FAITHFUL TEXTUAL EXPLANATIONS FOR INTERPRETABLE IMAGE AESTHETIC ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

In image aesthetic assessment (IAA), explanations typically rely on evaluative terms like *harmonious* or *nostalgic*. However, these terms describe the overall impression rather than specific regions of an image, and they lack ground-truth annotations for evaluation. Existing concept bottleneck models (CBMs) cannot be easily applied here because they rely on localized, verifiable visual concepts. Furthermore, selecting concepts based solely on their correlation with aesthetic ratings often fills explanations with paraphrases of "good image" or with descriptions of basic image content rather than aesthetic quality. To address these limitations, we introduce FaTeX, an interpretable IAA model that generates faithful, grounded, and image-specific textual explanations without requiring manual concept annotations. FaTeX calculates aesthetic scores by summing contributions from named concept pairs generated by a large language model (LLM) and scored using CLIP. It controls for image content, filters out redundant paraphrases, and uses concepts that apply both across datasets and within specific content categories. A gating step is designed to give no weight to level-specific concepts whose subject does not appear in the image, so that they drop out of the explanation. Evaluated across six single- and multi-genre datasets, FaTeX trails the best non-interpretable models by .035 SRCC on average (at most .058) and outperforms existing attribute-based approaches. In human and LLM-based evaluations measuring explanation quality and image identification, FaTeX produces better and more specific explanations than existing concept-based methods, and is on par only where a baseline draws on expert style annotations. Code and materials are available at `https://anonymous.4open.science/r/proj-FaTeX-E4FB`.

1 INTRODUCTION

Image aesthetic assessment (IAA) predicts how aesthetically pleasing a group of raters finds an image. It serves photo curation, creative support, and the evaluation of generative image models. For such applications, we argue that the interpretability of an IAA model matters as much as its predictive accuracy. Interpretability lets users check whether a model agrees with their own aesthetic values. This improves transparency, trust, and customizability (Liu & Wagemans, 2026). In creative support, it also makes it possible to give photographers and designers concrete feedback on their work.

Recent methods built on vision-language models (VLMs) (Zhou et al., 2025a;b) achieve high predictive accuracy, but their inference is a black box. Concept bottleneck models (CBMs) (Koh et al., 2020), which output a prediction as contributions of named concepts, were developed on tasks such as bird species classification, where a concept points to an object in the image, such as the color of a wing, and can be checked against annotations. Aesthetic assessment breaks these premises. Its vocabulary, such as "harmonious color scheme," is evaluative and emotional (Schindler et al., 2017); whether a word applies is not fixed by the objects in the image (Sibley, 1959); which words matter depends on the content (Li et al., 2023); and explanations have no ground truth. The CBM recipe of generating concepts with an LLM, scoring them with CLIP, and selecting them by correlation with the rating therefore fails: evaluative words collapse into paraphrases of "is this a good image?", and words that merely respond to the depicted content also enter the explanation. Existing interpretable IAA does not address this. Post-hoc explanations (Soydaner & Wagemans, 2026) need not reflect the model's computation, inherently interpretable methods (Liu & Wagemans, 2026) in practice rely on a few annotated attributes, and whether either helps people has rarely been measured.

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

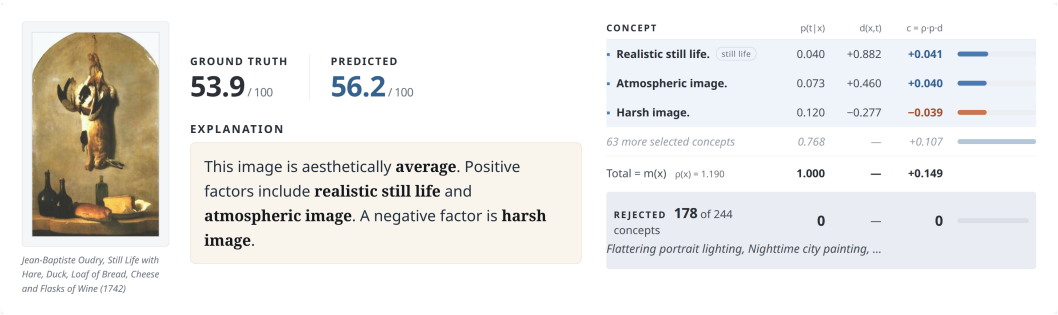

Figure 1: Example output of FaTeX. The contributions $c_t = \rho(x)\, p(t \mid x)\, d_\theta(x, t)$ sum to $m(x)$ (Eq. (1)); the explanation uses the top three by $|c_t|$. Rejected concepts get exactly zero weight.

Our goal is a model that, for IAA as a subjective task, produces language-based explanations useful to people. Language is our most familiar means of expression and supports intuitive understanding. It is also easy to use in applications such as feedback for creative support. We argue that such explanations need three properties. *Faithfulness*: the explanation is a decomposition of the prediction itself, so its arithmetic can be verified; it states how much each named concept contributed, not why the model weighted it so. *Fit to the image*: the features the explanation mentions are actually present in the image. *Specificity*: the explanation speaks about this particular image in vocabulary suited to its content, rather than offering generalities that hold for any image. The latter two concern the reader; with no ground truth for explanations, only people can judge them.

We propose FaTeX (**Fa**ithful **T**extual **eX**planations for interpretable image aesthetic assessment), which meets these requirements. FaTeX represents aesthetic quality as a *content-conditioned average* over a fixed pool of interpretable concepts. Each concept is an antonymic prompt pair ("Harmonious color palette." / "Clashing color palette."). The response $d(x, t) \in (-1, 1)$ tells which pole the image leans toward, and the weight $p(t \mid x)$ tells whether the concept applies to the image. The prediction is exactly this weighted sum, so it decomposes into the contributions of named concepts (*faithfulness*). Level-specific concepts judged not to apply receive exactly zero weight and are left out of the explanation (*fit to the image*). To build the concept pool, we score LLM-generated candidates with CLIP. We keep only pairs that are tied to the rating on their own even when content is controlled, and we remove pairs that paraphrase each other. Besides concepts that hold across the whole dataset, the pool contains concepts for each content level, e.g., landscapes or portraits (*specificity*). Building it requires no intermediate concept annotation. Filling a template with the top concepts yields a natural-language explanation without giving up *faithfulness* (Fig. 1).

Our contributions are threefold. **First**, for IAA, a subjective task, we propose FaTeX. It produces textual explanations that are faithful, fit the image, and are specific to it. **Second**, on six datasets spanning single- and multi-genre collections, we compare the predictive performance of FaTeX with state-of-the-art methods. Ablations examine the contribution of each component and reveal the trade-off between interpretability and predictive performance. **Third**, for a task without ground-truth explanations, we measure explanation quality directly with both human evaluation and LLM-as-a-judge (Zheng et al., 2023). Both answer the same questions under the same presentation. We therefore also measure how far an LLM can stand in for human raters when judging explanations.

2 Related Work

Interpretable AI. Explanation methods fall broadly into two groups. Post-hoc explanations approximate a trained model from the outside (Lundberg & Lee, 2017; Selvaraju et al., 2017). Inherently interpretable models build the explanation into the structure of the prediction (Rudin, 2019). The former do not guarantee *faithfulness*, i.e., that the explanation reflects the model's actual computation (Jacovi & Goldberg, 2020). Natural-language explanations from LLMs can also stay plausible while diverging from the real reasons for a prediction (Turpin et al., 2023). The concept bottleneck model (CBM) (Koh et al., 2020) is the representative of the latter. It outputs a prediction as a linear sum of the values of named concepts. LLMs and CLIP have since removed the need for concept

annotation (Oikarinen et al., 2023; Yang et al., 2023). However, these models have been validated only on tasks with concept annotations or well-defined classes (Sec. 1). Explanations are also ultimately used by people. Human explanations are contrastive and selective (Miller, 2019), and interpretability is ultimately best evaluated with humans (Doshi-Velez & Kim, 2017).

Interpretability in IAA. As noted in Sec. 1, VLM-based methods are black boxes in how they predict. Most attempts to explain them build the explanation outside the predictor. Some apply SHAP to regressors that predict the score from aesthetic attributes or from responses to antonymic prompt pairs (Soydaner & Wagemans, 2026; Viriyavisuthisakul et al., 2025). Others linearly probe the internal representations of a VLM (Ryu & Yanaka, 2026) or generate critiques (Li et al., 2025a; Cao et al., 2026). For textual explanations in particular, nothing ties the text to the predicted score. C2J (Liu & Wagemans, 2026) places a sparse linear read-out on concept directions obtained from positive and negative example images of each attribute, so in practice its vocabulary is limited to a few annotated attributes. PPA (Hara & Otsuka, 2026) mixes the responses to 260 LLM-generated antonymic prompt pairs with per-image weights; it yields per-concept contributions, but targets controlling the evaluation criteria through prompts. Neither measures whether its explanations help people.

Our position. We require explanations to be both faithful to the prediction and preferable to people. We measure the latter directly, through human evaluation of *fit to the image* and *specificity*. Methodologically, FaTeX is a CBM. It outputs the prediction as a sum of contributions of named concepts, and it obtains the concepts with LLMs and CLIP. The difference lies in how the vocabulary is built. The vocabulary of IAA is evaluative and does not point to objects (Sec. 1). Content can affect the rating, but we explain with words for differences within a content level rather than with the content itself. FaTeX therefore selects candidates by their association with the rating once content is controlled, and removes paraphrases. It builds a vocabulary per content level and explains each image with only the concepts that apply to it.

3 METHOD

Each component of FaTeX serves one property of Sec. 1. Formulating the prediction as a sum of concept contributions gives *faithfulness* (Sec. 3.1). Building the concept pool per content level gives *specificity* (Sec. 3.2). Weighting by applicability to the image gives *fit to the image* (Sec. 3.4).

3.1 PROBLEM FORMULATION

The training data are a set $\mathcal{D} = \{(x, y)\}$ of images paired with their mean rating by a group of raters. We assume y is normalized to $[0, 1]$ by the bounds of the rating scale. The goal of IAA is to predict y for unseen images. We represent aesthetic quality as a *content-conditioned average* over a fixed pool $\mathcal{T}$ of interpretable aesthetic concepts. Each concept $t \in \mathcal{T}$ is expressed as an antonymic prompt pair $t = (t_{\text{pos}}, t_{\text{neg}})$ that names the two poles of one concept. For an image x, we define

$$m(x) \;=\; \rho(x) \sum_{t \in \mathcal{T}} p(t \mid x)\, d_\theta(x, t) \;\in\; (-\rho_{\max}, \rho_{\max}). \tag{1}$$

Here $d_\theta(x, t) \in (-1, 1)$ is a *signed* response. It indicates which pole of concept t the image x leans toward, and by how much (Sec. 3.3). $p(t \mid x)$ is a distribution that weights how much concept t contributes to the aesthetic judgment of this image x. $\rho(x) \in (0, \rho_{\max})$ is a per-image learned *magnitude* (Sec. 3.4). The weights p set the *ratios* among concepts, and ρ sets the size of their sum. Both are non-negative, so only the response d_θ can be negative. The per-concept terms $c_t(x) = \rho(x)\, p(t \mid x)\, d_\theta(x, t)$ satisfy $\sum_t c_t(x) = m(x)$ exactly. The prediction therefore decomposes exactly into named contributions; this is the sense of *faithfulness* we claim. Why each weight takes its value is not itself explained (Appendix K). Concepts with $c_t > 0$ and $c_t < 0$ state "why it is aesthetic" and "why it is not" in the same units. The prediction is not m itself. It is obtained through an affine map to the rating scale, $\hat{y}(x) = \bar{y} + s\, m(x)$. The constants $\bar{y}$ and s are *determined by the split* and are neither learned nor tuned (Appendix E). Only θ is learned, with the mean squared error between $\hat{y}$ and y as the objective.

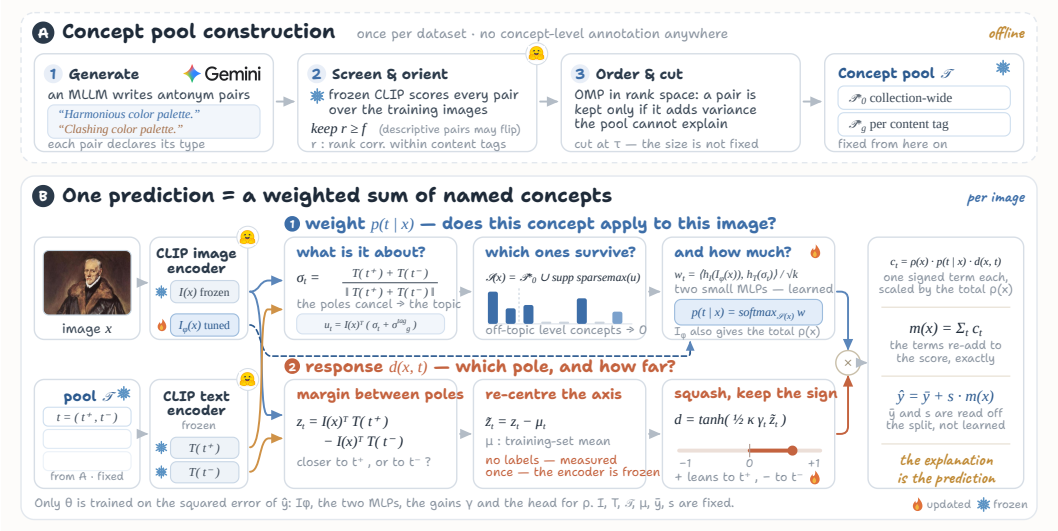

Figure 2: Overview of FaTeX. **A**: concept pool construction, once per dataset. **B**: prediction for one image: the weight $p(t \mid x)$ and response $d(x, t)$, scaled by $\rho(x)$, give terms c_t summing to $m(x)$.

3.2 Text Prompt Sampling

The concept pool $\mathcal{T}$ of Eq. (1) is constructed for each dataset in three steps. We generate a candidate set (Step 1). We then score the training images with a frozen CLIP. This decides whether each pair is associated with the rating *on its own*, and *in which direction* (Step 2). Finally, we order the remaining pairs by *how much of what they explain overlaps*, and truncate them into a low-redundancy pool (Step 3). Most datasets come with an L-level *content tag* G describing what an image shows (e.g., landscape or portrait; Sec. 4.1). We write $\mathcal{X}_g$ ($n_g = |\mathcal{X}_g|$) for the set of images at level g. Where no tags are distributed, we assign them ourselves (AADB; Appendix D). $g(t) = 0$ denotes a pair that applies to the whole collection, and $g(t) = g$ a pair that applies only within level g. We summarize each step below. The full rules are in Appendix B.

Step 1: Candidate generation. Candidates are drawn with Gemini-3.5 Flash from three sources. Source A consists of 40 *dataset-agnostic attribute seeds*. Source B consists of 90 *dataset-agnostic general candidates*. Both have $g(t) = 0$. They are generated once without example images and reused across all datasets (Hara & Otsuka, 2026), and their final noun is fixed to "image". Source C consists of 130 *content-tag-specific* candidates for each level g ($g(t) = g$), drawn with images of that level attached. All three sources share one generation template (Fig. S1). Each pair is returned with its declared type. An *evaluative* pair is one whose word meaning alone makes one pole better. A *descriptive* pair is one where the images decide the preferred pole. $\mathcal{T}_{\text{ev}}$ denotes the evaluative pairs.

Step 2: Filtering, orientation, and eligibility. We first filter the candidates by rules on the form of a pair and on duplicates. Duplicates are removed by dropping pairs that reuse a clause already taken. Each remaining pair is represented by the per-prompt projection $z(x, t) = I(x)^\top T(t_{\text{pos}}) - I(x)^\top T(t_{\text{neg}})$, where I and T are the CLIP image and text encoders. We define $r_g(t)$ as the *partial rank correlation between the projection z and the rating with the content tag G as a covariate*, measured over the image set $\mathcal{X}_g$. A candidate is thus judged by its contribution *within* the same content. Candidates with $g(t) = 0$ are measured over all of $\mathcal{X}$, and those with $g(t) = g$ within $\mathcal{X}_g$:

$$r_g(t) = \operatorname{corr}\big(z_t^{\perp},\ y^{\perp}\big)\big|_{\mathcal{X}_g},$$
$$\mathcal{T}_{\text{elig}} = \Big\{ t \ :\ r_{g(t)}(t) \ \geq\ f_{g(t)} \ \ (t \in \mathcal{T}_{\text{ev}}),\quad \big|r_{g(t)}(t)\big| \ \geq\ f_{g(t)} \ \ (t \notin \mathcal{T}_{\text{ev}}) \Big\}. \tag{2}$$

Here $z_t^{\perp}$ and $y^{\perp}$ are the residuals after converting to ranks and removing the component explained by the content tag. The floor f_g is the larger of the effect-size floor $f = 0.10$ and a sample-size guard (Appendix A). The test is two-sided only for descriptive pairs. For these, the polarity the generator assigned is not part of the name. If the data disagree, we swap the two clauses. For evaluative pairs,

word meaning already fixes the poles. A value $r_{g(t)}(t) < 0$ is then a contradiction, and the pair is rejected. Every retained pair thus satisfies $r_{g(t)}(t) \geq 0$, and t_{pos} names the high-rated side.

