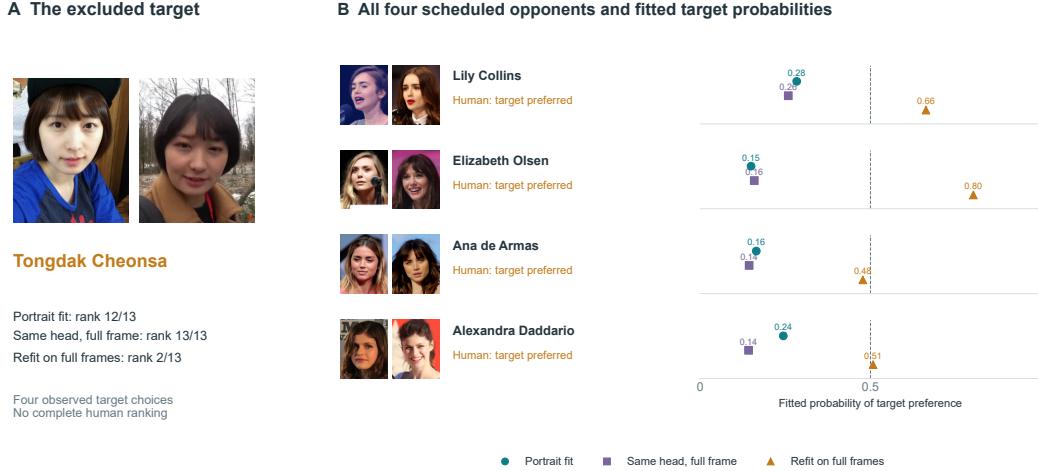


Figure 4: The excluded target and every scheduled opponent. Photographs show the canonical panels actually judged. Points are fitted Bradley–Terry probabilities for preferring the target, with a decision threshold at 0.5. The primary portrait head misses all four preferences. Transferring that head to original frames also misses all four; refitting on original framing matches three. These diagnostic conditions use the same original human labels, without claiming that full-frame preferences were separately observed.

Representation and photograph diagnostics. The second encoder, ViT-B/16, also yields 3/11 overall and 0/4 target choices for its portrait consistency head, ranking the target 13th. Framing produces larger changes: a B/32 refit on original frames yields 5/11 overall, 3/4 target choices, and target rank 2, whereas the B/16 original-frame refit yields 0/11 and target rank 13 (Table 6). Replacing only the target by alternate photographs changes even the B/32 original-frame refit from rank 2 to rank 6 and from three to one positive predictions against the four scheduled opponents. Holding the portrait-fitted weights and normalization fixed while changing only framing is a different diagnostic; its B/32 target rank is 13. Selecting the favorable full-frame result after inspection would conceal this representation sensitivity, so the prespecified portrait model remains primary.

The complete consistency-strength path is reported in Figure 7; it is not used to retune on test labels. Conditional resampling of the observed training comparisons gives a primary target-rank median of 12 and 5th–95th percentiles of 10–13, with no first-place fit among 1000 samples (Figure 10). These ranges keep features, photographs, and selected penalties fixed. They quantify fitting sensitivity within the case, not uncertainty over new observers or candidate collections.

6 Discussion and limitations

The experiment separates three outcomes: fitting personal comparisons, controlling score variation on observed photographs, and transferring to new photographs or a new person. The primary head attains 18/22 decisive training choices and nearly eliminates training-view variation, yet reaches only 3/11 on held-out choices. Its target score is stable across the two primary photographs but disagrees with every observed target preference. This is direct evidence that the optimized consistency criterion is insufficient for the requested transfer in this case.

The gap between training-pair validation and test performance is also informative. Validation withholds an edge among fixed panel means, while the test changes the photographs and introduces the target. Better leave-one-edge-out loss need not select a representation that survives either shift. The strong identity-only result on known candidates suggests that a useful component of this observer’s ordering persists across the new photos; it does not supply a visual explanation of target preference. The source and representation limitations remain confounded in this small design.

The portrait collection is purposive. The participant requested the actress pool and has an existing interest in the target; photographs also vary in year, expression, styling, capture conditions, and

resolution. Publicity or event photographs of actresses and the user-supplied target photographs are not controlled acquisitions. Cropping reduces some framing differences but does not isolate facial geometry or erase these source differences. Excluding target inputs from personalization does not remove such distribution shifts or make the participant blind to identity.

A single observer, four target comparisons, and four repeats cannot establish general human agreement or robust transfer to other people. The twelve test trials share candidates and are not independent population draws. The thirteen-way ordering contains many comparisons never asked of the participant. A broader study would require additional observers, additional excluded identities, independent portrait acquisition, and a new untouched evaluation set. Those extensions should test the same claimed tradeoff rather than choose candidates after seeing favorable rankings.

7 Conclusion

Two-portrait comparisons and a within-person score penalty give a compact, inspectable estimator with an exact substitution interpretation. In the completed individual study, that estimator reduces score variation without reliably transferring personal choices to new portraits or recovering the observed preference for Tongdak Cheonsa. The distinction is substantive: learning a stable score is not sufficient to learn this observer’s beauty preference. Further development requires a representation and validation design tested on new photographs and excluded identities, followed by a fresh human evaluation.

References

- Hani Altwaijry and Serge Belongie. Relative ranking of facial attractiveness. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 117–124, 2013. doi: 10.1109/WACV.2013.6475008. URL <https://vision.ucsd.edu/publications/2013/relative-ranking-facial-attractiveness>.
- Jiamu Bai, Jiaming Hu, Yanhong Wu, and Zellux Wang. Personalize at test time: Learning user preferences for image generation, 2026. URL <https://arxiv.org/abs/2610.09015>. arXiv preprint, version 1, 6 October 2026.
- Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(85):2399–2434, 2006. URL <https://jmlr.org/papers/v7/belkin06a.html>.
- Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.
- Google AI Edge. MediaPipe Face Landmarker: Task Documentation and Model Bundle. https://developers.google.com/edge/mediapipe/solutions/vision/face_landmarker, 2026. Accessed 23 October 2026; local package and model hash recorded in the study materials.
- Samuel Goree, Weslie Khoo, and David J. Crandall. Correct for whom? subjectivity and the evaluation of personalized image aesthetics assessment models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11818–11827, 2023. doi: 10.1609/aaai.v37i10.26395. URL <https://ojs.aaai.org/index.php/AAAI/article/view/26395>.
- Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804. URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC10499856/>.
- Jinwoo Kim, Jihye Yoo, and Seon Joo Kim. Learning personalized photographic style from pairwise user preferences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1134–1144, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Kim_Learning_Personalized_Photographic_Style_from_Pairwise_User_Preferences_CVPR_2026_paper.html.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.

- Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. Personalized image aesthetics. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 638–647, 2017. doi: 10.1109/ICCV.2017.76. URL https://openaccess.thecvf.com/content_iccv_2017/html/Ren_Personalized_Image_Aesthetics_ICCV_2017_paper.html.
- Sogand Salehi, Mahdi Shafiei, Teresa Yeo, Roman Bachmann, and Amir Zamir. ViPer: Visual personalization of generative models via individual preference learning. In *Computer Vision – ECCV 2024*, 2024. doi: 10.1007/978-3-031-72904-1_23. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/09492.pdf.
- Alexander Todorov and Jenny M. Porter. Misleading first impressions: Different for different facial images of the same person. *Psychological Science*, 25(7):1404–1417, 2014. doi: 10.1177/0956797614532474. URL <https://journals.sagepub.com/doi/10.1177/0956797614532474>.
- Shengqi Xu, Xinpeng Zhou, Yabo Zhang, Ming Liu, Tao Liang, Tianyu Zhang, Yalong Bai, Zuxuan Wu, and Wangmeng Zuo. PreferThinker: Reasoning-based personalized image preference assessment. In *International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=iFrDYBKMff>.

A Properties of the consistency penalty

For the two training photographs of identity i , let $u_{iv} = w^\top z_{iv}$, $\bar{u}_i = (u_{i1} + u_{i2})/2$, and $V_i = (u_{i1} - u_{i2})^2/4$. A uniform draw $K \in \{1, 2\}$ has mean $\bar{u}_i$ and variance V_i . Independent draws K, K' therefore satisfy

$$\mathbb{E}(u_{iK} - u_{iK'})^2 = 2V_i. \quad (7)$$

Averaging uniformly across identities gives $2w^\top Sw$. If the two draws must be different, the squared difference is instead exactly $4V_i$, with mean $4w^\top Sw$.

For a comparison of i and j , independent single-photo substitution has margin perturbation

$$\varepsilon_{ij} = (u_{iK} - \bar{u}_i) - (u_{jL} - \bar{u}_j), \quad \mathbb{E}\varepsilon_{ij} = 0, \quad \mathbb{E}\varepsilon_{ij}^2 = V_i + V_j. \quad (8)$$

The loss $\ell(a, y) = \log(1 + e^a) - ya$ is convex and has curvature $\sigma(a)(1 - \sigma(a)) \leq 1/4$. Jensen's inequality yields the lower bound in Eq. 6. The smoothness inequality

$$\ell(a + \varepsilon, y) \leq \ell(a, y) + \ell'(a, y)\varepsilon + \frac{1}{8}\varepsilon^2 \quad (9)$$

and zero-mean perturbation yield the upper bound. Let $L_{\text{single photo}}$ denote the loss averaged over comparisons and expected over the independent uniform photograph draws, while L_{panel} uses the panel means. The nonnegative difference concerns this expectation, not each individual substitution configuration. For M usable comparisons and identity degrees d_i , averaging gives

$$0 \leq L_{\text{single photo}} - L_{\text{panel}} \leq \frac{1}{8M} \sum_i d_i V_i. \quad (10)$$

With $M = 24$, $I = 12$, and all $d_i = 4$, the right-hand side is $W/4$. Skipping comparisons changes the degrees, so the general expression must be used. These statements concern the fixed observed training photographs and fixed panel labels. Photograph replacement may change the participant's real preference, and none of these identities ensures transfer to a new photograph or person.

The objective Hessian is

$$\nabla^2 \mathcal{L} = \frac{1}{T} D^\top \text{diag}(p_t(1 - p_t)) D + \lambda I + 2\gamma S \succeq \lambda I. \quad (11)$$

Strict convexity follows for every positive ridge in our grid. The visual baseline can be fitted efficiently in the row span of its training-difference matrix: an orthonormal basis preserves logits and the Euclidean ridge norm, while any orthogonal component increases the penalty without changing the supervised loss. The basis obtained from all training differences also contains each leave-one-pair-out fold's row span. This equivalence applies to the ridge-only visual baseline; it is not assumed for the semantic model with an additional consistency matrix. Numerical tests check the reduction, analytical gradients, tie/reversal symmetry, and training-only normalization.

B Baseline definitions, selected fits, and numerical diagnostics

Table 2 specifies each method and the assumption it makes when applied to the excluded target. Table 3 retains the selected settings for both encoders and both framing refits. All choices were made using training-pair validation; none was chosen from the held-out results. Numerical optimization completed for the saved models, but optimization convergence is not evidence of successful preference transfer. The primary consistency model's expected training-loss increase after uniform single-photo substitution is 0.0001020, below the bound 0.0001180. For the matched head the corresponding values are 0.04459 and 0.08135. These are exact finite enumerations over the four photographic outcomes for each training pair with its label fixed. Their calculation checks the interpretation of the fitted regularizer and does not enlarge the empirical human sample.

The primary scatter matrix has rank 12 in sixteen dimensions, using an eigenvalue cutoff of 10^{-10} times its largest eigenvalue. Its four-dimensional nullspace contains 96.12% of the fitted head's squared norm. Directions in this space assign equal scores to each observed training-photo pair but need not do so for new photos. This post-hoc geometric diagnostic describes the saved solution; it neither selected the model nor establishes a causal explanation for its errors.

Table 2: Prespecified comparison methods. All personalized methods use the same training responses. λ_* is selected jointly with consistency for the proposed scorer. The independently tuned semantic baseline is retained as a second control.

Method	Representation	Fitting and transfer assumption
Generic text contrast	Four beauty/neutral prompt pairs	No personal labels; raw cosine contrast.
Pair centroid	512 visual coordinates	Unit-length mean signed training-pair difference; ties contribute zero.
Identity BT	12 one-hot identities	Ridge-tuned utilities; the unseen target receives neutral utility zero.
Visual BT	512 standardized visual coordinates	Ridge-tuned shared linear head; numerical row-span reduction preserves the ridge solution.
Semantic BT, tuned	16 text contrasts	Independently selected ridge; no consistency term.
Semantic BT, matched	Same 16 contrasts	Same λ_* , same inputs; set $\gamma = 0$.
Consistency scorer	Same 16 contrasts	Select λ, γ from training pairs; penalize score variation between two training portraits of each person.

Table 3: Selected penalties and numerical convergence. Selection uses leave-one-training-comparison-out cross entropy only. The full tuning grid and fold losses are retained in the frozen prediction files.