Step 3: Orthogonal ordering and thresholding. The eligible candidates are ordered by orthogonal matching pursuit (OMP) in rank space. At each step, every remaining candidate is scored by its *residual relevance*. This is the absolute correlation between the rating and the component of the candidate's projection orthogonal to the columns already selected. The highest-scoring candidate is selected. Redundant paraphrases thus rank low without any similarity threshold. We keep the ordering up to just before the first candidate whose residual relevance falls below $\tau = 0.01$. The pool size therefore follows from the threshold rather than being fixed in advance. Ordering is done in two stages. The first stage orders the eligible candidates with $g(t) = 0$ over the whole of $\mathcal{X}$ to obtain $\mathcal{T}_0^\star$. The second stage handles each level g. It restricts the eligible candidates with $g(t) = g$ to $\mathcal{X}_g$. It then orders them *after projecting out the span of the columns of $\mathcal{T}_0^\star$ restricted to $\mathcal{X}_g$*, to obtain $\mathcal{T}_g^\star$. A level-specific concept is thus admitted only if, within its level, it explains variance that the collection-wide concepts *have not yet explained*. The pool is $\mathcal{T} = \mathcal{T}_0^\star \cup \bigcup_{g=1}^{L} \mathcal{T}_g^\star$ (sizes in Table S1).

3.3 Per-Concept Response

Following CLIP-IQA (Wang et al., 2023), we match an image against a pair t with the *frozen* CLIP image encoder I and text encoder T. We take a softmax over the similarities to the two clauses. This is the sigmoid of their difference. Recentering it so that the midpoint of the pair is 0 gives

$$\begin{gathered} d_\theta(x,t) \;=\; 2\,\mathrm{sigmoid}\big(\kappa\,\gamma_t\,\tilde{z}(x,t)\big) - 1 \;=\; \tanh\Big(\tfrac{1}{2}\kappa\,\gamma_t\,\tilde{z}(x,t)\Big) \;\in\; (-1,1), \\ \tilde{z}(x,t) \;=\; z(x,t) - \mu_t, \end{gathered} \tag{3}$$

where z is the projection of Step 2, κ is a fixed temperature, and $\mu_t = \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}[z(x,t)]$ is the per-concept center. μ_t is a constant determined by the frozen I and T. Subtracting it makes the sign of $d_\theta(x,t)$ say whether the image lies on the t_{pos} or the t_{neg} side of the training-set mean on that axis. $\gamma_t = \exp(a_t) > 0$ is a learnable per-concept gain. It is the only freedom on the response side, and it changes only how strongly each concept pushes relative to the others. With $\gamma_t \equiv 1$, the response equals an affine transform of the CLIP-IQA score (Appendix A).

3.4 Concept Selection by Applicability Weighting

We build the weight $p(t \mid x)$ in two stages. The first stage selects a *set* of concepts by whether they *apply* to this image. It is meant to give no weight to level-specific concepts whose subject the image does not show; those it rejects get exactly zero weight (Eq. (4)). The second learns the allocation *within* that set: which concepts *influence* the judgment, and how much (Eq. (5)).

Stage 1: Which concepts apply. Take a mountain landscape and the pair "Expressive portrait." / "Blank portrait.". With no person in the picture, asking which pole the image is closer to is meaningless. We therefore build the first-stage weight on *applicability*, i.e., how well the image matches the subject the concept names. Explanations thus favor concepts that apply to the image (*fit to the image*). We ask this only of the level-specific concepts $\mathcal{T}_g^\star$. The collection-wide concepts $\mathcal{T}_0^\star$ have passed the eligibility test over all of $\mathcal{X}$ with the content tag as a covariate (Sec. 3.2, Step 2). As a design choice, we do not ask again per image whether they apply. The construction follows from how pairs are written. The two clauses of t share the final noun and differ only in the polarity words (Sec. 3.2). The sum $T(t_{\text{pos}}) + T(t_{\text{neg}})$ therefore cancels the polarity and leaves the noun. We normalize it to unit length and call it the *subject direction* σ_t.

For a level-specific concept, we add the name of the selected level, read as a sentence, $\sigma_g^{\text{tag}} = T(\text{"An image of } g\text{."})$, to the subject direction and measure applicability on the sum (Appendix A). The weights project the applicability along σ_t onto the probability simplex. For this normalization we use sparsemax (Martins & Astudillo, 2016; Du et al., 2026) rather than softmax. Sparsemax sets the weights of the concepts it rejects to exactly zero.

$$\begin{gathered} \mathcal{S}(x) \;=\; \mathcal{T}_0^\star \;\cup\; \mathrm{supp}\;\mathrm{sparsemax}\big(u(x)|_{\mathcal{T}_{\text{lev}}^\star}\big), \qquad u_t(x) = I(x)^\top\big(\sigma_t + \sigma_{g(t)}^{\text{tag}}\big), \\ \mathrm{sparsemax}(v) \;=\; \operatorname*{arg\,min}_{p\in\Delta^{|\mathcal{T}_{\text{lev}}^\star|-1}} \|p - v\|_2^2, \end{gathered} \tag{4}$$

Table 1: GIAA performance on the test split.

			Single-genre						*Multi-genre*					
			LAPIS		APDDv2		PAMELA		AADB		TAD66K		PARA	
Method	Venue	Backbone	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
MLP	—	CLIP-B/16	.856	.858	.847	.877	**.806**	**.838**	.692	.687	.473	.505	.913	.942
		UniQA-B/16	.850	.842	.844	.863	.799	.835	.710	.700	**.480**	**.508**	.912	.942
NIMA	TIP '18	CLIP-B/16	**.863**	**.863**	—	—	—	—	—	—	—	—	.913	.942
		UniQA-B/16	.860	.861	—	—	—	—	—	—	—	—	**.917**	.945
		Incept.-v3†	.796	.799	—	—	—	—	—	—	—	—	.845	.898
Charm	CVPR '25	DINOv2-S/14†	.832	.838	.832	.857	.748	.795	.655	.675	.431	.458	.907	.939
Gamma-T	MM '25	CLIP-B/16	.859	.862	**.850**	**.883**	.790	.821	.694	.704	.454	.480	.916	**.945**
		UniQA-B/16†	.855	.858	.843	.868	.789	.810	**.712**	**.727**	.463	.489	.917	.943
PPA-T	CVPR '26	CLIP-B/16†	.853	.847	.737	.824	.787	.795	.648	.635	.452	.421	.908	.912
		UniQA-B/16	.852	.844	.732	.790	.799	.725	.673	.608	.459	.435	.907	.902
C2J	Pre '26	CLIP-B/16	.818	.815	—	—	—	—	.591	.585	—	—	.799	.833
		UniQA-B/16	.818	.816	—	—	—	—	.628	.631	—	—	.818	.855
		CLIP-RN50†	.796	.792	—	—	—	—	.606	.604	—	—	.777	.809
FaTeX (ours)	—	CLIP-B/16	.846	.845	.808	.837	.773	.798	.634	.651	.437	.462	.890	.924
		UniQA-B/16	.839	.839	.799	.828	.767	.806	.654	.669	.449	.480	.891	.929
Δ to column best			−.017	−.018	−.042	−.046	−.033	−.032	−.058	−.058	−.031	−.028	−.026	−.016

Each value is from the run with the median SRCC over three seeds; PLCC is from the same run. †: backbone of the original paper. —: undefined on the dataset (NIMA needs a discrete rating scale; C2J needs attribute annotations). **Bold underline** / underline: best / second best in the column, each method at its best backbone for that column; ties at three decimals are ranked by the unrounded values. Δ to column best: FaTeX's better backbone for that column minus the column best.

where sparsemax is taken over $\mathcal{T}^{\star}_{\text{lev}} = \bigcup_{g=1}^{L} \mathcal{T}^{\star}_{g}$. The image itself thus decides which level's concepts apply; its content tag is not given at inference. Concepts of $\mathcal{T}^{\star}_{0}$ get no tag term ($\sigma_0^{\text{tag}} = 0$).

Stage 2: Which matter, and how much. The weights that remain in $\mathcal{S}(x)$ measure how well a concept *matches* the image. They do not measure its influence on the aesthetic judgment. We therefore learn the allocation within the set separately:

$$p(t \mid x) \;=\; \begin{cases} \dfrac{\exp w_t(x)}{\sum_{t' \in \mathcal{S}(x)} \exp w_{t'}(x)} & (t \in \mathcal{S}(x)), \\ 0 & \text{otherwise}, \end{cases} \qquad w_t(x) \;=\; \frac{\langle h_I(I_\phi(x)),\, h_T(\sigma_t) \rangle}{\sqrt{k}}, \tag{5}$$

where h_I and h_T are two-layer MLPs with k-dimensional outputs (initialization in Appendix A). I_ϕ is a copy of the image encoder I, initialized from I and trained. It also gives the magnitude of Eq. (1), $\rho(x) = \rho_{\max} \operatorname{sigmoid}\big(v^\top I_\phi(x) + b\big)$. The response and the applicability read the frozen I, so training changes only the weights.

4 Experiments

4.1 Setup

Datasets. We evaluate on six public datasets that differ in both rating scale and content. LAPIS, APDDv2 and PAMELA all consist of art images. They measure how well fine aesthetic differences within one genre can be told apart. AADB, TAD66K and PARA gather images of any genre. They measure aesthetic assessment that holds across content. The composition, the content tags attached to each dataset, and their provenance and rating procedures are detailed in Appendix D (Table S4). We follow the public splits where they exist; splits are disjoint at the image level.

Implementation details. We use two backbones. One is CLIP (Radford et al., 2021), the standard VLM. The other is UniQA (Zhou et al., 2025a), which continues the pre-training of that CLIP on aesthetic and image quality assessment data. We update only I_ϕ, the copy of the image encoder read by the importance and the magnitude (Sec. 3.4). The parameters updated, the head sizes, the response temperature, the bound on the magnitude, the input pipeline, the optimizer and the learning-rate search are in Appendix E. Resizing, cropping and augmentation are shared by all trained methods.

Baselines and evaluation. We compare with six methods: **MLP**, a non-interpretable head; **NIMA** (Talebi & Milanfar, 2018); **Charm** (Behrad et al., 2025); **Gamma-T** (Zhou et al., 2025b); **PPA-T** (Hara & Otsuka, 2026); and **C2J** (Liu & Wagemans, 2026). Only PPA-T and C2J can split

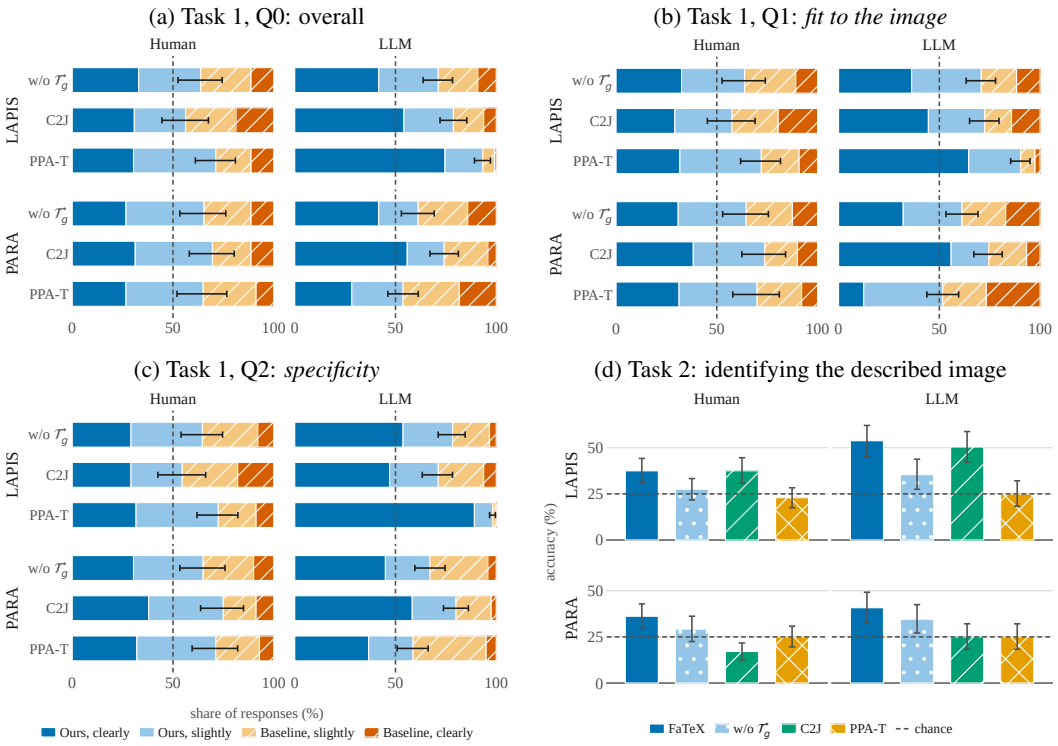

Figure 3: Interpretability on LAPIS and PARA. (a)–(c) Task 1: preference for FaTeX over each baseline, with its 95% CI. (d) Task 2: accuracy of identifying the described image (chance 25%).

their predictions into per-concept contributions. We add no CBM from outside IAA: the concept-based methods here already share its components, and what IAA changes is how concepts are defined, selected and evaluated. Appendix H details this and the settings of each method. All methods that involve training get the same budget. Their learning rates are searched by one common procedure, whose first trial is the setting of the original paper. Model selection uses only the argmax of the validation SRCC, except for C2J, which freezes the encoder. On the test set we measure **SRCC** and **PLCC**. Of three runs that differ only in the seed, we report the run with the median SRCC. PLCC is taken from that same run.

4.2 Results

Predictive performance. Table 1 shows the results on the six datasets. FaTeX does not reach the best method of any column (bold underline). Still, the SRCC gap between the best backbones is .035 on average (at most .058). The best methods are Gamma-T, MLP and NIMA, with two columns each. None of them can split its prediction into per-concept contributions. Of the two baselines that can, PPA-T is more accurate than FaTeX on five of the six datasets (by .007 to .026), and FaTeX leads only on APDDv2 (+.071); we return to this trade-off in Sec. 5. FaTeX outperforms C2J, the other one, on all three datasets where C2J is defined (+.026 to +.073). The gap to the best method is largest on AADB (−.058), the only dataset whose content tags we assigned with an LLM (Appendix D).

Interpretability. We measure explanation quality with human evaluation and have an LLM answer the same items. We measure two things. One is whether people judge an explanation to be better (*explanation quality*). The other is whether an explanation singles out its image among others (*discriminability*). *Specificity* is measured from two sides. One question of Task 1 (Q2) measures it through human judgment, and Task 2 measures it with a definite correct answer. The datasets are the single-genre LAPIS and the multi-genre PARA, the only two on which every compared method is defined and reaches an SRCC above .80 (Table 1). The baselines are PPA-T and C2J, which can produce concept-based explanations (Sec. 4.1). We also compare with FaTeX without the level-specific concepts, whose pool is $\mathcal{T}_0^\star$ alone (w/o $\mathcal{T}_g^\star$ below). The comparison with the former two

measures explanation quality relative to existing methods. The comparison with the latter measures the contribution of the level-specific concepts. Explanations are produced by a rule common to all methods. Methods thus differ only in where the vocabulary comes from and how concepts are selected (Fig. 1, center). Differences in predictive accuracy between methods should not leak into the evaluation of explanations. We therefore restrict the evaluation to images on which every compared method predicts close to the ground truth. From these, we select 120 images per dataset, stratified into the top, middle and bottom bands of the ground truth. Raters are recruited on Prolific. Each image is rated by 2 raters per pair in Task 1 and per method in Task 2. The LLM (Claude Opus 5) is given the same images, explanations and questions as the raters, in the same presentation. It answers again with the presentation swapped, and we report the mean of both. For Task 1 we report the preference rate for FaTeX. For Task 2 we report the difference in accuracy (FaTeX $-$ baseline). Both are given per pair and question, with 95% CIs over images. See Appendix I for details.