Encoder	Framing	Model	λ	γ	LOO CE	$\ \nabla L\ _\infty$
B/32	portrait	Identity BT	0.1	0.0	0.6237	1.8e-10
B/32	portrait	Visual BT	10.0	0.0	0.6346	3.8e-08
B/32	portrait	Semantic BT, tuned	1.0	0.0	0.6249	1.3e-08
B/32	portrait	Consistency scorer	0.1	10.0	0.5426	5.7e-07
B/32	full	Identity BT	0.1	0.0	0.6237	1.8e-10
B/32	full	Visual BT	10.0	0.0	0.6107	2.6e-08
B/32	full	Semantic BT, tuned	0.1	0.0	0.5893	2.2e-08
B/32	full	Consistency scorer	0.1	10.0	0.5945	3.0e-07
B/16	portrait	Identity BT	0.1	0.0	0.6237	1.8e-10
B/16	portrait	Visual BT	10.0	0.0	0.6320	1.5e-07
B/16	portrait	Semantic BT, tuned	0.1	0.0	0.6059	6.7e-08
B/16	portrait	Consistency scorer	0.1	0.1	0.6109	9.7e-08
B/16	full	Identity BT	0.1	0.0	0.6237	1.8e-10
B/16	full	Visual BT	10.0	0.0	0.6144	9.5e-07
B/16	full	Semantic BT, tuned	1.0	0.0	0.6160	4.6e-09
B/16	full	Consistency scorer	1.0	0.1	0.6176	9.8e-09

Table 4: Exact diagnostics of the saved portrait-fitted heads. With/without refers to the consistency term at matched ridge. The final columns enumerate the four single-photo outcomes for each training pair while keeping its original human panel label fixed.

Encoder	Consistency	$\ w\ _2$	W_{train}	W_{test}	$\Delta\ell$	Bound
B/32	Without	0.988	0.3254	0.6704	0.04459	0.08135
B/32	With	0.942	0.0004718	0.4036	0.000102	0.000118
B/16	Without	1.121	0.8647	1.136	0.1412	0.2162
B/16	With	0.911	0.2077	0.7912	0.04095	0.05192

W is the mean within-person variance of the two saved scores. The nonnegative expected loss increase is bounded by Eq. 6; this numerical identity does not add human judgments or establish transfer to new photos.

C Complete predicted ranking and photo sensitivity

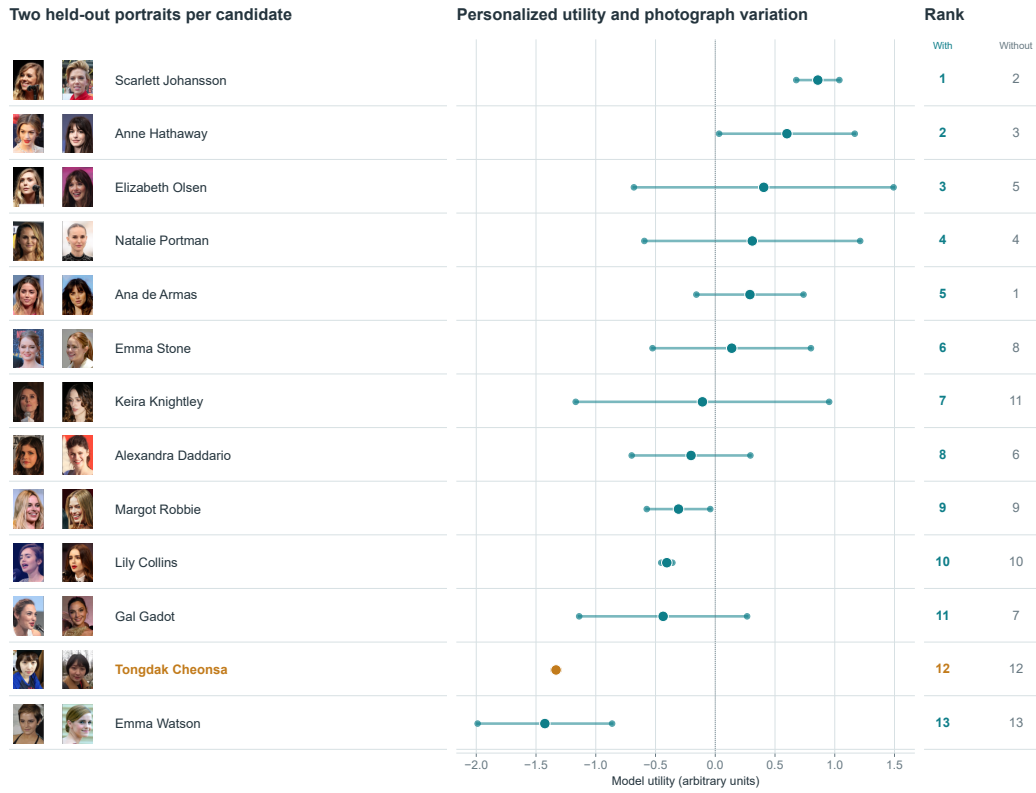


Figure 5: All thirteen test panels under the primary consistency head. Large dots are two-photo means; small dots are individual-photo utilities, joined only to show their difference. Right columns compare ranks with and without consistency at the same ridge. Utilities have no universal beauty scale. The target–Scarlett comparison and many other pairs were never asked of the participant, so the entire ordering is not an observed human ranking.

Table 5: Tongdak Cheonsa as the excluded target. Ranks use thirteen candidates; one is highest. Alternate photos replace only the target panel. The single-photo range enumerates all 8192 choices of one primary test photo per candidate.

Method	Primary rank	Alternate rank	Single-photo range	Target choices
Generic text	12.0	11.0	5–13	0/4
Pair centroid	11.0	11.0	8–13	2/4
Identity BT	7.0	7.0	7–7	3/4
Visual BT	9.0	8.0	7–13	2/4
Semantic BT, tuned	11.0	13.0	9–13	0/4
Consistency scorer	12.0	12.0	12–13	0/4
Semantic BT, matched	12.0	12.0	9–13	0/4

The alternate panel and single-photo configurations have no additional human labels. Their ranks are sensitivity diagnostics, not additional accuracy observations.

Stability on the fitted photographs does not transfer uniformly



Primary B/32 model and matched-ridge ablation. Each cell is an exact difference of the two saved image scores. The target has no training photographs; low variation is not evidence of agreement with the observer.

Figure 6: Exact absolute differences between the two photograph scores for each candidate. All cells use the same primary B/32 representation and matched-ridge ablation. Training-view differences mostly shrink sharply, while held-out differences can increase. The target is excluded from the training columns; her small test-photo gap coexists with four incorrect preference predictions.