Task 1: Pairwise comparison of explanations. We show one image with the explanations of two methods side by side. Raters answer three questions on a four-point scale from "clearly A" to "clearly B", with no neutral option (Kayser et al., 2021). Q0 asks which is the better explanation overall. Q1 asks whether the features mentioned are actually in the image (*fit to the image*). Q2 asks whether the description could apply to any image or is about this image (*specificity*). The pairs are FaTeX versus PPA-T, versus C2J, and versus w/o $\mathcal{T}_g^\star$. Fig. 3(a)–(c) shows the results. On PARA, raters chose FaTeX's explanation for all three pairs and all three questions (preference 64.6–75.0%, every 95% CI excluding 50%). On LAPIS, too, FaTeX was chosen on all three questions against PPA-T and against w/o $\mathcal{T}_g^\star$. The level-specific concepts thus raise explanation quality. Q2 alone could favor an explanation that merely names the content; Task 2 rules this out. Only against C2J on LAPIS was there no significant difference (54.6–57.5%). A likely reason is that the 26 annotated style names give C2J highly specialized explanations that name a painting's art movement ("Negative factors include Fauvism style and Expressionism style."), which raters found as convincing as FaTeX's. Task 2 shows the same result. The LLM mostly agreed with the raters but disagreed in two places. Against C2J on LAPIS, it chose FaTeX on all three questions. Against PPA-T on PARA, it showed no significant difference on Q0 and Q1. The latter arises from the top band of PARA. There PPA-T returns "sharp image, focused foreground image and stable image", which holds for almost any highly rated photo. The LLM judged that this loses nothing on *fit to the image*, but people did not regard it as a good explanation. On Q0, humans and the LLM chose the same side in 68.8% of cases on LAPIS and 63.2% on PARA. Results by band are in Table S16.

Task 2: Identifying the image from its explanation. We show one explanation and four images and ask which image it describes (chance 25%). The other three images are drawn from images with the same content tag as the correct one and a similar rating. The content or the rating level alone therefore hardly leads to the answer (an example is in Fig. S5, Appendix I). The sets of four images are shared by all methods, and only the explanations differ. Fig. 3(d) shows the results, and Table S17 in the appendix gives the differences between methods. From FaTeX's explanations, raters identified the correct image 37.5% of the time on LAPIS and 36.2% on PARA. In both cases the 95% CI lies above chance. PPA-T stays at chance on both datasets (22.9%, 25.0%), and C2J falls below it on PARA (17.1%). On LAPIS C2J matches FaTeX at 37.5%. As in Task 1, C2J on LAPIS can use expert style labels from the annotations (e.g., Fauvism); FaTeX equals it with only the coarse genre tag (e.g., portrait), which merely organizes its vocabulary. The difference from w/o $\mathcal{T}_g^\star$ is significant on LAPIS ($+10.0$ points). On PARA it is $+7.1$ points, and its CI includes 0. Methods whose vocabulary is fixed or limited to annotations return the same explanation for many images. Among the 40 images of a ground-truth band, PPA-T and C2J have bands with only 1–3 distinct explanations. FaTeX produces 35–39. The LLM is more accurate than humans (FaTeX: LAPIS 53.8%, PARA 40.8%). Yet the direction and significance of the differences between methods agreed with humans in all six comparisons. Q2 of Task 1 and Task 2 point the same way: FaTeX's explanations are more specific than those of existing methods.

Faithfulness checks. An explanation of FaTeX consists of the selected set $\mathcal{S}(x)$ and the contribution $c_t(x)$ of each concept. We test whether the prediction rests on it by editing the set without retraining. The weights are recomputed with the softmax of Eq. (5) over the edited set, and nothing else, including $\rho(x)$, changes. We use the FaTeX runs of Table 1 (Appendix E) and give PARA in parentheses after LAPIS. Deleting concepts in descending order of $|c_t(x)|$ lowers SRCC steeply: with half gone, it falls from .846 to .543 (.890 to .654). Deleting the same half at random costs only .032 (.033), and the smallest first barely matter (Fig. 4). Conversely, reading out with only the top 10% (9.1 of 85.7

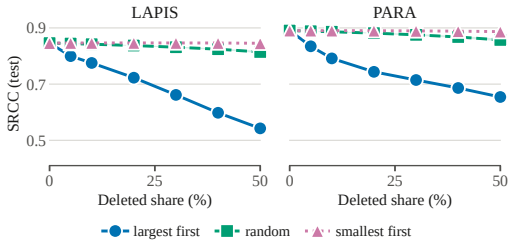

Figure 4: Test SRCC vs. share of $\mathcal{S}(x)$ deleted; random is the mean of 20 draws.

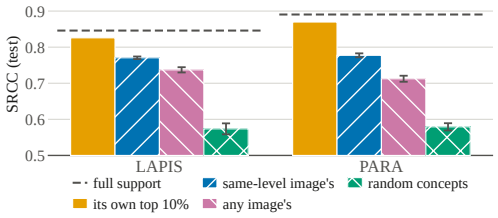

Figure 5: SRCC when the top 10% is taken from another image or at random. Dashed: full $\mathcal{S}(x)$; bars: SD, 20 draws.

concepts; 10.9 of 104.6) costs only .020 (.020). The three concepts that the explanation names hold 31.5% (34.0%) of $\sum_t |c_t(x)|$ (Appendix F). An equal-size set borrowed from another image lowers SRCC even when that image is at the same level, and a uniform draw from the pool brings it down to .574 (.579; Fig. 5). The concepts with the largest contributions thus carry most of the prediction, and applicability picks a set that differs from image to image.

Ablations. Retraining with each weight component removed (Appendix F) shows that the prediction rests on the allocation within the set; removing the gate does not lower it. Without the gate, however, 60–70% of the top three concepts in an explanation are concepts of other levels that do not fit the image content (under 20% for FaTeX). The gate is needed not for prediction but to fit the explanation to the image. Likewise, content control in building the pool leaves the prediction unchanged. On LAPIS, where content and rating are strongly linked, it keeps paraphrases of "what is depicted" out of the explanations; without it, they enter those of 48% of the test images.

5 Discussion and Conclusion

To our knowledge, FaTeX is the first IAA model whose textual explanations are required to be faithful to the prediction and are verified with people. Content-controlled selection, a vocabulary per content level and a per-image selection by applicability extend concept bottleneck models to a task whose vocabulary is subjective and whose explanations have no ground truth.

Built solely from concept contributions, FaTeX stays within .035 of the best methods on average and outperforms C2J. PPA-T, the other decomposable baseline, is more accurate on five datasets, but its top concepts barely change across images, raters preferred FaTeX's explanations on every question, and its explanations identified the image only at chance. Applications that need explanations thus have little reason to settle for a black box. An explanation states how much the prediction relied on which concepts; not the aesthetic reasons of the image, nor why the model weighted each concept as it did. The level-specific concepts raised *specificity*, clearly on LAPIS and as a trend on PARA, and a concept of another level is not necessarily wrong for the image, as with a plants concept on a PARA portrait of a child holding flowers (Appendix F). Since generic explanations from a fixed vocabulary score well on *fit to the image*, explanations of subjective judgments must also be measured for *specificity*; that humans and an LLM agreed on the differences between methods suggests part of this can be delegated. As limitations, VLMs miss terms like the rule of thirds, the pool stays fixed once built and is built per dataset, the level-specific vocabulary requires content tags, leakage through the weights has not been tested rigorously, and human evaluation covers two datasets (Appendix K).

An antonymic-pair concept names what lowered the rating and, through its opposite pole, a direction for improvement. As the prediction is a sum of named contributions, users can check it against their own values concept by concept, or personalize it by removing or reweighting concepts. This is a step toward applications where people and models examine and align each other's judgments.

AI Use Statement

LLMs enter this work in two ways. *As part of the method and its evaluation*, Gemini-3.5 Flash generates the candidate concepts and assigns the content tags of AADB (Sec. 3.2, Appendix D),

and Claude Opus 5 serves as the LLM judge (Sec. 4.2); all prompts are given in Appendix C. *As research assistants*, the authors used Claude Opus 5 and Claude Opus 5.5 to implement the method, the baselines, the experiments and the figure scripts, and to draft, translate into English and edit the manuscript for readability. The research questions, method design, choice of experiments and conclusions were made by the authors. The authors reviewed and verified all code, text and references produced with these tools and take full responsibility for the content of the paper.

Ethics Statement

All authors have read and adhere to the ICLR Code of Ethics. The human evaluation was approved by the ethics review board of the authors' institution (details withheld for anonymous review). Raters were recruited on Prolific, gave informed consent, and were paid at the equivalent of £10 per hour, including those excluded by the attention checks, whose criteria were fixed in advance. The recruitment notice stated that paintings containing nudity would be shown, and sexually explicit works were withheld (Appendix I). No personally identifying information was collected, and the responses we release are anonymized. All datasets are public and are used under their original terms. Aesthetic ratings reflect the preferences of each dataset's rater population, and a model trained on them can inherit their biases. Explanations of FaTeX state what the prediction relied on, which makes such biases easier to inspect, but they do not remove them. The authors declare no conflicts of interest.

Reproducibility Statement

An anonymized repository at `https://anonymous.4open.science/r/proj-FaTeX-E4FB` provides the code for building the concept pool, training, evaluation and explanation generation; the generated candidate concepts, concept pools and the content tags we assigned; the data splits; and the materials and anonymized responses of the human and LLM evaluations. The images and annotations of the six public datasets are not redistributed; a script rebuilds the splits from their official releases. The trained models will be released upon acceptance because of their size. The concept-pool construction is described in Sec. 3.2, with its full rules and all prompts in Appendices B and C. The datasets, splits and content tags are in Appendix D; the training budget, hyperparameters, learning-rate search and run selection in Sec. 4.1 and Appendix E; the baseline settings in Appendix H; and the design of the human and LLM evaluations, including question texts and exclusion criteria, in Appendix I.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for CommonsenseQA: New Dataset and Models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3050–3065, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.238. URL `https://aclanthology.org/2021.acl-long.238/`.

Fatemeh Behrad, Tinne Tuytelaars, and Johan Wagemans. Charm: The missing piece in vit fine-tuning for image aesthetic assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 7815–7824, June 2025.

Leo Breiman, Jerome Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1st edition, 1984. doi: 10.1201/9781315139470.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajuin Yao, and Yihao Liu. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15313–15322, June 2026.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:11319376.

Xingbo Du, Qiantong Dou, Lei Fan, and Rui Zhang. Flexible concept bottleneck model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Mateo Espinosa Zarlenga. In defense of information leakage in concept-based models, 2026. URL https://arxiv.org/abs/2606.10669.

Takayuki Hara and Yuya Otsuka. Probabilistic prompt adaptation for unified image aesthetics and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37236–37246, June 2026.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York, 2nd edition, 2009. URL https://link.springer.com/book/10.1007/978-0-387-84858-7.

Marton Havasi, S. Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems 35*, 2022. URL https://api.semanticscholar.org/CorpusID:258509699.

Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 942–948. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/132. URL https://doi.org/10.24963/ijcai.2022/132. Main Track.

ITU-R. Methodologies for the subjective assessment of the quality of television images. Recommendation ITU-R BT.500-14, 2019.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Xin Jin, Qianqian Qiao, Yi Lu, Huaye Wang, Heng Huang, Shan Gao, Jianfei Liu, and Rui Li. Apddv2: Aesthetics of paintings and drawings dataset with artist labeled scores and comments, 2024. URL https://arxiv.org/abs/2411.08545.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. E-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1244–1254, October 2021.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Jun Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:51737170.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proc. Int. Conf. Machine Learning (ICML)*, pp. 5338–5348, 2020.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016.

Shuitsu Koyama, Yuiga Wada, Daichi Yashima, and Komei Sugiura. MLLM-as-a-judge exhibits model preference bias. arXiv preprint arXiv:2604.11589, 2026.

Leida Li, Yipo Huang, Jinjian Wu, Yuzhe Yang, Yaqian Li, Yandong Guo, and Guangming Shi. Theme-aware visual attribute reasoning for image aesthetics assessment. *IEEE TCSVT*, 33(9): 4798–4811, 2023.

Leida Li, Xiangfei Sheng, Pengfei Chen, Jinjian Wu, and Weisheng Dong. Towards explainable image aesthetics assessment with attribute-oriented critiques generation. *IEEE TCSVT*, 35(2): 1464–1477, 2025a.

Zhuo Li, Xingao Yan, Xuebin Wei, and Feng Shao. IAACLIP: Image aesthetics assessment via CLIP. *Electronics*, 14(7):1425, 2025b.

Xiao-Chang Liu and Johan Wagemans. From concepts to judgments: Interpretable image aesthetic assessment, 2026. URL https://arxiv.org/abs/2603.18108.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

Anne-Sofie Maerten, Li-Wei Chen, Stefanie De Winter, Christophe Bossens, and Johan Wagemans. Lapis: A novel dataset for personalized image aesthetic assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 6356–6365, June 2025.

Anne-Sofie Maerten, Juliane Verwiebe, Shyamgopal Karthik, Ameya Prabhu, Johan Wagemans, and Matthias Bethge. Personalizing text-to-image generation to individual taste. arXiv preprint arXiv:2604.07427, 2026.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. In *ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*, 2021. arXiv:2106.13314.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1614–1623, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/martins16.html.

Adam W. Meade and S. Bartholomew Craig. Identifying careless responses in survey data. *Psychological Methods*, 17(3):437–455, 2012.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *ICLR*, 2023.

Enrico Parisini, Tapabrata Chakraborti, Chris Harbron, Ben D. MacArthur, and Christopher R. S. Banerji. Leakage and interpretability in concept-based models, 2025. URL https://arxiv.org/abs/2504.14094.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rolf Reber, Piotr Winkielman, and Norbert Schwarz. Effects of perceptual fluency on affective judgments. *Psychological Science*, 9(1):45–48, 1998.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Koki Ryu and Hitomi Yanaka. What do vision–language models encode for personalized image aesthetics assessment? In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 34146–34167, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1706. URL https://aclanthology.org/2026.findings-acl.1706/.

Ines Schindler, Georg Hosoya, Winfried Menninghaus, Ursula Beermann, Valentin Wagner, Michael Eid, and Klaus R. Scherer. Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PLoS ONE*, 12(6):e0178899, 2017.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017.

Frank Sibley. Aesthetic concepts. *The Philosophical Review*, 68(4):421–450, 1959.

Derya Soydaner and Johan Wagemans. Unveiling the factors of aesthetic preferences with explainable ai. *British Journal of Psychology*, 117(2):444–478, 2026. doi: https://doi.org/10.1111/bjop.12707. URL https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/bjop.12707.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Trans. Image Process.*, 27(8):3998–4011, August 2018.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, 2023.

Supatta Viriyavisuthisakul, Shun Yoshida, Kaede Shiohara, Ling Xiao, and Toshihiko Yamasaki. Explainable AI for image aesthetic evaluation using vision-language models. In *Conference on Artificial Intelligence x Multimedia (AIxMM)*, pp. 62–65, 2025. doi: 10.1109/AIxMM62960.2025.00017.

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*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i2.25353. URL https://doi.org/10.1609/aaai.v37i2.25353.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

M. K. Ward and Adam W. Meade. Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, 74:577–596, 2023.

Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks. In *Proc. Int. Conf. Machine Learning (ICML)*, pp. 11205–11216, 2021.

An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition. In *ICCV*, pp. 3090–3100, 2023.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, pp. 19187–19197, 2023.

Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19829–19837, 2022. doi: 10.1109/CVPR52688.2022.01924.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *ICLR*, 2023.

Robert B. Zajonc. Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2, Pt.2):1–27, 1968.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Hantao Zhou, Longxiang Tang, Rui Yang, Guanyi Qin, Yan Zhang, Yutao Li, Xiu Li, Runze Hu, and Guangtao Zhai. UniQA: Unified vision-language pre-training for image quality and aesthetic assessment. arXiv preprint arXiv:2406.01069, 2025a. URL https://arxiv.org/abs/2406.01069.

Hantao Zhou, Rui Yang, Longxiang Tang, Guanyi Qin, Runze Hu, and Xiu Li. Gamma: Toward generic image assessment with mixture of assessment experts. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2025b.

Appendix

A Method Details **15**

B Rules for Building the Concept Pool **17**

C Prompts **17**

D Dataset Details **22**

E Training and Evaluation Protocol **23**

F Analysis of the Weight Components **23**

G Analysis of the Concept Pool Construction **26**

H Baseline Details **29**

I Human Evaluation Details **32**

J Human Evaluation: Participants and Released Data **35**

K Further Limitations and Future Work **36**

A Method Details

This section supplements design decisions of Sec. 3 that are not essential for understanding the method.