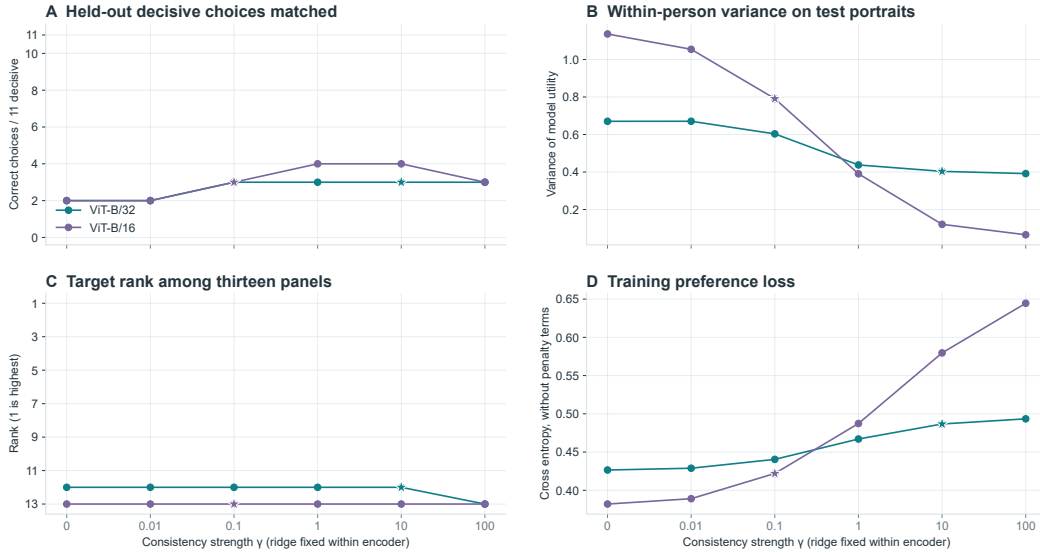
D Framing, representation, and regularization paths

The two original-framing experiments answer different questions. A full refit changes training normalization, panel differences, within-person scatter, and the selected head. A fixed-head transfer preserves the portrait-fitted normalization and weights. Both are evaluated against the original canonical-portrait labels, because no separate full-frame judgments were collected.

Table 6: Representation and framing sensitivity. “With/without” denotes the consistency term; the no-consistency head uses the matched ridge in each fit. The fixed-head middle condition preserves the portrait normalization and weights.

Encoder	Fit / evaluated input	Consistency	All test	Target	Rank	Alt. rank
B/32	Portrait fit / portrait input	Without	2/11	0/4	12.0	12.0
B/32	Portrait fit / portrait input	With	3/11	0/4	12.0	12.0
B/32	Portrait fit / original input	Without	1/11	0/4	12.0	11.0
B/32	Portrait fit / original input	With	3/11	0/4	13.0	13.0
B/32	Original fit / original input	Without	3/11	1/4	9.0	9.0
B/32	Original fit / original input	With	5/11	3/4	2.0	6.0
B/16	Portrait fit / portrait input	Without	2/11	0/4	13.0	13.0
B/16	Portrait fit / portrait input	With	3/11	0/4	13.0	13.0
B/16	Portrait fit / original input	Without	0/11	0/4	13.0	13.0
B/16	Portrait fit / original input	With	2/11	0/4	13.0	13.0
B/16	Original fit / original input	Without	0/11	0/4	13.0	13.0
B/16	Original fit / original input	With	0/11	0/4	13.0	13.0

Every diagnostic is compared with the same canonical-portrait human answers. Rank columns use primary and alternate target photographs. Refitting changes the training representation and selected penalties; it is distinct from transferring a fixed head.

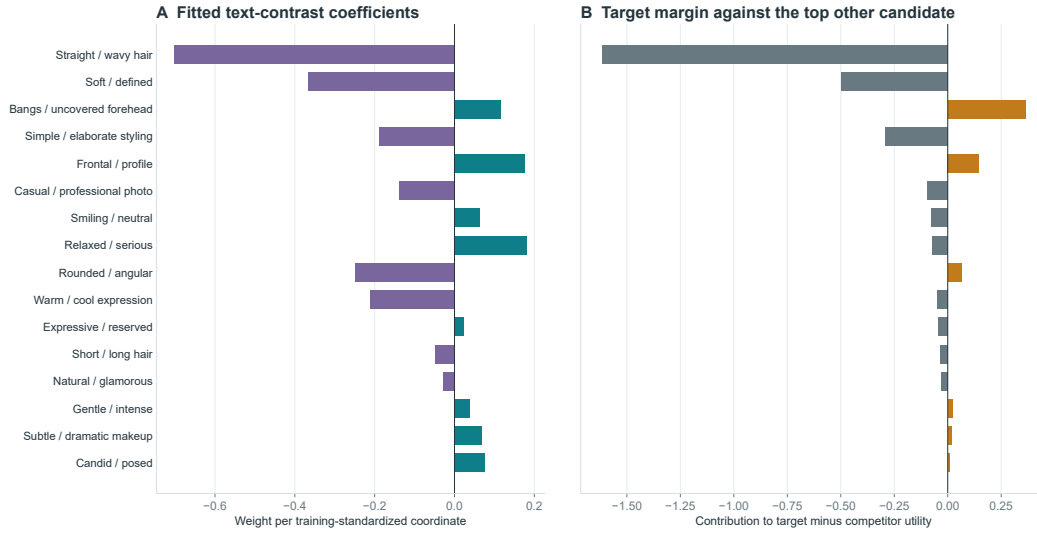


Stars mark penalties chosen only by training-pair validation. The full path was specified before responses; test performance does not select γ . Smaller variance can arise through score shrinkage and is interpreted alongside choice accuracy, not as success by itself.

Figure 7: The prespecified diagnostic consistency path at the selected ridge for each encoder. Stars mark the training-validation choices. All test outcomes were computed only after predictions were frozen; the displayed paths do not select a new model. Reduced raw score variance can coexist with poor held-out agreement and low target rank. The final panel shows supervised training loss, excluding both penalty terms.

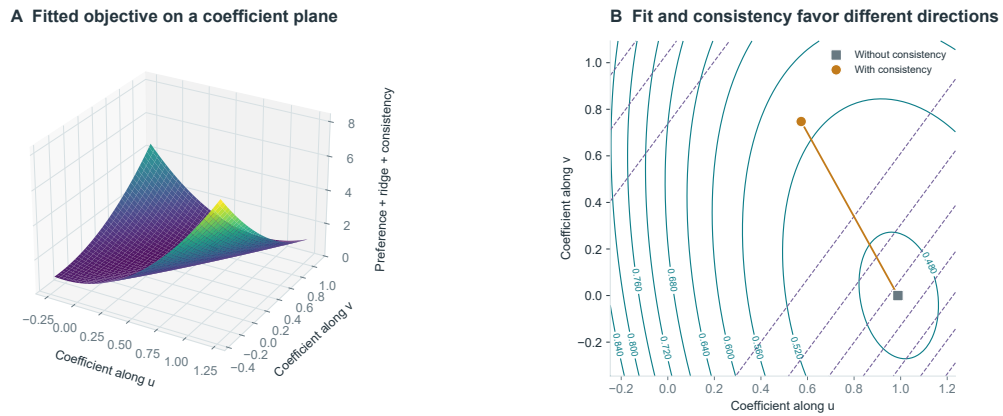
E Inspecting the fitted head

The linear form permits an exact additive decomposition. For the target and highest-ranked non-target panel, the margin is $\sum_k w_k(m_{\text{target},k} - m_{\text{other},k})$. Figure 8 displays all sixteen terms, ordered by their absolute contribution. The largest negative term is the straight-versus-wavy-hair text contrast. This is a property of the fitted representation and coefficients; it neither validates those text probes as independent attributes nor reveals the participant’s reason for a choice. In particular, no target–Scarlett human comparison was collected.



Competitor: Scarlett Johansson. Sum of all contributions = -2.1907. Order follows absolute contribution, without dropping coordinates. These are exact terms in the fitted score, not measured facial attributes or causal explanations of the participant's choice.

Figure 8: Exact coefficients and score contributions for the primary head. Left: coefficient per standardized coordinate. Right: each coordinate's contribution to the target-minus-Scarlett margin. Terms sum to -2.1907 , and none is omitted. The displayed comparison is a model diagnostic without a corresponding human label.



Plane contains both fitted 16-dimensional heads; $\lambda = 0.1$, $\gamma = 10$. Solid contours: preference + ridge; dashed contours: consistency. Surface coordinates are coefficient directions, not image features or a recorded optimization trajectory.

Figure 9: The actual primary training objective on an orthonormal plane containing the matched and consistency heads. The first direction follows the matched head; the second is the orthogonal component of the consistency head. Left: total selected objective. Right: preference-plus-ridge contours and consistency contours, with the two saved optima marked. Their connecting segment is a geometric comparison, not a recorded training trajectory.

F Conditional training-resampling sensitivity

Each resample draws 24 training comparisons with replacement. The same feature representation, normalization, photographs, and selected penalties are retained; only the fitted weights change. The procedure preserves the target’s exclusion from preference fitting. Its distributions describe sensitivity to this resampling convention, rather than independently sampled people, test trials, or observers.

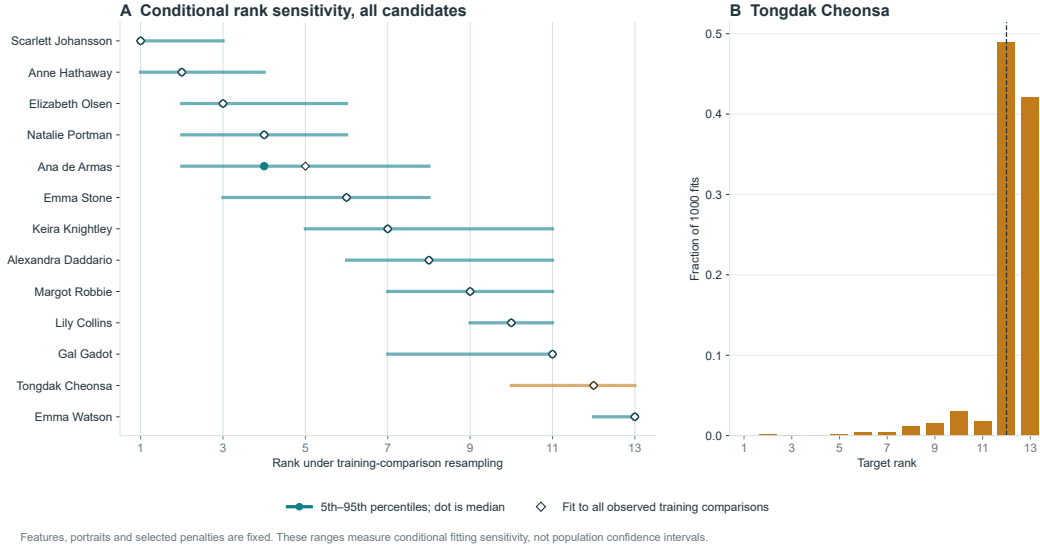


Figure 10: All candidate rank distributions across 1000 conditional training resamples. Bars span the 5th–95th empirical percentiles, circles mark medians, and diamonds show the original fit. The target histogram retains all observed ranks, including rare tails; its median is 12. These are sensitivity summaries, not population confidence intervals.

G Response ledger and reproducibility

The complete response sequence is listed below, including ties and all reversed repeats. The target trials are shaded. Recorded sides are screen positions rather than identity labels; the evaluator resolves the selected identity through the frozen schedule. Timing is retained to document collection and never used to exclude a response.

Both feature extractors and all preference fits ran on CPU. Feature extraction for the two framing variants of all 52 images took 9.718 seconds for B/32 and 22.633 seconds for B/16 in the recorded environment. Fitting all baselines, tuning folds, framing conditions, and 1000 conditional resamples per encoder took 18.470 seconds. These are local run times, not hardware-independent performance guarantees. The pinned checkpoint revisions, Python package versions, model-asset hash, file hashes, and regeneration commands are included in the distributed materials.

The original sources and recorded crop rectangles regenerate all 104 canonical and full-frame JPEGs byte-for-byte under the recorded image-library version. Numerical checks cover the loss gradient, the within-person variance identity, the visual baseline’s orthonormal reduction, response reversal, and train-only normalization. The paper’s plots are derived from the saved predictions and evaluation; their editable sources and export manifests accompany the PDF. The participant supplied the preferences; coding, literature synthesis, and manuscript preparation used AI assistance.

Table 7: The fixed collection schedule. A response compares two two-photograph panels. Photographs are separated within each actress; all target photographs are excluded from personalization. Repeat answers are never training labels.

Phase	Pairs	Portraits	Purpose
Personalization	24	24	Twelve identities, two photos each; four comparisons per identity.
New-photo test	8	24	Same twelve identities, different photos and different comparison edges.
Target test	4	2 target	Target vs. four prespecified opponents using their test panels.
Reversed repeats	4	Reused	Check four earlier choices after reversing screen position.
Target source swap	0	2 target	Additional photographs for model sensitivity, with no new human labels.
Total	40	52 unique	One participant; thirteen candidate identities.

Table 8: Provenance pointers. SHA-256 prefixes are shown for readability; complete hashes and all frozen file relationships are distributed in the linked manifests.

Artifact	SHA-256 prefix
Final protocol	88754a9c5b4160ae
Input freeze	d67a445e998761da
Human responses	cfaf481649d3d2f3
Prediction freeze	398999cb183ed297
Evaluation	bcc2b9cb08ae30cc

Table 9: Complete response ledger in collection order. Screen sides and trial phases are reconstructed from the frozen protocol; phases were hidden in the participant interface.