Why a magnitude ρ. With $\rho \equiv 1$ fixed, the read-out is a convex combination, and m is nothing but an *average* of bounded responses. For an average to reach the ends of the rating scale, all concepts with non-zero weight must point to the same pole. ρ adds the freedom for the read-out to reach the ends of the scale, without breaking the additive decomposition. It changes only the total and leaves the ratios of the weights unchanged. Contributions therefore stay additive and proportional to p, and the additive decomposition $\sum_t c_t(x) = m(x)$ holds exactly under ρ. All that is lost is the guarantee that $\hat{y}$ always stays inside the rating scale, and only up to $\rho_{\max}$. Indeed, under $\rho \equiv 1$ we have $m \in (-1, 1)$, and $\hat{y}$ reaches the whole rating scale while staying inside it. The scale s of Sec. 3.1 is chosen that way. The trained models do represent the ends of the scale through ρ: fixing the read-out to an average shrinks their predictions toward the center. A model retrained without ρ, however, plays the same role with its remaining parts, and its accuracy shows no consistent difference (Appendix F). ρ is thus not an essential part but a degree of freedom that comes at no cost.

Initializing the importance and the magnitude. The h_I and h_T of Eq. (5) are randomly initialized. Their output layer is initialized small, however, so $w \approx 0$ at the start of training. The weights thus start from a uniform allocation over $\mathcal{S}(x)$. For the magnitude $\rho(x)$, v is initialized to zero and b is set to the value that gives $\rho \equiv 1$. Training therefore starts from a uniform convex-combination read-out over $\mathcal{S}(x)$. Any departure of the allocation from uniform, or of ρ from 1, is learned from the data.

Center of the response μ_t. With $\mu_t = 0$, since image and text features are both L_2-normalized, $\|I(x) - T(t_{\text{pos}})\|^2 - \|I(x) - T(t_{\text{neg}})\|^2 = -2z(x,t)$. Hence $d_\theta = 0$ is equivalent to x being *equidistant* from the two clauses. This is the neutral point defined by the concept itself, not the mean of the dataset. Subtracting μ_t as in Eq. (3) lets the sign of $d_\theta(x,t)$ be read against the mean over the training set instead. We make the latter the default for the readability of explanations.

Table S1: Candidates and resolved pools.

Dataset	L	Returned	Passed rules (by source) A	B	C	Backbone	$\lvert\mathcal{T}_0^\star\rvert$	$\sum_g \lvert\mathcal{T}_g^\star\rvert$	$\lvert\mathcal{T}\rvert$
Single-genre									
LAPIS	7	1,040	40	81	799	CLIP-B/16	43	201	244
						UniQA-B/16	59	203	262
APDDv2	5	780	40	81	570	CLIP-B/16	48	178	226
						UniQA-B/16	63	180	243
PAMELA	6	910	40	81	684	CLIP-B/16	52	154	206
						UniQA-B/16	54	159	213
Multi-genre									
AADB	8	1,170	40	81	969	CLIP-B/16	46	199	245
						UniQA-B/16	55	229	284
TAD66K	6	910	40	81	677	CLIP-B/16	36	48	84
						UniQA-B/16	34	48	82
PARA	8	1,170	40	81	942	CLIP-B/16	56	262	318
						UniQA-B/16	64	232	296

Returned: candidates the generator returned. Step 1 requests $130 + 130L$ pairs; A and B are dataset-independent and asked once for all datasets, so only the $130L$ of C are per dataset. Passed rules: candidates passing the pair rules of Step 2, by source (A: attribute seeds, B: general, C: per level); A and B are the same pairs for every dataset. The last three columns are after Step 3.

What the gain γ_t can do. Being positive, γ_t cannot change the ranking of images *within* a concept. Scaling a concept's responses by a constant leaves correlations unchanged, so it cannot fit concepts to one another either. γ_t only changes how strongly the concept contributes relative to the others.

Normalizing the subject direction σ_t. Applicability depends only on the *direction* of the subject, not on the length of the vector. Without normalization, the length, which differs from pair to pair, would multiply the logit directly. The comparison between concepts would then mix in a scale difference unrelated to the image.

The tag term. The tag term of Eq. (4) is a correction that lets concepts of the level an image shows pass the gate more easily. The subject direction σ_t alone does not identify the level a concept was selected in, since the shared final noun need not name it (PARA's portraits level yields "Balanced exposure."). The term is added unweighted and enters only the applicability, so the response, the contribution c_t and the additive decomposition are unchanged.

Eligibility floor and ordering threshold. The ranks of Eq. (2) are taken within $\mathcal{X}_g$. For $g \neq 0$ the tag is constant on $\mathcal{X}_g$, so r_g reduces to the plain Spearman correlation there. Partialling acts only for $g = 0$, where the ranks are regressed on the dummy-coded levels of G and the residuals are kept. The floor of Eq. (2) is $f_g = \max\big(f,\ \zeta/\sqrt{n_g - 3}\big)$. Here $f = 0.10$ is the effect-size floor. The second term is a guard derived from the sample size under Fisher's z transform ($\zeta = 4$), which demands a larger correlation from sparser levels. For the same reason, the threshold of Step 3 is rescaled per level as $\tau_g = \tau\sqrt{N/n_g}$ ($N = |\mathcal{X}|$).

B Rules for Building the Concept Pool

This section gives the rules of the three steps of Sec. 3.2 that were left out of the main text. How the floor f_g and threshold τ are set is given in Appendix A. The wording used for generation is in Appendix C.

Breakdown of candidate generation (Step 1). The 40 pairs of A are drawn two each for 11 aesthetic attributes (e.g., Color harmony) and 9 image-quality attributes (e.g., Blur). The 90 pairs of B consist of 50 pairs on general aesthetic quality and 40 on general image quality. The generation template shared by the three sources differs only in {WHAT}, the presence of images, and whether the final noun is specified. Each call generates at most 20 pairs, so we make several calls. Each call within a source is shown the positive clauses already produced as exclusions. A and B both have $g(t) = 0$ and compete for the same $\mathcal{T}_0^\star$, so calls for B are also shown the exclusions from A, across sources. This is only part of the instruction: every pair returned is kept as a candidate.

Evaluative and descriptive pairs (Step 1). The generator also declares, for each pair, whether it is *evaluative* or *descriptive*. In an evaluative pair, the meaning of the words alone makes one pole better ("Sharp photo." / "Blurry photo."). A descriptive pair only names an axis, and the images decide which pole is preferred ("Centered subject." / "Off-center subject."). The set $\mathcal{T}_{\text{ev}}$ of pairs declared evaluative is tested one-sided in the eligibility test of Step 2 (Eq. (2)).

Rules on pair form and duplicates (Step 2). The form rules require that the positive and negative clauses share the same final noun, and that each clause has at most three words. In addition, neither clause may have two or more words left once the shared part is removed. The duplicate rule first merges pairs with the same two clauses. It then drops a pair when a clause that could become its positive after orientation is already used by an earlier pair. For a descriptive pair, this covers both clauses. Whether a clause is already used is judged on its wording. A final noun naming only the medium (image, photo, painting, etc.) is disregarded, however, so "Candid photo." and "Candid image." count as one clause. The reason is that the explanation renders this noun as image (Appendix I), and keeping the two apart could let one clause appear twice in it. The rule is applied in the order the candidates arrive: A, B, and then each level. A clause is thus kept by the source that proposed it first, i.e., the more general one.

C Prompts

This section reproduces verbatim the text we sent to MLLMs. It consists of three prompts. The first is the generation prompt used to build the concept pool of Sec. 3.2 (Fig. S1, with slot values in Table S2). The second is the prompt with which we assign content tags in domains where the dataset distributes none (Fig. S2). The third is the prompt of the LLM evaluation of Sec. 4.2 (Fig. S3). The first two are sent to Gemini with temperature 1.0, a thinking budget of 0, and images downscaled to 512 px on the long side. For the LLM evaluation, Claude Opus 5 receives the same images as the raters (800 px on the long side). Decoding is left at the API's defaults, and answers are returned through structured output.

Prompt-bank generation prompt (one template; four versions of {WHAT})

System prompt (sent with every generation call)

```
You are a researcher specializing in empirical aesthetics, skilled at
naming the distinct visual concepts that drive aesthetic judgments.
```

User prompt — (1) sample images (C only; *A and B attach no images. {k} images, redrawn for every call)*

[unlabeled sample image] × {k}

C (per level; drawn from {content level} only)

```
The attached images are {k} random samples of a single content type
in this collection - the one named below - shown so you can see what
those images actually look like.
```

User prompt — (2) task

```
Create {n} antonymic text-prompt pairs describing {WHAT}.  Each pair
contrasts a positive concept with its antonymic negative, each prompt
3 words or fewer (including the shared trailing noun); across the
pairs, cover mutually distinct concepts spanning different dimensions,
with no repeats or paraphrases.
```

User prompt — (3) exclusion clause *(appended to (2) only when some concepts are already covered)*

```
This applies to the already-covered concepts below as well - propose
genuinely new ones, not these:
{covered}
```

User prompt — (4) pair format and output

A and B

```
IMPORTANT: every prompt must end with the noun "image", and within
each pair "pos" and "neg" must be identical except for the contrasting
concept word(s) (e.g.  "Sharp image." / "Blurry image.").
```

C

```
IMPORTANT: within each pair, "pos" and "neg" must be identical except
for the contrasting concept word(s); keep the SAME ending noun and
sentence structure (e.g.  "Sharp photo." / "Blurry photo.", NOT
"Blurry picture.").  You may choose any trailing noun, but "pos" and
"neg" of a pair must share the same one.
```

All sources, following the above

```
Use plain, common, everyday vocabulary; avoid rare jargon, obscure or
highly technical terms.  For each pair also set "evaluative":  true
if one end is better than the other by the MEANING OF THE WORDS ALONE,
in any collection of images ({evaluative examples}), and false if the
pair merely names an axis whose two ends are both legitimate and whose
better end depends on the images ({descriptive examples}).  Judge
the words on their own - do NOT consider which end you expect to be
rated higher in this particular collection.  When "evaluative" is true,
"pos" must be the end that is better by meaning.  Return {"pairs":
[{"pos":  "...", "neg":  "...", "evaluative":  true|false}]}.  Respond
ONLY with a JSON object of the requested form.  No prose, no markdown.
```

Figure S1: Prompt used to generate the candidate pairs. Blue { }: slots (values in Table S2); orange [ ]: image positions. The sources A (attribute seeds), B (general) and C (per level) differ in {WHAT}, in (1), in whether (4) fixes the final noun, and in what {covered} lists; the wording of (2) and (3) is shared. B is asked twice, once for aesthetics and once for image quality, which gives the fourth version. A and B name no collection and attach no images, so both are generated once and reused for every dataset.

Table S2: Slot values for Fig. S1. Only {collection phrase} and {content levels} depend on the dataset. {content levels} lists the content-tag levels in plain words, omitting levels that name no content.

A, B, C	
{covered}	Positive clauses proposed so far, joined by `"; "` (kept or not, in full). The collection-wide sources share one list: each of the 20 calls of A lists the clauses of the others, and B lists all of A's clauses as well as its own across its two scopes. A per-level call of C lists only that level's earlier calls. C is therefore shown neither A nor B nor another level, whose vocabulary it may legitimately restate under a different $g(t)$; a clause that does repeat one is removed by the de-duplication rule of Step 2 either way
A, B	
{evaluative examples}	`"Sharp image." / "Blurry image."; "Clean image." / "Noisy image."`
{descriptive examples}	`"Centered image." / "Off-center image."; "Glossy image." / "Matte image."; "Warm image." / "Cool image."`
C	
{evaluative examples}	`"Sharp photo." / "Blurry photo."; "Clean image." / "Noisy image."`
{descriptive examples}	`"Centered subject." / "Off-center subject."; "Glossy surface." / "Matte surface."; "Warm tones." / "Cool tones."`
A (attribute seeds)	
{WHAT}	`image aesthetics for the attribute "{attribute}" (e.g. "Vibrant image." / "Faded image."; "Sharp image." / "Blurry image.")`
{n}	2 per attribute (20 attributes $\times$ 2 $=$ 40 pairs)
{attribute}	11 `aesthetic attributes: Interesting content, Subject emphasis, Good lighting, Color harmony, Vivid colors, Depth of field, Motion blur, Rule of thirds, Balancing element, Repetition, Symmetry` 9 `quality attributes: Blur, Color-related, Contrast, Compression, Noise, Over-exposure, Quantization, Under-exposure, Spatially-localized`
B (general)	
{WHAT}	*aesthetics*: `general aesthetic quality, excluding the following elements:` the 11 aesthetic attributes of A, joined by `", "` *image quality*: `general image quality, excluding the following degradation factors:` the 9 quality attributes of A, joined by `", "`
{n}	$\leq$ 20 (3 calls for aesthetics and 2 for image quality; $50 + 40 = 90$ pairs in total)
{k}	— (no images)
C (per level)	
{WHAT}	`image aesthetics specifically for {collection phrase}, specifically those whose content is: {content level}`
{n}	$\leq$ 20 (7 calls per level; $130L$ pairs per dataset)
{content level}	Each level of {content levels} in turn
{k}	Number of content-tag levels: LAPIS 7, APDDv2 5, PAMELA 6, AADB 9, TAD66K 6, PARA 8 (AADB's other is counted). Drawn from that level only, redrawn for every call (with replacement)

Dataset	{collection phrase}	{content levels}
LAPIS	`artwork`	`abstract works; cityscapes; flower paintings; landscapes; nude paintings; portraits; still life`
APDDv2	`paintings and drawings`	`portraits; landscapes; still life; flowers and birds; mountains and water`
PAMELA	`AI-generated artwork`	`abstract works; cityscapes; flower paintings; landscapes; portraits; still life`
AADB	`photographs spanning many subjects`	Same as PARA (the eight `semantic_8` levels, assigned by us)
TAD66K	`images spanning many themes`	`landscapes; man-made subjects; colours; people; animals; plants`
PARA	`images spanning many themes`	`portraits; outdoor scenes; animals; food; still life; indoor scenes; buildings; plants`

Calibration prompt for the content tagger (PARA `semantic_8`; no slots)

System prompt — PARA

```
You label photographs for a research dataset.  You judge only what is
visible in the image.
```

User prompt — PARA

[image to label]

```
Label the MAIN SUBJECT of this photograph - what the photograph is
OF. When several things are in frame, judge the one the photograph was
taken for:  largest, in focus, and placed as the subject.
- portrait:  a person or people.
- scene:  an outdoor view - landscape, cityscape, street - by day or
night.
- animal:  an animal.
- food:  food or drink.
- stilllife:  an inanimate object is the subject - an artifact,
product, ornament, sculpture, or a close detail of one.  Any number
of them, arranged or not, indoors or outdoors.
- indoor:  an interior space.
- building:  a building or structure.
- plant:  plants or flowers.
- other:  the subject is none of these.
'other' is for a subject outside the list, not for hesitation between
two labels on it.  Choose exactly one label; never give more than one.
```

— constrained by structured output to —

```
{"label":  {enum of levels}}
```

Figure S2: Prompt for scoring the content tagger against a taxonomy with public ground truth, PARA `semantic_8` (Table S3). The same prompt is used to tag AADB.

Table S3: Agreement of the content tagger with PARA's public taxonomy, on 100 training images per level.

	PARA `semantic_8`
Levels / images	8 / 800
Accuracy	.851
Cohen's κ	.830
macro-F1	.855
η^2 (ground-truth tags)	.083
η^2 (tagger)	.068
Removal ratio	0.81
Removed beyond content	.002

η^2: share of rating variance explained by the tags, on the same images. Removal ratio: null-corrected η^2 (tagger) / η^2 (ground-truth tags). Removed beyond content: variance the tagger's tags still explain after the ground-truth tags are removed from the ratings (0 = exactly the content).