No.	Phase	Left candidate	Right candidate	Response	Seconds
1	Train	Alexandra Daddario	Natalie Portman	right	742.0
2	Target test	Lily Collins	Tongdak Cheonsa	right	1.8
3	Train	Margot Robbie	Keira Knightley	right	3.1
4	Train	Elizabeth Olsen	Scarlett Johansson	right	2.0
5	Train	Alexandra Daddario	Scarlett Johansson	right	2.0
6	Other test	Lily Collins	Elizabeth Olsen	right	1.7
7	Train	Alexandra Daddario	Ana de Armas	right	1.6
8	Other test	Keira Knightley	Ana de Armas	left	3.6
9	Other test	Gal Gadot	Emma Watson	left	3.7
10	Train	Keira Knightley	Emma Stone	left	2.5
11	Train	Ana de Armas	Lily Collins	left	2.5
12	Target test	Tongdak Cheonsa	Elizabeth Olsen	left	1.8
13	Train	Ana de Armas	Margot Robbie	right	12.6
14	Train	Scarlett Johansson	Margot Robbie	left	1.3
15	Other test	Emma Watson	Natalie Portman	left	1.7
16	Train	Gal Gadot	Margot Robbie	right	2.6
17	Repeat	Natalie Portman	Alexandra Daddario	left	2.4
18	Train	Natalie Portman	Elizabeth Olsen	tie	3.2
19	Train	Natalie Portman	Emma Stone	right	1.7
20	Train	Emma Stone	Gal Gadot	right	5.9
21	Train	Ana de Armas	Anne Hathaway	left	5.5
22	Other test	Natalie Portman	Keira Knightley	right	2.6
23	Train	Gal Gadot	Alexandra Daddario	left	1.4
24	Other test	Emma Stone	Anne Hathaway	left	8.1
25	Repeat	Keira Knightley	Margot Robbie	left	1.7
26	Train	Scarlett Johansson	Lily Collins	left	1.6

No.	Phase	Left candidate	Right candidate	Response	Seconds
27	Other test	Ana de Armas	Scarlett Johansson	right	2.0
28	Train	Emma Watson	Emma Stone	right	1.6
29	Train	Emma Watson	Elizabeth Olsen	left	2.3
30	Train	Gal Gadot	Anne Hathaway	right	2.7
31	Train	Natalie Portman	Lily Collins	tie	2.2
32	Train	Lily Collins	Emma Watson	right	3.3
33	Repeat	Scarlett Johansson	Elizabeth Olsen	left	1.8
34	Train	Elizabeth Olsen	Keira Knightley	right	4.7
35	Target test	Ana de Armas	Tongdak Cheonsa	right	1.5
36	Train	Keira Knightley	Anne Hathaway	left	2.3
37	Other test	Alexandra Daddario	Margot Robbie	tie	2.5
38	Train	Anne Hathaway	Emma Watson	right	2.0
39	Target test	Tongdak Cheonsa	Alexandra Daddario	left	0.8
40	Repeat	Scarlett Johansson	Alexandra Daddario	left	2.1

Browser elapsed time includes any time spent away from the page and is not a controlled response-time measure. No choice is excluded using a time or repeat-consistency threshold.

H Portrait preprocessing and full input collection

Detection and deterministic cropping. MediaPipe Face Landmarker [Google AI Edge, 2026] is used only to locate a crop. We run its image mode with at most five detected faces and detection/presence thresholds of 0.5. Detection images have a maximum dimension of 1600 pixels. If the full frame yields no face, a fixed cascade tries upper-center windows with normalized bounds $(.125, 0, .875, .65)$ and $(.25, 0, .75, .5)$. A source is rejected if no face is detected, if the second-largest landmark box exceeds 30% of the largest box’s area, or if either native dimension of the selected landmark box is below 120 pixels. The retained foreground face is visually checked before collection; this check is not an attractiveness filter.

For a landmark box of width f_w , height f_h , and center (c_x, c_y) , the desired crop height is $h = \max(1.72f_h, 1.80f_w/0.8)$, width $0.8h$, and center $(c_x, c_y - 0.12f_h)$. The rectangle is reduced uniformly if needed to fit the original, translated within image bounds, rounded to integer coordinates, and resized with Lanczos to 512×640 . These coordinates and both source and output hashes are retained. No rotation alignment or facial deformation is applied. The encoder then uses its official resize and center-crop operations, as illustrated in Figure 11. The original-frame diagnostic starts from an aspect-preserving image whose maximum dimension is 1024 pixels.

Source selection before responses. The candidate pool contains Anne Hathaway, Natalie Portman, Scarlett Johansson, Margot Robbie, Ana de Armas, Emma Watson, Emma Stone, Alexandra Daddario, Keira Knightley, Lily Collins, Elizabeth Olsen, and Gal Gadot. Wikimedia Commons supplies four files from distinct recorded source groups per actress. Source records include a pinned Commons page revision, original-file SHA-1 and SHA-256, metadata dates, credited creator, license, and adult-at-capture check. A video still is allowed when its stated source has a compatible license. Two initial international-panel files were replaced because they failed the fixed detection or prominent-face criteria; their source revisions and rejection reasons are retained. This occurred before the final protocol freeze and before any human responses.

The source collection mixes years, events, photographers, image quality, expressions, and prior source edits. It is not a controlled acquisition dataset. Within-person source separation and exact-file independence are verified; photographer independence and global event separation are not assumed. The four target files retain the provenance of the requester’s earlier photo collection. Their reuse does not transfer the earlier study’s labels or outcomes to this experiment.

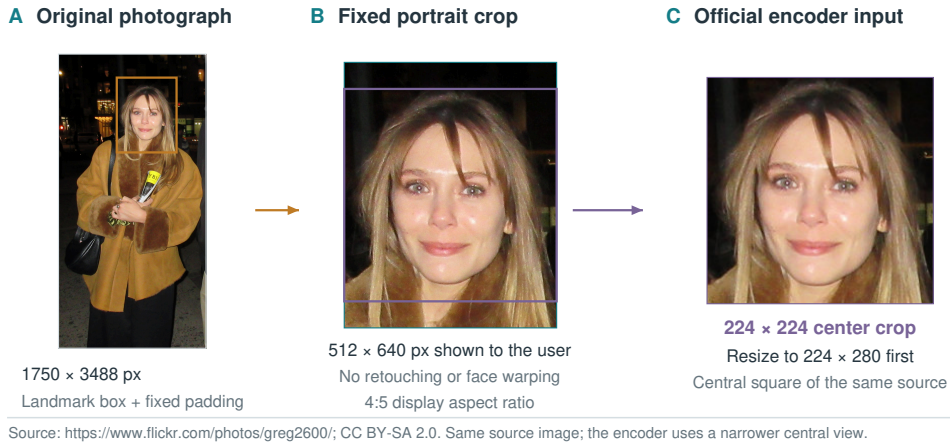


Figure 11: Exact preprocessing illustrated on a retained source. Orange marks the fixed face-centered crop in the original photograph; purple marks the encoder’s central square within the displayed portrait. Aspect ratios are preserved in this diagram. The human sees the complete canonical portrait, while CLIP receives its central square. The source record and license appear in Table 11.