LLM evaluation prompt (R, Task 1, Task 2; the slots take the raters' wording)

System prompt (sent with every call)

```
You are a participant in a study about explanations of image
aesthetics.  Each description you will read was written about
the aesthetic quality of one image.  Look at the images carefully
and answer each question from your own judgment, as an attentive
participant would.
```

R — user turn

```
[image]
{title} -- {prompt}
Answer on the scale from {min} to {max} in steps of {step}
({anchors}).
```

— constrained by structured output to —

```
{"rating":  {scale value}}
```

Task 1 — the R turn, the judge's own R answer as the assistant turn, then

```
Here are two descriptions of this image.
A: {left explanation}
B: {right explanation}
Judge each description as an explanation -- not by whether you like
the image or agree with the rating.
Answer each question with one of:  Clearly A, Slightly A, Slightly B,
Clearly B.
Q0.  Overall, which description is the better explanation?
Q1.  Which description lists features that are actually present in
this image?
Q2.  Which description says something specific about this image,
rather than something that could be said of any image?
```

— constrained by structured output to —

```
{"Q0":  {choice}, "Q1":  {choice}, "Q2":  {choice}}
```

Task 2 — user turn

```
The description below was written about one of the four images.  Which
one is it?
Description:  {explanation}
Image 1:  [image]  Image 2:  [image]  Image 3:  [image]  Image 4:  [image]
Answer with the number of the image (1-4).
```

— constrained by structured output to —

```
{"answer":  {1-4}}
```

Figure S3: LLM evaluation prompt. The slots are filled with the wording of the raters' screens: for PARA, {title} is "Aesthetics score", {prompt} "Comprehensive aesthetic judgement", and the scale runs from 1 to 5 in steps of 0.5 with the anchors written as "1 = Much worse; . . . ; 5 = Much better" (Table S15). Each Task 1 pair and Task 2 trial is asked in a presentation its raters saw, chosen so that FaTeX's side and the answer's position stay balanced, and once more with that presentation swapped.

Table S4: Dataset statistics. Image counts are per split.

Dataset	Domain	Scale	Raters	Ratings/img	L	Images			
						Train	Val	Test	Total
Single-genre									
LAPIS	Painting	0–100	568	24.2	7	8,205	1,173	2,345	11,723
APDDv2	Painting, drawing	0–100	37	6.3	5	8,019	1,002	1,002	10,023
PAMELA	AI-generated art	1–5	126	13.3	6	1,279	258	440	1,977
Multi-genre									
AADB	Photo	0–1	190	5.0	8	8,458	500	1,000	9,958
TAD66K	Photo	1–10	—	$\approx$1200	6	41,797	10,451	14,079	66,327
PARA	Photo	1–5	438	25.9	8	22,576	5,644	3,000	31,220

Rater counts are not released for TAD66K (—). L: number of content-tag levels used in Eq. (2).

D Dataset Details

This section describes the provenance, rating procedure and content tags of the six datasets used in Sec. 4.1. Image counts and splits are in Table S4. Whenever we draw a split, we stratify by the mean rating binned into 7 equal-width bins.

LAPIS (Maerten et al., 2025). LAPIS consists of 11,723 paintings selected from WikiArt. They were rated by 568 raters on a 0–100 slider, and per-rater records are available. We use the official train/val/test splits. We use the seven genres (abstract, cityscape, flower painting, landscape, nude painting, portrait, still life) as the content tag ($L = 7$).

APDDv2 (Jin et al., 2024). APDDv2 consists of 10,023 oil paintings, drawings and Chinese paintings. They were rated by 37 art experts in three panels (18 for oil, 10 for drawing and 9 for Chinese painting); half of the experts are art teachers. Each image has at least 6 raters. The images come from ConceptArt, DeviantArt, WikiArt and the China Art Museum. About a quarter are works by art-school students, included to widen the range of ratings. The overall score, on 0–100, is given against a table of reference images prepared for each category. Only per-image means are released, with neither per-rater records nor rating histograms, so NIMA is undefined. We use the official train/val/test splits. We use the five subjects (portraiture, landscapes, still life, floral and avian, mountains and water) as the content tag ($L = 5$).

PAMELA (Maerten et al., 2026). PAMELA consists of outputs of text-to-image models (FLUX.2 and Nano Banana), rated on a continuous 1–5 slider with five anchors (bad, poor, fair, good, excellent). The raters give one aesthetic judgment that combines beauty, liking and attraction. The ratings are continuous, so there are no discrete levels for a rating histogram and NIMA is undefined. We use its art collection (1,977 images, 126 raters) and the official train/val/test splits. We use the six genres (abstract, cityscape, flower painting, landscape, portrait, still life) as the content tag ($L = 6$).

AADB (Kong et al., 2016). AADB is a collection of photos of any genre, collected from Flickr. Each photo was rated on 1–5 by 5 of 190 raters, and the ratings are normalized to $[0, 1]$. Only per-image means are released, with no rating histograms, so NIMA is undefined. We use the official train/val/test splits. We use the eight categories of PARA (portrait, scene, animal, food, stilllife, indoor, building, plant) as the content tag ($L = 8$), assigned to all 9,958 images with an MLLM (Gemini-3.5 Flash). The tagging uses, unchanged, the prompt with which we calibrate the tagger on PARA (Fig. S2). Its accuracy is scored against PARA's ground truth (Table S3). The label other marks subjects outside the eight categories (29 training and 6 test images). It names no content, so we exclude it from the range of g and from the count L, but keep it in the covariate. These labels have not been reviewed by humans.

TAD66K (He et al., 2022). TAD66K consists of 66,327 photos collected from Flickr and rated on 1–10. It is the largest and most densely rated dataset we use (about 1,200 opinions per image). The images were rated independently per theme. Only per-image means are released, with no rating histograms, so NIMA is undefined. We use the official train/test split and hold out 20% of the training

split for validation. We use six themes (Landscape, Artifact, Color, Human, Animal, Plant) as the content tag ($L = 6$).

PARA (Yang et al., 2022). PARA consists of $31{,}220$ photos rated by 438 raters on 1–5 in steps of 0.5 (9 levels). Each image has 25.9 raters on average, the densest among datasets with per-rater records. We use the official train/test split and hold out 20% of the training split for validation. We use eight categories (portrait, scene, animal, food, stilllife, indoor, building, plant) as the content tag ($L = 8$), with nightScene merged into scene.

E Training and Evaluation Protocol

This section gives the values of the protocol of Sec. 4.1 that are left out of the main text.

Hyperparameters of FaTeX. Only the copy I_ϕ is updated. Its trainable parameters are the LayerNorm affine parameters of all 12 blocks and the output LayerNorm of the image tower. On top of these come the MLPs h_I, h_T of Eq. (5) (hidden width 128, output $k = 32$), the per-concept gains γ_t (one scalar per concept), and the magnitude head (v, b) (Sec. 3.4). The temperature of the response (Eq. (3)) is fixed to $\kappa = 30$, and the upper bound of the magnitude to $\rho_{\max} = 3$.

Map to the rating scale. The affine map $\hat{y}(x) = \bar{y} + s\, m(x)$ (Sec. 3.1) takes $m(x)$ of Eq. (1) to the rating scale. Its constants are the mean rating of the training set, $\bar{y}$, and the distance from there to the farther end of the rating scale, $s = \max(\bar{y} - y_{\min},\ y_{\max} - \bar{y})$. Both are determined by the split and are neither learned nor tuned.

Input. Images are resized so that the short side is 224 px and center-cropped to a square. During training, we randomly crop 224 px from a 256 px frame, a resize-then-crop scheme common in IAA (Kong et al., 2016; Li et al., 2025b), and add horizontal flips. Resizing, cropping and augmentation are shared by all methods that involve training.

Optimization. We use Adam with the learning rate split into two groups: η for the fine-tuned encoder and 10η for the newly initialized parts (h_I, h_T, γ_t, magnitude head). Only η is searched. This is the same $1 : 10$ ratio as in NIMA (Talebi & Milanfar, 2018).

Budget and search. Training runs for 20 epochs with batch size 128. The model at the final epoch is the result of the trial. The learning rate is searched over $[10^{-7}, 10^{0}]$ with a budget of 15 trials of log-uniform random sampling. The first trial is the setting each method reported in its original paper. The three reported runs differ only in the random seed ($\mathrm{seed} = 42 + \mathrm{run} - 1$).

Verifying the explanations. The backbone in Sec. 4.2 is the one with the higher test SRCC in Table 1 (CLIP-B/16 for LAPIS, UniQA-B/16 for PARA), and the run is the median of the three seeds. Since $|\mathcal{S}(x)|$ differs from image to image, the number of concepts deleted is a fraction of $|\mathcal{S}(x)|$ (5–50%) rather than a fixed count. In the swaps, the set size is matched to the image's own $\lceil 0.1|\mathcal{S}(x)| \rceil$ under every condition. Even when a set is borrowed, the response and importance are those of x itself. The top 10% of concepts account for 57% of the total $|c_t|$ on average (66% on PARA). Conditions that involve random sampling are drawn 20 times, and we report the mean. These analyses and those of Appendices F and G encode the test images anew. The full model on PARA then reads .890 rather than the .891 of Table 1 (.8905 vs. .8907), and all differences are taken from the re-encoded value.

F Analysis of the Weight Components

The weight $p(t \mid x)$ and the magnitude $\rho(x)$ are built from three parts. The first is the set $\mathcal{S}(x)$ that applicability selects (its support) (Eq. (4)). The second is the importance $w_t(x)$ (Eq. (5)), which sets the allocation within that set. The third is the magnitude $\rho(x)$. Whether each part is *needed* is decided by retraining the model without it (Table S5). *How* the trained model uses each part is examined by replacing one part at a time and evaluating the model again (Table S7). Where the two disagree, the role of the part is one that the remaining parts take over after retraining. We study the same two runs as in Sec. 4.2, and each retrained condition is run once. Across the three seeds of the same configuration, the SRCC spans .007 on LAPIS and .003 on PARA. We do not distinguish differences smaller than this. On the explanation side, we take the top three concepts by $|c_t(x)|$, which the explanation text uses (Appendix I). We record which stage they come from ($\mathcal{T}_0^\star$, the image's own

Table S5: Each part of the weighting removed and the model retrained, on the runs of Sec. 4.2: one run per condition, at the full model's learning rate and seed.

						Top-3 from				Top-3 overlap		
Condition	SRCC	Δ	PLCC	$\lvert\mathcal{S}\rvert$	Share	$\mathcal{T}_0^\star$	Own	Other	Distinct	Same	Other	Neg.
LAPIS (CLIP-B/16)												
Full model	.846	—	.845	85.7	.315	.576	.250	.174	100	.207	.125	.642
No gate	.847	+.001	.846	244.0	.178	.298	.107	.595	70	.223	.172	.652
W/o tag term	.840	−.006	.841	103.4	.276	.529	.141	.330	98	.214	.156	.658
Polarity direction	.842	−.004	.844	79.3	.313	.668	.138	.194	127	.209	.081	.606
Uniform allocation	.826	−.020	.824	85.7	.165	.556	.260	.184	67	.367	.230	.666
$\rho \equiv 1$	.832	−.014	.833	85.7	.328	.640	.220	.141	114	.221	.128	.665
PARA (UniQA-B/16)												
Full model	.890	—	.929	104.6	.340	.540	.324	.136	123	.171	.080	.571
No gate	.897	+.006	.932	296.0	.229	.132	.161	.707	94	.163	.111	.542
W/o tag term	.891	.000	.927	123.1	.348	.575	.212	.213	120	.172	.103	.599
Polarity direction	.894	+.004	.927	100.8	.271	.634	.172	.195	141	.153	.106	.586
Uniform allocation	.862	−.029	.907	104.6	.112	.508	.325	.168	77	.299	.166	.625
$\rho \equiv 1$	.897	+.006	.929	104.6	.386	.489	.333	.177	122	.183	.092	.557

Δ: SRCC against the full model, taken before rounding. Columns after PLCC are those of Table S7, measured on each retrained model's own explanations. No gate puts every concept on the support, so its $\lvert\mathcal{S}\rvert$ is the pool size.

level, or another level). We also measure their overlap, separately between images of the same level and between images of different levels.

- **The prediction depends on the allocation.** Making the allocation uniform lowers the SRCC beyond the seed range on both datasets. The allocation of the trained model contains per-image information that cannot be shared even with images of the same level.
- **The parts that set the support act on the explanation, not on the prediction.** Retraining without the gate, the tag term or the subject direction does not lower the SRCC beyond the seed range. Each removal, however, either lets concepts of other levels into the top three or reduces the concepts of the image's own level.
- **The effect of the magnitude is inconsistent.** The trained model expresses the ends of the scale through ρ. When it is retrained without ρ, the remaining parts play the same role.

Retraining without a part. We train five models (Table S5). The first has no gate and takes all concepts as the support. The second drops the tag term. The third reads applicability along the polarity direction $T(t_{\text{pos}}) - T(t_{\text{neg}})$ of PPA (Hara & Otsuka, 2026) instead of the subject direction σ_t. The fourth makes the allocation uniform over the support. The fifth fixes the magnitude to $\rho \equiv 1$. The learning rate, random seed and number of epochs are those of the FaTeX run, and nothing is searched.

The gate fits the concepts in the explanation to the image. Removing the gate does not lower the SRCC (+.001; PARA +.006). Among the top three concepts, however, the share from other levels rises (.174 → .595; PARA .136 → .707). Even when retrained over all concepts, the importance does not by itself exclude concepts that do not apply to the image. What decides which concepts may appear in an explanation is the support, not the allocation. Dropping the tag term moves the explanation the same way, to a smaller degree (.174 → .330; PARA .136 → .213). A concept of another level is not necessarily foreign to the image, however. Fig. S4 shows a PARA portrait of a child holding flowers: "High-contrast plant image.", selected for the plants level, ranks second and is part of the explanation, and it describes the image. Table S6 lists the concepts of other levels on the support of randomly drawn images, as a qualitative check.

The subject direction keeps concepts of the image's own level. Selecting along the polarity direction does not lower the SRCC beyond the seed range. Among the top three concepts, however, the share from the image's own level nearly halves (.250 → .138; PARA .324 → .172). The polarity direction measures which pole the image leans to, not whether the concept applies to the image.

The allocation is needed for both prediction and explanation. Uniform allocation is the only one of the five conditions that lowers the SRCC beyond the seed range on both datasets (−.020; PARA

Table S6: Concepts of other levels on the support of five randomly drawn images per dataset (qualitative).

Image	Concepts of other levels on the support (top 5 by $\lvert c_t(x) \rvert$)
LAPIS (CLIP-B/16)	
	**Painterly landscape art**; Crowded landscape painting; **Realistic landscape painting**; **Impasto canvas**; Inspiring landscape painting
	**Abstract landscape painting**; Abstract nude painting; Amateurish nude painting; **Loose landscape painting**; Abstracted nude painting
	**Biomorphic abstraction**; **Diluted paint**; **Washed canvas**; **Transparent shapes**; Derivative painting
	Classical portrait composition; Inventive painting; Fragmented pieces; Inspired artwork; Regal portrait attire
	Simplistic town artwork; Simplistic abstraction; Sparsely arranged artwork; Agitated strokes; Geometric abstraction
PARA (UniQA-B/16)	
	**Harmonious color palette**; **Cluttered space photograph**; **Awkward composition**
	Cool-toned portrait; Defined jawline; **Monochrome portrait**; **Dark portrait**; **Faint building silhouette**
	**Overcast outdoor scene**; Flat outdoor horizon; Unappealing outdoor setting; Ordinary outdoor vibe; Monochromatic outdoor colors
	*None.*
	*None.*

A qualitative check, not a measurement. Five images per dataset are drawn with seed 42 from the 120 images of the human evaluation, and the models are those that produced its explanations. For each image, the concepts selected for a content level other than the image's own that are on its support are listed, at most five, in descending order of $\lvert c_t(x) \rvert$, each given by the pole the image leans to. **Orange**: the clause describes the image, in the authors' judgment. Purely evaluative clauses (e.g., *inspiring*, *captivating*) are never marked and are shown in gray. The judgments are subjective and only show what such concepts look like; we give no count. Many of the unmarked ones share their modifier with the image and differ from it only in the subject (e.g., the abstraction of "Abstract nude painting" in the second image, the classical arrangement of "Classical portrait composition" in the fourth), so the confusion is an understandable one.

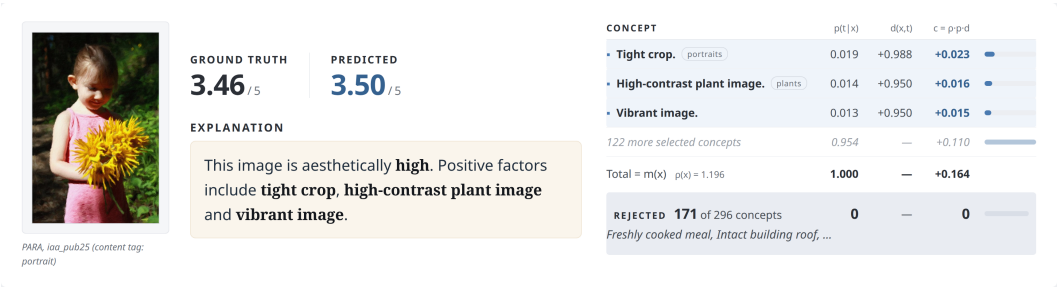

Figure S4: A concept of another level that applies, in the format of Fig. 1. The image is tagged portrait; "High-contrast plant image." was selected for the plants level.

$-.029$). The explanation no longer concentrates on a few concepts: the share of $|c_t|$ held by the top three falls ($.315 \rightarrow .165$; PARA $.340 \rightarrow .112$). Images of the same level also all receive similar explanations: the overlap within a level rises ($.207 \rightarrow .367$; PARA $.171 \rightarrow .299$).

The effect of the magnitude is inconsistent. Under $\rho \equiv 1$, the SRCC difference is $-.014$ on LAPIS and $+.006$ on PARA, so the sign is split. Neither the errors at the high and low ends of the ratings nor the explanation metrics change in one direction.

Replacing a part of the trained model. Without retraining, we replace one part at a time and evaluate the model again (Table S7). In N1, we rebuild the support under the same rules as above, or as a random set of the same size. In N2, we make the allocation uniform over the image's own support, or fix the magnitude to its mean. In N3, we borrow the importance logits from another image (the donor), or replace them with their mean over all images. In every case the other parts remain the image's own. A drop therefore also includes the part due to the remaining parts not matching the replaced one.

The allocation contains per-image information. Borrowing another image's allocation lowers the SRCC by about .05 (PARA .10). The drop is the same when the donor is an image of the same level. With the mean allocation, which does not depend on the image, the drop is smaller. The explanation, however, becomes a fixed pattern per level: the overlap within a level rises ($.207 \rightarrow .383$; PARA $.171 \rightarrow .350$). This does not mean that the allocation favors the image's own level. Compared with uniform allocation, the learned allocation favors concepts of the image's own level on PARA. On LAPIS it instead favors $\mathcal{T}_0^\star$.

The magnitude expresses the ends of the scale. Fixing ρ at its mean lowers the SRCC by .026 (PARA .016), and the prediction shrinks toward the middle of the scale. On the top third of the ratings, the mean error moves ($-.035 \rightarrow -.049$; PARA $-.012 \rightarrow -.028$). After retraining, the remaining parts take over this role.

Replacing the support changes the explanation in the same direction as retraining. With the support built without the gate or the tag term, concepts of other levels increase, as in the retrained models. The SRCC drop (without the gate, $-.016$; PARA $-.052$) disappears after retraining. It is therefore a mismatch arising because the importance was learned on the support of FaTeX. With the support of the polarity direction, fewer images have a negative contribution among their top three ($.642 \rightarrow .543$; PARA $.571 \rightarrow .463$). This difference, too, disappears after retraining, as the importance compensates for it.

G Analysis of the Concept Pool Construction

We examine what the three steps of Sec. 3.2 do. We use the same two dataset–backbone settings as in Sec. 4.2 (LAPIS $\times$ CLIP-B/16 and PARA $\times$ UniQA-B/16). Whether a component is needed is decided with models retrained on a swapped pool (Table S8). What each step selects and drops we study without training. We start from all candidates the generator returned, with their projections on the training images, and rerun Steps 2 and 3 with a single change (Table S9). In this case we use a proxy for prediction accuracy: a ridge regression that predicts the rating from the concept responses of the pool. Its inputs also include dummies for the content tag, and a level-specific concept takes

Table S7: What each part of the weighting does, read off the trained models of Sec. 4.2 with one part replaced and nothing retrained.

					Top-3 from				Top-3 overlap			
Condition	SRCC	Δ	$\lvert\mathcal{S}\rvert$	Share	$\mathcal{T}_0^\star$	Own	Other	Distinct	Same	Other	Neg.	Kept
LAPIS (CLIP-B/16)												
Full model	.846	—	85.7	.315	.576	.250	.174	100	.207	.125	.642	—
Selection (N1): the support rebuilt under another rule												
Own rule, rebuilt	.846	.000	85.7	.315	.576	.250	.174	100	.206	.125	.642	.999
No gate	.830	−.016	244.0	.176	.306	.129	.565	78	.201	.148	.701	.507
W/o tag term	.843	−.003	103.4	.278	.483	.197	.321	98	.192	.120	.661	.806
Polarity direction	.826	−.020	79.3	.317	.632	.134	.234	108	.218	.134	.543	.662
Random, same size	.828	−.018	85.7	.306	.611	.067	.322	119	.174	.135	.705	.585
Stages (N2): on the image's own support												
Uniform allocation	.813	−.033	85.7	.158	.340	.449	.211	81	.233	.090	.638	.316
ρ at its mean	.820	−.026	identical to the full model									1.000
Allocation (N3): importance logits from elsewhere												
Donor, same level	.798	−.048	85.7	.309	.594	.239	.168	103	.200	.124	.652	.371
Donor, other level	.791	−.055	85.7	.298	.604	.227	.169	105	.170	.129	.679	.324
Donor, any image	.793	−.052	85.7	.301	.609	.222	.169	105	.173	.131	.676	.330
Mean over images	.823	−.023	85.7	.209	.791	.091	.119	42	.383	.301	.675	.419
PARA (UniQA-B/16)												
Full model	.890	—	104.6	.340	.540	.324	.136	123	.171	.080	.571	—
Selection (N1): the support rebuilt under another rule												
Own rule, rebuilt	.891	.000	104.6	.340	.540	.324	.136	123	.171	.080	.571	.999
No gate	.839	−.052	296.0	.171	.338	.134	.528	95	.147	.116	.735	.496
W/o tag term	.878	−.012	123.1	.309	.469	.248	.283	122	.168	.099	.653	.795
Polarity direction	.870	−.021	100.8	.346	.548	.205	.247	128	.136	.081	.463	.703
Random, same size	.840	−.051	104.6	.342	.564	.076	.360	136	.109	.091	.737	.551
Stages (N2): on the image's own support												
Uniform allocation	.833	−.057	104.6	.131	.908	.060	.032	55	.416	.398	.610	.154
ρ at its mean	.875	−.016	identical to the full model									1.000
Allocation (N3): importance logits from elsewhere												
Donor, same level	.796	−.095	104.6	.336	.554	.310	.136	126	.164	.081	.620	.294
Donor, other level	.794	−.096	104.6	.336	.562	.276	.162	127	.138	.086	.642	.253
Donor, any image	.795	−.095	104.6	.335	.562	.280	.158	128	.138	.085	.643	.258
Mean over images	.849	−.041	104.6	.208	.482	.350	.168	76	.350	.192	.655	.380

Δ: SRCC against the full model, and for the selection rows against the model's own rule rebuilt the same way (the rules are recomputed in fp32 from the cached embedding, which moves about one image in ten by one concept on the sparsemax boundary and the SRCC by at most .0004). $|\mathcal{S}|$: concepts on the support. Share: fraction of $\sum_t |c_t(x)|$ held by the top three. Top-3 from: fraction of the top-three slots held by $\mathcal{T}_0^\star$, by the image's own level and by another level. Distinct: concepts appearing in any top three over the test split. Top-3 overlap: mean $|A \cap B|/3$ over 20,000 random pairs of images of the same level or of different levels. Neg.: images whose top three hold a negative contribution. Kept: mean overlap with the full model's top three. Rows with a random draw are the mean of 20 draws.

values only on images of its own level. We report its 5-fold cross-validated SRCC on the training images, called the proxy SRCC below.

- **Whether selecting the vocabulary by the rating helps prediction depends on the dataset.** Replacing it with PPA's fixed vocabulary lowers the SRCC beyond the seed range on LAPIS, but not on PARA.
- **Content control helps the explanations, not the prediction, and only where content and rating are strongly linked.** Removing it leaves the SRCC unchanged. On LAPIS, however, concepts that merely respond to "what is depicted" then enter the explanations of about half of the test images. On PARA, no such concept appears.
- **OMP is what removes redundant paraphrases.** Taking candidates in order of $|r|$ leaves paraphrases in the pool.

Retraining with a swapped pool. We train two models. The first replaces the vocabulary with the fixed 260 pairs of PPA (Hara & Otsuka, 2026) and keeps the read-out of FaTeX. This vocabulary has no levels, so every concept belongs to $\mathcal{T}_0^\star$ and is always in the set. The second uses a pool re-solved without content control. Its eligibility test and ordering use the marginal correlation instead of the correlation partialled on the content tag. Level-specific concepts are tested within their level, where

Table S8: The pool swapped and the model retrained, on the runs of Section 4.2: one run per condition, at the full model's learning rate and seed.

						Top-3 from				Top-3 overlap			
Condition	SRCC	Δ	PLCC	$\lvert\mathcal{S}\rvert$	Share	$\mathcal{T}_0^\star$	Own	Other	Distinct	Same	Other	Neg.	Uncontr.
LAPIS (CLIP-B/16)													
Full model	.846	—	.845	85.7	.315	.576	.250	.174	100	.207	.125	.642	.000
PPA's fixed 260 pairs	.833	−.013	.834	260.0	.297	1.000	.000	.000	61	.209	.162	.685	—
W/o content control	.843	−.003	.842	92.6	.250	.656	.235	.109	115	.165	.121	.694	.192
PARA (UniQA-B/16)													
Full model	.890	—	.929	104.6	.340	.540	.324	.136	123	.171	.080	.571	.000
PPA's fixed 260 pairs	.894	+.004	.931	260.0	.348	1.000	.000	.000	77	.183	.146	.584	—
W/o content control	.892	+.002	.929	104.6	.358	.564	.303	.134	108	.194	.105	.587	.000

Columns up to Neg. are those of Table S5. PPA's pool names no content level, so every concept is on every support and no concept is an image's own. Uncontr.: share of the top-three slots held by a collection-wide concept whose content-partialled correlation with the rating on the training images falls below the floor f that its marginal correlation clears, i.e. one the content-controlled test would have refused. It is zero for the full model by construction and undefined for PPA's pool.

the content tag is constant, so only the $\mathcal{T}_0^\star$ stage changes. On LAPIS, 43 pairs become 51: 14 drop out and 22 come in. On PARA, 7 of the 64 pairs are replaced. Training follows Appendix F, with one run per condition.

Whether selecting the vocabulary by the rating helps prediction depends on the dataset. Replacing the vocabulary with PPA's fixed one lowers the SRCC by .013 on LAPIS, beyond the seed range (.007). On PARA it instead rises by .004 (seed range .003). On both, fewer distinct concepts appear among the top three of the explanations (LAPIS 100 → 61, PARA 123 → 77).

Content control helps the explanations, not the prediction. Without content control, the change in SRCC stays within the seed range (LAPIS −.003, PARA +.002). However, 18 of the 51 pairs in $\mathcal{T}_0^\star$ of the uncontrolled pool would not reach the floor once content is controlled for. These pairs respond to what is depicted: whether it is figurative or abstract, and what it is (e.g., "Two-dimensional image." / "Dimensional image."). They hold .192 of the top-three slots, and at least one of them appears in the explanations of 48% of the test images. Content control thus removes paraphrases of "what is depicted" from the explanations without lowering prediction. On PARA, removing the control swaps only 7 of the 64 pairs in $\mathcal{T}_0^\star$, and none of them would fail the floor under the control. Content control matters for the explanations on datasets like LAPIS, where content and rating are strongly linked.

Filtering (Step 2). Table S1 gives the numbers of candidates and retained pairs. Among the pairs declared evaluative, the correlation in the data runs against the declaration for 24% on LAPIS and 8% on PARA. Keeping them with the poles swapped would give explanations in which the pole the wording calls better is not the highly rated side, so we reject them. Content control changes the filtering most on LAPIS, where content and rating are strongly tied. There, 24 pairs are eligible only under the marginal correlation ("Stable image." / "Unstable image." has .377 marginally and .047 after control; PARA has only 1 such pair). Among the level-specific candidates, 97 pairs on LAPIS and 110 on PARA are eligible only within their level ("Clean floral background." / "Distracting floral background." has .428 within flower paintings and .007 over the whole collection). These would be lost without the per-level test.

Ordering (Step 3). Taking the same number of candidates in order of $|r|$ leaves redundant paraphrases (Table S9). In PARA's $\mathcal{T}_0^\star$, the number of pairs with $|r| \geq 0.7$ rises from 58 to 103. For example, "Rich image." / "Drab image.", which the $|r|$ order would take, correlates at .844 with the selected "Exquisite image." / "Crude image.". Among the level-specific concepts, too, more of them nearly paraphrase $\mathcal{T}_0^\star$ (LAPIS 27 → 36, PARA 79 → 127). The difference is small in LAPIS's $\mathcal{T}_0^\star$, which has few eligible candidates. Whether the span of $\mathcal{T}_0^\star$ is projected out before ordering also matters little. What suppresses redundancy is mainly OMP itself.

Floor and threshold. Over the grid of floors $f \in \{0.05, 0.10, 0.15\}$ and thresholds $\tau \in \{0.005, 0.01, 0.02, 0.04\}$, the pool size ranges from tens to hundreds of pairs. The proxy SRCC, however, stays within .844–.849 on LAPIS and .873–.882 on PARA for $\tau \leq 0.02$. It drops only at $\tau = 0.04$, where the pool shrinks to a few dozen pairs. The default $(f, \tau) = (0.10, 0.01)$ lies in this flat region.

Table S9: The ordering of Step 3 against orderings that ignore redundancy, from the candidates alone.

	LAPIS (CLIP-B/16)				PARA (UniQA-B/16)			
Ordering	Mean $\lvert r\rvert$	Near	Eff. rank	CV R^2	Mean $\lvert r\rvert$	Near	Eff. rank	CV R^2
$\mathcal{T}_0^\star$: the same number of eligible candidates								
OMP (the method)	.173	0	.337	.350	.298	58	.112	.723
Largest $\lvert r\rvert$	.193	0	.295	.335	.419	103	.072	.710
Random	.182	—	.316	—	.290	—	.119	—
$\mathcal{T}_g^\star$: the same number per level, eligible inside the level								
Ordering	Max $\lvert r\rvert$ to $\mathcal{T}_0^\star$	Repeats	R^2 gain	Pool SRCC	Max $\lvert r\rvert$ to $\mathcal{T}_0^\star$	Repeats	R^2 gain	Pool SRCC
OMP after $\mathcal{T}_0^\star$ (the method)	.556	27	.083	.847	.619	79	.024	.880
OMP, $\mathcal{T}_0^\star$ kept in	.551	26	.076	.849	.626	79	.017	.879
Largest $\lvert r\rvert$	.594	36	.063	.843	.721	127	.014	.875

All on the training images, in rank space. $\mathcal{T}_0^\star$: mean off-diagonal $|r|$ between the chosen concepts (content-partialled), pairs with $|r| \geq 0.7$ (Near), effective rank over the set size, and the cross-validated R^2 of a ridge on them for the content-partialled rating; Random is the mean of 20 draws. $\mathcal{T}_g^\star$: inside each level, the mean over the chosen concepts of the largest $|r|$ with a $\mathcal{T}_0^\star$ concept, the concepts with such an $|r| \geq 0.7$ (Repeats, summed over levels), and the in-sample R^2 added to $\mathcal{T}_0^\star$, all weighted by the level's images; and the cross-validated SRCC of the whole pool (Pool SRCC).

Table S10: Components the concept-based methods share with the CBM literature. ✓: used; —: not used.

Component	FaTeX	PPA-T	C2J	In CBMs
Prediction linear in named concepts	✓	✓	✓	Koh et al. (2020)
Vocabulary generated by an LLM	✓	✓	—	Oikarinen et al. (2023); Yang et al. (2023)
Concept scored by CLIP similarity	✓	✓	—	Oikarinen et al. (2023); Yang et al. (2023)
Concept direction from examples	—	—	✓	Kim et al. (2017); Yuksekgonul et al. (2023)
Candidate filtering	✓	—	—	Oikarinen et al. (2023)
Compact, non-redundant subset	✓	—	—	Yang et al. (2023); Yan et al. (2023)
Sparse read-out	✓	—	✓	Du et al. (2026); Wong et al. (2021)
Weights from concept embeddings	✓	—	—	Du et al. (2026)

H Baseline Details

For the six methods named in the main text (Sec. 4.1), we describe the read-out and our settings. The training budget, learning-rate search, model selection and input pipeline are the same for all methods (Sec. 4.1, Appendix E). What follows are the remaining differences.