Different photographs for personalization and evaluation

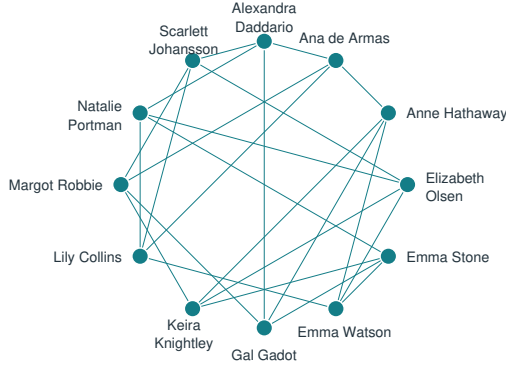
	Personalization: 2 portraits	Held-out panel: 2 new portraits
Anne Hathaway Illustrative candidate	 	 
Natalie Portman Illustrative candidate	 	 
Scarlett Johansson Illustrative candidate	 	 
Tongdak Cheonsa Prespecified target	No target portraits or target comparisons in training	 

Four distinct source groups per actress. The target has no training photograph. This design holds out photographs within known people and one target identity.

Figure 12: Photo-level separation for three illustrative actresses and the target. Every actress contributes two training and two test photographs. Tongdak Cheonsa contributes no training portrait; primary test photographs A and D are separate from alternate diagnostic photographs B and C. The full collection follows in Figures 15 and 16.

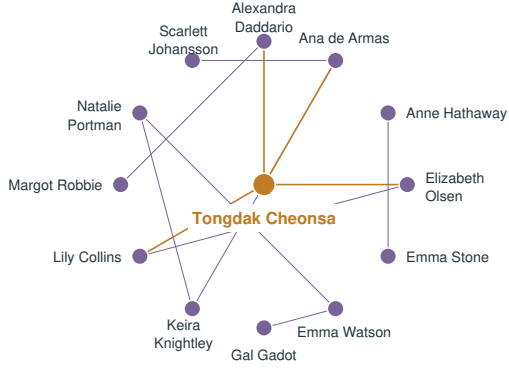
I Comparison schedule and frozen feature geometry

A Training comparisons: 24



12 people; 4 comparisons per person

B Held-out comparisons: 12

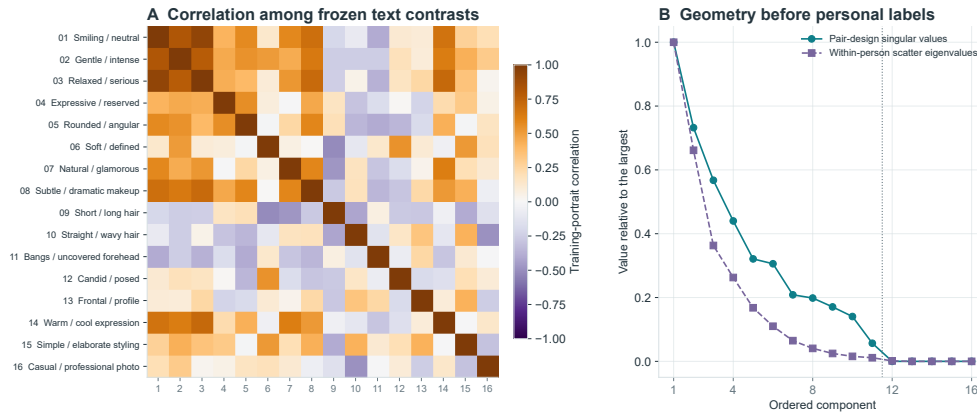


8 other pairs + 4 target pairs

Edges show the frozen question schedule, not observed preference or model predictions. The four reversed repeats are excluded from both graphs.

Figure 13: The complete scheduled comparison graphs. Edges denote questions, without expressing a response or a prediction. The training graph has degree four at every identity; the twelve held-out comparisons include four target edges. Names and phase membership are hidden in the human interface. The reversed-repeat trials are kept outside both graphs and outside fitting.

The sixteen prompt contrasts in Table 10 were chosen before responses. All models use their complete assigned representation; the feature display does not guide pruning or penalty selection. Figure 14 describes the primary encoder's actual frozen training features. Their correlations explain why a coefficient should be read conditional on the other probes. The 24 training differences have numerical rank 11, consistent with their construction from twelve panel means. Increasing the number of questions among those same twelve means would not remove this algebraic rank bound.



These descriptions use only the fixed training photographs. Correlated prompts are not independent attributes; the 24-pair design has rank at most 11. No dimension was removed and no model was selected using this display.

Figure 14: Structure of the fixed primary representation before personal labels. Left: Pearson correlations among sixteen text-contrast coordinates over the 24 training portraits. Right: singular values of the scheduled panel-difference matrix and eigenvalues of the within-person scatter, each divided by its own largest value. These are separate spectra, not a common-coordinate decomposition. No held-out photograph or response enters this description.

Table 10: The sixteen fixed text contrasts. Unless the full prompt is shown, the shared prefix is “a portrait of a woman with” and a period is appended. The protocol contains every prompt verbatim.

ID	Positive text	Negative text
01	a smiling face	a neutral expression
02	a gentle expression	an intense expression
03	a relaxed expression	a serious expression
04	an expressive face	a reserved expression
05	rounded facial features	angular facial features
06	soft facial features	sharply defined facial features
07	a natural appearance	a glamorous appearance
08	subtle makeup	dramatic makeup
09	short hair	long hair
10	straight hair	wavy hair
11	bangs	her forehead uncovered
12	a candid portrait of a woman	a posed portrait of a woman
13	a front-facing portrait of a woman	a profile portrait of a woman
14	a warm expression	a cool expression
15	simple styling	elaborate styling
16	a casual snapshot of a woman	a professional portrait of a woman

Coordinates subtract negative from positive text–image cosine similarity, then use training-portrait standardization. A coefficient is conditional on these correlated probes; the labels are not validated measurements of facial traits.

J Complete portrait plates

The plates display all 52 canonical portrait files, including the four target photographs. Names here identify the sources; they are absent from the comparison screen. Image suffixes join each portrait to the credit table and machine-readable manifest. Dates reproduce available source metadata at year precision; “n.d.” means the date is not recorded. The two target alternatives are diagnostic inputs, without new human labels.

The portrait collection

Every displayed file is fixed before choices; image IDs link to full credits and source hashes.

Alexandra Daddario	 train 04 2022	 train 01 2015	 test 02 2016	 test 03 2017
Ana de Armas	 train 04 2024	 train 02 2025	 test 01 2017	 test 03 2018
Anne Hathaway	 train 03 2024	 train 04 2014	 test 02 2011	 test 01 2023
Elizabeth Olsen	 train 03 2024	 train 02 2019	 test 01 2013	 test 04 2025
Emma Stone	 train 01 2010	 train 04 2024	 test 02 2018	 test 03 2024
Emma Watson	 train 04 2017	 train 02 2012	 test 01 2010	 test 03 2013

Figure 15: Complete collection, part 1 of 2. Each row contains two training and two held-out photographs of one actress. Colors indicate assigned input roles, not a preference or model outcome.

The portrait collection

Every displayed file is fixed before choices; image IDs link to full credits and source hashes.