Why no CBM from outside IAA. We do not add general-purpose CBMs as baselines, because the concept-based methods we compare already build on their components (Table S10). FaTeX and PPA-T take their vocabulary from an LLM instead of annotations (Oikarinen et al., 2023; Yang et al., 2023; Menon & Vondrick, 2023) and score each concept by CLIP similarity. FaTeX also filters the candidates by form and duplicates (Oikarinen et al., 2023), keeps a compact, non-redundant subset (Yang et al., 2023; Yan et al., 2023), selects concepts per image with sparsemax, and computes the weights from the concept's text embedding (Du et al., 2026). C2J obtains concept directions from positive and negative examples (Kim et al., 2017) and fits a sparse linear read-out on a frozen encoder (Yuksekgonul et al., 2023; Wong et al., 2021); it is close to a post-hoc CBM applied to aesthetic attributes. The details differ, but the architectural gap to the CBM literature is small. What IAA changes lies elsewhere, in three points. *Definition*: the vocabulary is evaluative and names no object (Sec. 1), so a concept cannot be checked against annotations of what the image contains. FaTeX and PPA-T write a concept as an antonymic prompt pair (Wang et al., 2023), and C2J limits it to annotated attributes. *Selection*: selecting by association with the rating admits paraphrases of overall quality and words that merely track content. FaTeX selects with content controlled and removes paraphrases (Sec. 3.2). *Evaluation*: explanations of IAA have no ground truth, so their quality has to be measured with people (Sec. 4.2). A general-purpose CBM is designed for object-level classes and addresses none of these three points. The baselines instead cover the choices already made for IAA: a fixed LLM vocabulary (PPA-T) and annotated attributes (C2J).

MLP. A one-hidden-layer MLP outputs the score from the L^2-normalized image embedding and is fitted with the mean squared error. The hidden width is fixed to 256 and not tuned. Only the depth of fine-tuning follows the baselines: like PPA, it updates the last 4 blocks. The learning rate is split into

two groups as in Appendix E, and the randomly initialized head is trained at 10η. The backbone, the loss and the embedding the read-out reads are shared with FaTeX; the fine-tuning scope is not.

NIMA (Talebi & Milanfar, 2018). NIMA outputs logits for each level of the rating scale and is fitted to the rating histogram with the squared EMD. The score is the expectation of the predicted distribution. It thus needs a histogram over discrete levels and is undefined on datasets that do not provide one (Table 1).

Charm (Behrad et al., 2025). Charm is a *tokenization* method rather than a score formulation. It encodes images at their original resolution and aspect ratio, keeping only some regions at high resolution. It leaves the encoder unchanged, so the read-out is a single fully connected layer on the class token. Since Charm is designed for ViTs pre-trained on images alone, such as DINOv2, we use only its original backbone, DINOv2-S/14, and do not port it to the two CLIP backbones.

Gamma-T (Zhou et al., 2025b). Gamma-T keeps both encoders frozen and replaces the FFNs of the last 6 layers with a Mixture of Assessment Experts. The score is the expectation of learnable level values under the similarities to five-level prompts. Its only concept-like vocabulary is the five-level prompts. Since contributions pass through the mixture of experts, they cannot be decomposed per concept.

PPA-T (Hara & Otsuka, 2026). PPA-T is the only baseline that, like ours, composes the score from a distribution over prompts. It takes a convex combination of per-pair responses with softmax weights. It uses the fixed 260 pairs released by the original authors as they are. It applies neither the eligibility test nor the orthogonal ordering of Sec. 3.2, so the pool is the same for every dataset (Table S12). Its structural differences from FaTeX are that no weight ever becomes zero and that the convex combination has no per-image magnitude. The original paper focuses on controlling the evaluation criteria through prompts, not on interpretability. That we can extract per-concept contributions from PPA-T follows from its construction as a convex combination of per-pair responses.

C2J (Liu & Wagemans, 2026). C2J is a concept bottleneck on a frozen encoder. Concept directions are fitted with linear SVMs to the dataset's attribute annotations (26 styles and 7 genres for LAPIS, 9 aesthetic attributes for PARA; Table S12), and an Elastic Net read-out h is fitted on top. We use h alone, which the original paper calls *interpretable*. The search covers only the two Elastic Net hyperparameters (Sec. 4.1), whose values and selection the original paper does not state. We do not take the argmax of the validation SRCC. The concept directions are nearly collinear (on PARA the first eigenvalue of their Gram matrix holds 88% of the total), so as the penalty vanishes the weights split into cancelling positive and negative parts and their signs become arbitrary, while the validation SRCC barely changes (2×10^{-6} on PARA). We instead use the *one-standard-error rule* (Breiman et al., 1984; Hastie et al., 2009): among grid points within one standard error of the best validation SRCC, we take the one with the fewest non-zero weights. The standard error comes from resampling the single validation split ($B = 1000$) in place of cross-validation folds. The resulting weights are close to those of the original paper: positive on AADB and PARA at the reported backbones, largest for good lighting (AADB) and composition (PARA); on LAPIS, styles and genres mix signs. The argmax would raise C2J's test SRCC by .003–.015 (Table S11).

Table S11: C2J under the other selection rule: the argmax of the validation SRCC over the same grid. Table 1 reports the same settings under the one-standard-error rule.

Dataset	Backbone	λ	α	Live / negative	SRCC	PLCC
AADB	CLIP-B/16	1.0×10^{-7}	1.0	$11/11, 2$	.616	.617
	UniQA-B/16	1.8×10^{-5}	1.0	$10/11, 1$	.644	.649
	CLIP-RN50	1.0×10^{-4}	1.0	$9/11, 0$	.624	.622
LAPIS	CLIP-B/16	1.8×10^{-8}	0.25	$33/33, 15$	.823	.820
	UniQA-B/16	3.2×10^{-7}	1.0	$32/33, 16$	.825	.824
	CLIP-RN50	3.2×10^{-8}	1.0	$33/33, 18$	.807	.805
PARA	CLIP-B/16	0	—	$9/9, 2$	.803	.835
	UniQA-B/16	1.0×10^{-8}	1.0	$9/9, 3$	.823	.861
	CLIP-RN50	0	—	$9/9, 3$	.781	.810

LAPIS and PARA are the median of three seeds, as in Table 1; AADB is one seed. Live / negative: concepts with a non-zero weight, and how many of those weights are negative. α is undefined at $\lambda = 0$, where every α gives the same least-squares fit. The rule selects a penalty at or next to the unpenalized end in six of the nine settings, and the weights there cancel: $\sum_i |w_i| / |\sum_i w_i|$ reaches 14.8 on LAPIS and 2.9 on PARA.

I HUMAN EVALUATION DETAILS

This section details the human evaluation of Sec. 4.2. We give the compared methods (Table S12), the explanation templates (Table S14), the questions of Task 1 (Table S15), the exclusion criteria (Table S19), and the scale and estimation (Table S20).

Compared methods. All four methods of Table S12 decompose the prediction into a sum of per-concept contributions c_t, so explanations are built by the same procedure for all of them. For each method, we take the backbone with the highest median test SRCC in Table 1 and explain its median run. w/o $\mathcal{T}_g^\star$ uses FaTeX's backbone and its own median run. Its vocabulary coincides with FaTeX's $\mathcal{T}_0^\star$, so the two differ only in the level-specific concepts.

Definition of the contributions. Explanations are built from the exact additive decomposition $\hat{y} = a + \sum_t c_t$ (Table S13), with the baseline a placed as in each method's original paper. PPA-T's paper defines no per-image explanation, so we take its top three concepts by $|c_t|$, as for the other methods.

Explanations. The top three concepts by $|c_t|$ are filled into the templates of Table S14 according to their signs (example in Fig. 1, center). Templates are shared by all methods. We show neither the predicted value, nor weights or responses, nor any chart. The level word is taken from the ground-truth band, not the prediction, and is the same for every method, so the methods differ only in their three concepts. With predicted bands, the level words of the two sides would disagree in 14–27% of the Task 1 pairs on PARA, and the side closer to the rater's own R could win regardless of the concepts. On LAPIS, each concept phrase is set in bold for every method, with the instruction "Each bold phrase is one factor.", so that a modifier such as C2J's "absence of" is not read as applying to every phrase after it. The LLM judge receives the same text.

Concept phrases. For FaTeX, w/o $\mathcal{T}_g^\star$ and PPA-T, the phrase is the concept's clause without its initial capital and final period. A final noun naming the medium (photo, photograph, picture, painting, artwork) is replaced with image, e.g., "Harsh lighting photo." becomes harsh lighting image. Nothing else is rewritten. Within a dataset the medium is constant (LAPIS: paintings; PARA: photos), yet PPA-T's fixed prompts call paintings photos and only FaTeX's $\mathcal{T}_g^\star$ can say painting; unifying the noun keeps this from becoming a cue. We replace rather than drop it because it is the head the adjectives modify (dropping it from "Noisy animal picture." would make noisy modify the animal). C2J's concepts are annotation names, so we name their two sides after the examples each concept is trained to separate: for PARA, the top against the bottom 100 images on the rating (good / poor composition, etc.); for LAPIS, a class against its siblings (landscape subject / absence of a landscape subject, etc.).

Selection of evaluation images. With scores normalized to $[0, 1]$, we use only test images on which every method satisfies $|\hat{y} - y| \leq \delta$, where δ is 0.5 times the test standard deviation of y, fixed in advance. This adapts to regression the procedure of Label-free CBM (Oikarinen et al., 2023, arXiv:2304.06129, App. B). Such images cluster around the mean, so we stratify the selection into the bottom, middle and top bands of the ground truth (40 images each, 120 per dataset). Within a band, images are spread as evenly as possible over the levels of the content tag, starting from the sparsest; drawing at random would, e.g., make 86% of LAPIS's bottom band abstract.

Task 1. The screen proceeds image → R → two explanations (left and right) → Q0 → Q1 and Q2 (Table S15). R is asked before the explanations and cannot be revisited. Its wording and scale follow each dataset's paper: a 0–100 slider for LAPIS (Maerten et al., 2025) and a 1–5 slider in steps of 0.5 for PARA (Yang et al., 2022). Q0–Q2 use a four-point forced choice (Clearly A / Slightly A / Slightly B / Clearly B) without a neutral option (Kayser et al., 2021), collapsed to A / B for the analysis. Below the explanations we add "Judge each description as an explanation — not by whether you like the image or agree with the rating." Each rater receives the three pairs in equal numbers, mixed, and sees each image once, so that FaTeX's explanation is never read repeatedly (Zajonc, 1968; Reber et al., 1998). Left and right are balanced per pair within a set. Results are in Fig. 3(a)–(c) and, by band, in Table S16.

Task 2. We show one explanation and four images and ask "The description below was written about one of the four images. Which one is it?". The three wrong images share the content tag (LAPIS: seven genres; PARA: eight subject categories) and the ground-truth band of the correct one, so the content or the rating level alone hardly leads to the answer. The sets of four are shared by all methods. The correct image's position is balanced over the four positions within a set. Fig. S5 shows an item.

Table S12: Methods compared in the human evaluation. Vocabulary sizes are for the backbone each method uses.

Method	Vocabulary source	Vocabulary size (LAPIS / PARA)	Concept selection	Decomposition
C2J (Liu & Wagemans, 2026)	Dataset annotations (LAPIS: 26 styles, 7 genres; PARA: 9 aesthetic attributes)	33 / 9	None (linear weight on every concept)	Additive
PPA-T (Hara & Otsuka, 2026)	Fixed prompt pairs released with the paper (shared by all datasets)	260 / 260	softmax weights (never zero)	Convex combination
FaTeX	LLM generation → eligibility test → orthogonal ordering (Sec. 3.2)	244 / 296	sparsemax + learned importance	Additive
w/o $\mathcal{T}_g^\star$	As FaTeX, but $\mathcal{T}_0^\star$ only (no level-specific concepts)	43 / 64	As FaTeX	Additive

Table S13: How each method's prediction is decomposed for the explanation, $\hat{y} = a + \sum_t c_t$. Here c_t is on the rating scale, so FaTeX's is s times that of Sec. 3.1. $q(x, t)$: PPA-T's two-way softmax score of pair t.

Method	Baseline a	Contribution c_t	What the sign is read against
FaTeX, w/o $\mathcal{T}_g^\star$	$\bar{y}$	$s\,\rho(x)\,p(t\|x)\,d_\theta(x,t)$	The mean response over the training set, which the method's own read-out is built on
PPA-T (Hara & Otsuka, 2026)	$1/2$	$p(t\|x)\,\big(q(x,t) - 1/2\big)$	The pair's own midpoint: Eq. (1) of the original paper is a two-way softmax over the two clauses
C2J (Liu & Wagemans, 2026)	b	$w_i\, f_C^i(x)$	Zero, as the original paper plots the concept projections (its Figs. 3, 5, 7)

Wrong images are drawn from outside the 120 evaluation images, each used once. They are not restricted to the test split, as sparse genres would run short, and no model is run on them; the test split's tertiles define their bands. Each rater receives the four methods in equal numbers and sees each correct image once. Results by band are in Table S18.

Assignment. Each rater is assigned one fixed set, which fixes the images, explanations, left and right positions, and the order of presentation. A set contains the pairs (Task 2: methods) and the bands in equal numbers, with no image twice. Sets are issued in random order, and an excluded rater's set is re-issued as is. The three practice trials include one of each pair (method). Task 1 raters are excluded from Task 2, which uses the same images.

Controls and exclusion. The first three trials of each task are practice, drawn from outside the evaluation images and left out of the analysis. A session takes about 30 minutes (ITU-R, 2019) and is paid at £10 per hour, including for excluded raters. The exclusion criteria (Table S19) are fixed in advance (Meade & Craig, 2012; Ward & Meade, 2023). LAPIS keeps the genre nude painting, but works depicting sexual acts, close-ups of genitals or sexual violence (30 in LAPIS) are never shown. Since titles do not reveal them, every image used is checked by eye. The recruitment notice states that paintings containing nudity will be shown.

Attention-check items. Check items look like real trials, use images outside the evaluation, practice and wrong images, match bands so the level word gives no cue, and use FaTeX's explanations from the reported run. They are chosen so that the explanation names the image's content (portrait, landscape, food, animal, etc.). In Task 1, one side is replaced with the explanation of an image with different content; only Q0 and Q1 are scored. In Task 2, the four images show mutually different content. Per dataset there are four Task 1 pairs and three Task 2 trials. Task 1 uses four pairs, two with the genuine side on the left and two on the right, so that a rater who always presses one side fails two and is excluded.

Table S14: Explanation templates. {level}: low, average or high, from the band of the ground-truth score (the three bands the images are drawn from), the same word for every method; $\{t_1\}$–$\{t_3\}$: top three concepts by $|c_t|$, positive ones first and each sign in descending $|c_t|$.

Signs of the top three	Template
3 positive	`This image is aesthetically {level}. Positive factors include {`t_1`}, {`t_2`} and {`t_3`}.`
2 positive, 1 negative	`This image is aesthetically {level}. Positive factors include {`t_1`} and {`t_2`}. A negative factor is {`t_3`}.`
1 positive, 2 negative	`This image is aesthetically {level}. A positive factor is {`t_1`}. Negative factors include {`t_2`} and {`t_3`}.`
3 negative	`This image is aesthetically {level}. Negative factors include {`t_1`}, {`t_2`} and {`t_3`}.`

Table S15: Task 1 questions. R follows each dataset's own paper verbatim, in both wording and scale; Q0–Q2 compare the two explanations. —: no prior item; defined by us.

	Aspect	Question shown	Prior work
R	Own aesthetic rating (covariate)	LAPIS: `How aesthetic do you find this image? Please consider how beautiful the image is, how much you like it, prefer it and are drawn to it.` PARA: `Aesthetics score` — `Comprehensive aesthetic judgement`	Maerten et al., 2025; Yang et al., 2022
Q0	Overall	`Overall, which description is the better explanation?`	Oikarinen et al., 2023, App. B; Aggarwal et al., 2021
Q1	*Fit to the image*	`Which description lists features that are actually present in this image?`	Oikarinen et al., 2023, App. B; Kayser et al., 2021
Q2	*Specificity*	`Which description says something specific about this image, rather than something that could be said of any image?`	—

Rater information. After the task, we ask for age, gender, color vision, and experience in art and in photography. Gender and color vision use the options of LAPIS (Maerten et al., 2025); experience uses PARA's four levels (Yang et al., 2022).

Scale and analysis. Raters are recruited on Prolific separately for the two tasks: per dataset, 30 for Task 1 and 16 for Task 2, plus replacements (Table S20). The unit of analysis is the image. 95% CIs use a two-way cluster bootstrap over images and raters for Task 1 and per-image paired differences for Task 2. Q0–Q2 are treated alike. CIs are reported per pair and question, without a decision rule or multiplicity correction.

LLM evaluation. The LLM (Claude Opus 5) receives the same images, explanations and questions in the same presentation (Fig. S3). We do not use Gemini, which generated the concept pool, since MLLMs tend to prefer text from their own family (Koyama et al., 2026). Against position bias (Zheng et al., 2023), each item is asked again with the presentation swapped (Task 1: sides exchanged; Task 2: images reversed), and we report the mean of the two (Wang et al., 2024). Intervals resample images with both answers kept together; otherwise estimation follows the human evaluation. Table S21 gives each presentation separately.