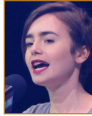
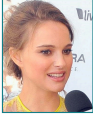
Gal Gadot	 train 03 2016	 train 02 2018	 test 04 2015	 test 01 2017
Keira Knightley	 train 03 2012	 train 02 2011	 test 01 2015	 test 04 2008
Lily Collins	 train 03 2019	 train 04 2013	 test 02 2015	 test 01 2013
Margot Robbie	 train 01 2013	 train 03 2018	 test 02 2016	 test 04 2026
Natalie Portman	 train 02 2011	 train 04 2015	 test 01 2019	 test 03 2025
Scarlett Johansson	 train 04 2017	 train 01 2012	 test 02 2013	 test 03 2016
Tongdak Cheonsa	 alternate B n.d.	 alternate C n.d.	 test A n.d.	 test D n.d.

Figure 16: Complete collection, part 2 of 2. The target row contains alternate sources B and C and primary sources A and D. All other rows contain two training and two held-out photographs.

K Photograph credits

Table 11: Photograph sources. Image suffixes match the complete portrait plates. Dates are source metadata and can be publication dates rather than camera-capture dates. Canonical crops are adaptations of the credited originals.

Candidate / image	Date	Credit	License	Source
Alexandra Daddario / 01	2015-04-04	Gage Skidmore	CC BY-SA 2.0	record
Alexandra Daddario / 02	2016-10-15	DannyB Photos	CC BY 2.0	record
Alexandra Daddario / 03	2017-05-17	Eva Rinaldi	CC BY-SA 2.0	record
Alexandra Daddario / 04	2022-09-27	Pansvalnac	CC BY-SA 4.0	record
Ana de Armas / 01	2017-07-22	Gage Skidmore from Peoria, AZ, United States of America	CC BY-SA 2.0	record
Ana de Armas / 02	2025-04-03	Gage Skidmore from Surprise, AZ, United States of America	CC BY-SA 2.0	record
Ana de Armas / 03	2018-03-01	Eric Longden	CC BY 3.0	record
Ana de Armas / 04	2024-09-07	Jay Dixit / WikiPortraits	CC BY-SA 4.0	record
Anne Hathaway / 01	2023-02-16	Harald Krichel	CC BY-SA 3.0	record
Anne Hathaway / 02	2011-02-27	Mingle MediaTV	CC BY-SA 2.0	record
Anne Hathaway / 03	2024-10-08	Jay Dixit	CC BY-SA 4.0	record
Anne Hathaway / 04	2014-04-26	Mingle MediaTV	CC BY-SA 2.0	record
Elizabeth Olsen / 01	2013-07-20	Gage Skidmore	CC BY-SA 2.0	record
Elizabeth Olsen / 02	2019-04-10	https://www.flickr.com/photos/greg2600/	CC BY-SA 2.0	record
Elizabeth Olsen / 03	2024-09-08	John Sears / WikiPortraits	CC BY-SA 4.0	record
Elizabeth Olsen / 04	2025-09-27	Kevin Paul	CC BY 4.0	record
Emma Stone / 01	2010-10-15	Promiflash.de	CC BY-SA 3.0	record
Emma Stone / 02	2018-09-13	Patrick L.	CC BY-SA 2.0	record
Emma Stone / 03	2024-05-18	Ariela Ortiz-Barrantes / WikiPortraits	CC BY-SA 4.0	record
Emma Stone / 04	2024-10-05	Bryan Berlin	CC BY-SA 4.0	record
Emma Watson / 01	2010-11-15	Joella Marano	CC BY-SA 2.0	record
Emma Watson / 02	2012-04-21	David Shankbone	CC BY 3.0	record
Emma Watson / 03	2013-05-29	Georges Biard	CC BY-SA 3.0	record
Emma Watson / 04	2017-06-12	Mars Movies	CC BY 3.0	record
Gal Gadot / 01	2017-11-14	Mark Neyman / Government Press Office of Israel	CC BY-SA 3.0	record
Gal Gadot / 02	2018-07-21	Gage Skidmore from Peoria, AZ, United States of America	CC BY-SA 2.0	record
Gal Gadot / 03	2016-07-23	Gage Skidmore from Peoria, AZ, United States of America	CC BY-SA 2.0	record
Gal Gadot / 04	2015-07-11	Gage Skidmore from Peoria, AZ, United States of America	CC BY-SA 2.0	record
Keira Knightley / 01	2015-02-02	The Standard	CC BY 3.0	record
Keira Knightley / 02	2011-09-10	GabboT	CC BY-SA 2.0	record
Keira Knightley / 03	2012-09-04	Paul Bird	CC BY 2.0	record
Keira Knightley / 04	2008-02-11	Chloe from Reading, England	CC BY 2.0	record
Lily Collins / 01	2013-03-30	Gage Skidmore	CC BY-SA 2.0	record
Lily Collins / 02	2015-04-23	We Day, Free The Children	CC BY-SA 4.0	record
Lily Collins / 03	2019-05-07	Montclair Film	CC BY 2.0	record
Lily Collins / 04	2013-07-19	Gage Skidmore	CC BY-SA 3.0	record
Margot Robbie / 01	2013-08-08	P. Lovell	CC BY-SA 2.0	record
Margot Robbie / 02	2016-07-23	Gage Skidmore from Peoria, AZ, United States of America	CC BY-SA 2.0	record
Margot Robbie / 03	2018-01-23	Eva Rinaldi from Sydney Australia	CC BY-SA 2.0	record
Margot Robbie / 04	2026-02-02	Ondine Goat	CC0	record
Natalie Portman / 01	2019-07-20	Gage Skidmore from Peoria, AZ, United States of America	CC BY-SA 2.0	record
Natalie Portman / 02	2011-02-26	LG Electronics	CC BY 2.0	record
Natalie Portman / 03	2025-05-16	Martin Kraft	CC BY-SA 4.0	record
Natalie Portman / 04	2015-02-07	Siebbi	CC BY-SA 3.0	record

Candidate / image	Date	Credit	License	Source
Scarlett Johansson / 01	2012-04-19	Paul Bird	CC BY 2.0	record
Scarlett Johansson / 02	2013-07-20	Gage Skidmore	CC BY-SA 3.0	record
Scarlett Johansson / 03	2016-09-11	GabboT	CC BY-SA 2.0	record
Scarlett Johansson / 04	2017-03-16	Dick Thomas Johnson; derivative work: Minerva97	CC BY 2.0	record
Tongdak Cheonsa / A	n.d.	User-supplied photograph	Unverified rights	–
Tongdak Cheonsa / B	n.d.	User-supplied photograph	Unverified rights	–
Tongdak Cheonsa / C	n.d.	User-supplied photograph	Unverified rights	–
Tongdak Cheonsa / D	n.d.	User-supplied photograph	Unverified rights	–

Full URLs, license links, revision identifiers, source groups, adult-at-capture checks, original hashes, and crop coordinates are retained in the machine-readable manifest. The target files are supplied by the requester; a Creative Commons license is not inferred for them.