Table S16: Task 1 Q0 preference (%) for FaTeX, by ground-truth band.

FaTeX vs.	Band	LAPIS Human	LAPIS LLM (Claude Opus 5)	PARA Human	PARA LLM (Claude Opus 5)
w/o $\mathcal{T}_g^\star$	Low	57.5 [38.1, 76.2]	56.2 [41.2, 70.0]	61.3 [42.0, 78.7]	55.0 [40.0, 68.8]
	Middle	66.2 [47.5, 83.3]	80.0 [68.8, 90.0]*	67.5 [48.9, 84.0]	65.0 [51.2, 78.8]*
	High	67.5 [48.2, 85.2]	77.5 [66.2, 88.8]*	67.5 [48.9, 83.8]	63.7 [50.0, 76.2]
C2J (Liu & Wagemans, 2026)	Low	50.0 [32.5, 67.9]	71.2 [57.5, 82.5]*	57.5 [37.6, 76.4]	48.8 [35.0, 62.5]
	Middle	53.8 [32.6, 73.7]	71.2 [57.5, 83.8]*	72.5 [53.4, 89.0]*	80.0 [68.8, 91.2]*
	High	65.0 [44.9, 82.7]	93.8 [85.0, 100.0]*	78.8 [62.9, 91.6]*	93.8 [86.2, 100.0]*
PPA-T (Hara & Otsuka, 2026)	Low	61.3 [41.6, 80.0]	90.0 [81.2, 97.5]*	52.5 [33.3, 71.4]	62.5 [50.0, 75.0]
	Middle	80.0 [63.9, 93.0]*	95.0 [88.8, 100.0]*	77.5 [58.6, 92.4]*	50.0 [36.2, 63.7]
	High	72.5 [53.5, 88.0]*	95.0 [87.5, 100.0]*	65.0 [42.3, 84.4]	48.8 [36.2, 61.3]

Each band has 40 images. 95% CIs are from a bootstrap clustered by image and, for humans, also by rater. *: CI excludes 50%. LLM: the mean of the presentation shown to the raters and its swap, each image's two answers resampled together; each presentation is in Table S21. Human–LLM agreement over all bands, as the share of items on which both favor the same explanation: 68.8% on LAPIS and 63.2% on PARA. The LLM columns are measured on the reported models.

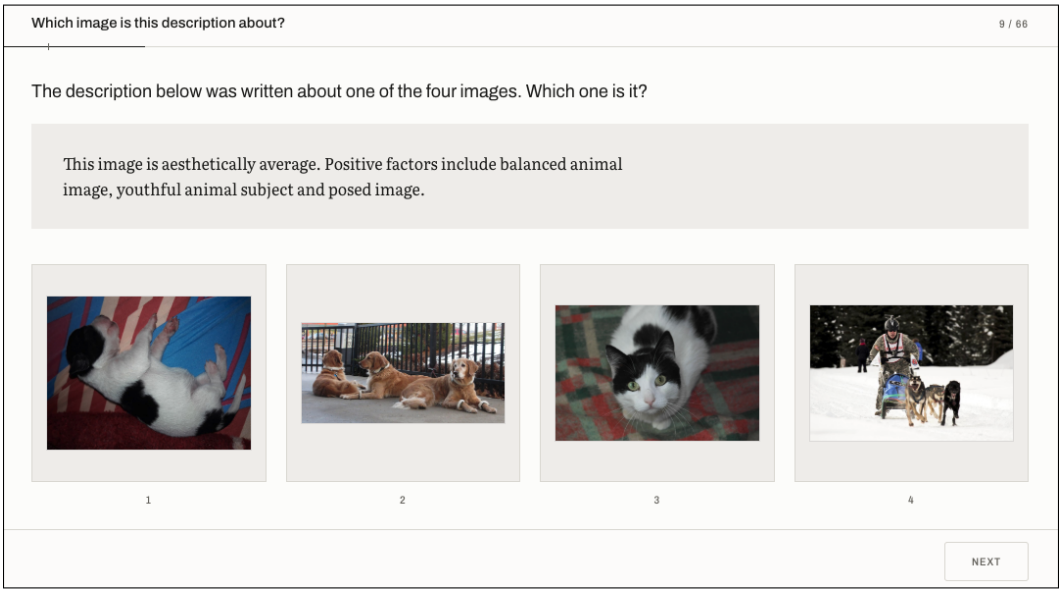

Figure S5: A Task 2 item on PARA, as shown to raters. The explanation comes from one of the four methods, which are not named on the screen. The three wrong images share the content tag (animal) and the ground-truth band of the correct one.

J Human Evaluation: Participants and Released Data

Participants. Table S22 counts the participants of each study. Those released before finishing their set abandoned it; those who finished but met a criterion of Table S19 were excluded. Every freed set was re-issued, so each study ends with the planned 30 (Task 1) or 16 (Task 2) raters. Fig. S6 shows who the 92 included raters were (median age 32, range 19–67). None reported a color vision deficiency.

Task 2 by method and band. Table S17 gives the accuracies of Fig. 3(d) with their differences from FaTeX, and Table S18 splits them by ground-truth band. In the bottom band, where the four images resemble each other more, FaTeX's human accuracy is below that in the middle band (LAPIS 36.2% vs. 45.0%, PARA 31.2% vs. 41.2%), and on LAPIS every difference between methods has a CI including 0.

Table S17: Task 2 (identifying the described image): accuracy (%) and difference from FaTeX.

	LAPIS					PARA				
		Human		LLM (Claude Opus 5)			Human		LLM (Claude Opus 5)	
Method	SRCC	Acc.	Diff. [95% CI]	Acc.	Diff. [95% CI]	SRCC	Acc.	Diff. [95% CI]	Acc.	Diff. [95% CI]
FaTeX	.846	37.5		53.8		.891	36.2		40.8	
w/o $\mathcal{T}_g^\star$	.832	27.5	$+10.0$ [1.7, 18.3]*	35.4	$+18.3$ [8.8, 27.9]*	.891	29.2	$+7.1$ [-1.7, 15.8]	34.6	$+6.2$ [-3.3, 15.8]
C2J (Liu & Wagemans, 2026)	.818	37.5	$+0.0$ [-9.6, 10.0]	50.4	$+3.3$ [-7.1, 13.3]	.818	17.1	$+19.2$ [10.8, 27.1]*	25.0	$+15.8$ [6.2, 25.4]*
PPA-T (Hara & Otsuka, 2026)	.853	22.9	$+14.6$ [7.9, 21.2]*	25.0	$+28.7$ [18.8, 38.3]*	.908	25.0	$+11.2$ [3.3, 18.8]*	25.0	$+15.8$ [7.1, 24.2]*
Human–LLM agree.		39.7					37.1			

SRCC: test SRCC of the explained model, the median-of-three run on the backbone whose median is best in Table 1 (w/o $\mathcal{T}_g^\star$: FaTeX's backbone, its own median run). Chance is 25%. Diff.: FaTeX − method, with 95% CIs from an image-level paired bootstrap. *: CI excludes 0. LLM: the mean of the two presentations, as in Fig. 3. Agreement: share of trials in which humans and the LLM chose the same image. All columns are measured on the reported models. The accuracies are drawn in Fig. 3(d); results by band are in Table S18.

Table S18: Task 2 accuracy (%) and difference from FaTeX, by ground-truth band.

		LAPIS				PARA			
		Human		LLM (Claude Opus 5)		Human		LLM (Claude Opus 5)	
Band	Method	Acc.	Diff. [95% CI]	Acc.	Diff. [95% CI]	Acc.	Diff. [95% CI]	Acc.	Diff. [95% CI]
Low	FaTeX	36.2		38.8		31.2		36.2	
	w/o $\mathcal{T}_g^\star$	25.0	$+11.2$ [-1.2, 23.8]	32.5	$+6.2$ [-10.0, 22.5]	31.2	$+0.0$ [-13.8, 13.8]	37.5	-1.2 [-18.8, 16.2]
	C2J	35.0	$+1.2$ [-15.0, 16.3]	41.2	-2.5 [-20.0, 15.0]	12.5	$+18.8$ [6.2, 31.2]*	11.2	$+25.0$ [11.2, 40.0]*
	PPA-T	28.7	$+7.5$ [-5.0, 20.0]	28.7	$+10.0$ [-3.8, 25.0]	28.7	$+2.5$ [-11.2, 16.2]	22.5	$+13.8$ [1.2, 27.5]*
Middle	FaTeX	45.0		68.8		41.2		52.5	
	w/o $\mathcal{T}_g^\star$	25.0	$+20.0$ [5.0, 33.8]*	31.2	$+37.5$ [21.2, 52.5]*	38.8	$+2.5$ [-12.5, 17.5]	38.8	$+13.8$ [0.0, 28.7]
	C2J	36.2	$+8.8$ [-8.8, 26.2]	56.2	$+12.5$ [-6.2, 31.2]	16.2	$+25.0$ [11.2, 38.8]*	31.2	$+21.2$ [6.2, 37.5]*
	PPA-T	21.2	$+23.8$ [12.5, 35.0]*	25.0	$+43.8$ [26.2, 61.3]*	22.5	$+18.8$ [7.5, 30.0]*	35.0	$+17.5$ [0.0, 33.8]
High	FaTeX	31.2		53.8		36.2		33.8	
	w/o $\mathcal{T}_g^\star$	32.5	-1.2 [-16.2, 15.0]	42.5	$+11.2$ [-5.0, 27.5]	17.5	$+18.8$ [2.5, 35.0]*	27.5	$+6.2$ [-8.8, 21.2]
	C2J	41.2	-10.0 [-26.2, 6.2]	53.8	$+0.0$ [-16.2, 16.2]	22.5	$+13.8$ [-2.5, 30.0]	32.5	$+1.2$ [-17.5, 20.0]
	PPA-T	18.8	$+12.5$ [2.5, 22.5]*	21.2	$+32.5$ [16.2, 48.8]*	23.8	$+12.5$ [0.0, 25.0]	17.5	$+16.2$ [1.2, 31.2]*

Each band has 40 images. Notation as in Table S17.

K Further Limitations and Future Work

Dependence on the concepts a VLM can capture. Whether an explanation of FaTeX is correct about an image depends on how well the backbone captures each concept. Both the response and the applicability are built from its embeddings, without per-concept supervision. Zero-shot CLIP scores capture to some degree the level of brightness or noise and abstract impressions such as happy/sad. Antonymic-pair responses also agree with human judgments on focus and exposure (Hara & Otsuka, 2026), and our Q1 results of Task 1 are consistent with this. Photographic terms such as the rule of thirds or shallow depth of field, however, are hard for them to capture (Wang et al., 2023). Such terms are among our attribute seeds. UniQA continues the pre-training of CLIP on sentences describing image quality and aesthetics. Its effect, however, has been verified only on overall score prediction and on a few quality words such as blurry (Zhou et al., 2025a). Whether it captures technical terms has not been tested. Future work should map, concept by concept, what the backbone is good and bad at, and compensate for the aspects it is weak at.

Information leakage through the weights. In concept bottleneck models, concept values can carry task information beyond the concepts they name, so the explanation no longer reflects what the prediction relies on (Mahinpei et al., 2021; Havasi et al., 2022). In FaTeX, the responses are comparatively safe from this: they read the frozen encoder, and training adds only a per-concept gain γ_t and center μ_t shared by all images. The weights are not. The importance $w_t(x)$ and the magnitude $\rho(x)$ read the trained copy I_ϕ, and borrowing the allocation from another image lowers the SRCC by .05 on LAPIS and .10 on PARA (Appendix F). We cannot tell whether this is legitimate per-image relevance or unnamed task information, and we have not run a dedicated leakage test (Parisini et al., 2025); some leakage is hard to avoid with an incomplete concept set such as aesthetic vocabulary (Havasi et al., 2022; Espinosa Zarlenga, 2026). It did not, however, make the explanations less useful to people (Sec. 4.2).

Choosing the levels of the content tag. We chose the levels of the content tag at a granularity where the images of each level are not too sparse and still separate image content to some extent. Finer levels could make concepts more closely tied to content. However, the smaller n_g, the higher the eligibility floor f_g (Appendix A), and the harder it becomes for level-specific concepts to be admitted.

Table S19: Rater exclusion criteria.

Stage	Procedure	Excluded if
Recruitment	Prolific filters: approval rate $\geq 98\%$ (Oikarinen et al., 2023, App. B), ≥ 100 approved submissions, fluent English, desktop only	Not eligible to participate
Task 1	4 check pairs in which one explanation is replaced by that of an image showing different content (the genuine one on the left in two, on the right in two)	Picks the replaced side on Q0 or Q1 in 2 or more of the 4 pairs
Task 2	3 check trials whose four images each show different content	2 or more of the 3 wrong
Both tasks	Time per trial (participant's median)	Below $1/3$ of the median over all participants

Table S20: Scale and estimation of the human evaluation, per dataset.

	Task 1: pairwise comparison	Task 2: image identification
Per set	24 images (R + 3 questions; 8 per pair) + 4 check pairs	60 trials (15 per method) + 3 checks
Duration	About 30 min	About 30 min
Design	120 images $\times$ 2 raters $\times$ 3 pairs $= 720$ ratings	120 images $\times$ 4 methods $\times$ 2 raters $= 960$ ratings
Sets, one per rater	30	16
Estimand	Preference for FaTeX	Accuracy difference (FaTeX $-$ method)
Smallest effect of interest	60:40	$+10$ points
Expected 95% CI half-width	± 9 points	± 9 points

The smallest effect of interest sets the scale only; results are reported as intervals, with no test of difference or equivalence. An excluded rater is replaced by re-issuing their set, so the number recruited grows with the exclusion rate.

Our levels are just one instance of this trade-off. Researchers and practitioners can choose levels to suit the purpose of their own models. Tags that are not distributed with a dataset can be assigned with an MLLM, as we did for AADB.

Table S21: LLM evaluation under the two presentations.

Method	LAPIS		PARA	
	As shown	Swapped	As shown	Swapped
Task 1, Q0: preference (%) for FaTeX				
w/o $\mathcal{T}_g^\star$	72.5	70.0	59.2	63.3
C2J (Liu & Wagemans, 2026)	80.0	77.5	72.5	75.8
PPA-T (Hara & Otsuka, 2026)	92.5	94.2	50.8	56.7
Task 2: accuracy (%)				
FaTeX	55.0	52.5	40.0	41.7
w/o $\mathcal{T}_g^\star$	35.0	35.8	32.5	36.7
C2J (Liu & Wagemans, 2026)	53.3	47.5	24.2	25.8
PPA-T (Hara & Otsuka, 2026)	24.2	25.8	20.0	30.0

As shown: the presentation the raters saw. Swapped: Task 1 with the two explanations exchanged, Task 2 with the four images in reverse order. The LLM values in Fig. 3 and Table S17 are the mean of the two. The same answer under both presentations: in Task 1 Q0, 90.0% of the items on LAPIS and 82.2% on PARA; in Task 2, 70.0–80.0% and 71.7–86.7% of the trials across methods. Chance is 25% in Task 2.

Table S22: Participants per study.

		Started	Abandoned	Excluded (checks)	Excluded (time)	Included
LAPIS	Task 1	39	2	5	2	30
	Task 2	16	0	0	0	16
PARA	Task 1	31	0	1	0	30
	Task 2	16	0	0	0	16

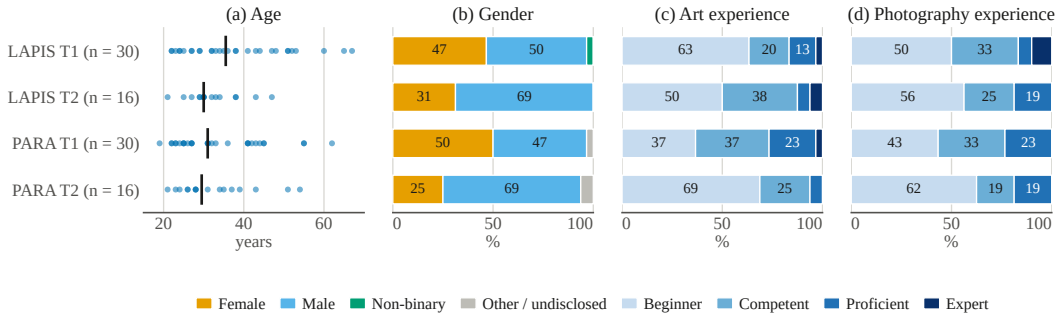

Figure S6: Included raters of the human evaluation, per dataset and task (T1, T2). (a) Each dot is a rater; the bar is the median. (b)–(d) Share of raters (%) per answer.