Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TAR-Net: Theme-Aware Aesthetic Ranking for No-Reference Quality Assessment of Cross-Theme Visual Designs

Xiaohua Wang

wangxiaohua@sdxiehe.edu.cn

Shandong Xiehe University

Xupeng Yao

Qilu University of technology

Bin Yu

Ping Dingshan University, Urban-Rural Integration Demonstration Zone

Research Article

Keywords: no-reference image quality assessment, image aesthetic assessment, cross-theme generalisation, vision–language models, feature-wise modulation, pairwise ranking learning

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11592279/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

TAR-Net: Theme-Aware Aesthetic Ranking for No-Reference Quality Assessment of Cross-Theme Visual Designs

Xiaohua Wang[1]*, Xupeng Yao[2], and Bin Yu[3]

[1] College of Computer Science, Shandong Xiehe University, 6277 Jiqing Road, Licheng District, Jinan 250109, China

[2] College of Arts and Design, Qilu University of technology, 3501 Daxue Road, Changqing District, Jinan 250353, China

[3] Ping Dingshan University, South Section of Weilai Road, Urban-Rural Integration Demonstration Zone, Ping Dingshan 467000, China

*Correspondence: wangxiaohua@sdxiehe.edu.cn

Abstract

Aesthetic ranking of visual designs deteriorates when the test theme differs from the themes a model was trained on, because the perceptual weight of colour density, whitespace, object prominence and typography hierarchy varies with communication intent. We introduce TAR-Net, a theme-aware aesthetic ranking network for no-reference scoring of cross-theme visual designs. A frozen text encoder converts a natural-language theme descriptor into a compact vector that drives a low-rank modulation at three intermediate transformer blocks. An aesthetic-decomposition module fuses colour-harmony, learned-saliency and typography-alignment streams before a margin-ranking head, and a paraphrase-consistency regulariser stabilises the theme-invariant representation while leaving the theme-aware branch free to specialise. On a six-theme benchmark of 6,985 design images (editorial, e-commerce, education, finance, healthcare, travel), evaluated under random-split and held-out-theme protocols, TAR-Net raises the cross-theme Spearman correlation from 0.719 to 0.743 over the strongest baseline — a substantial and consistent improvement across five seeds and all six held-out themes, supported by paired, permutation, bootstrap and mixed-effects analyses. The benchmark protocol, theme descriptors, per-seed metrics and reference implementation are released for independent verification.

Keywords: no-reference image quality assessment; image aesthetic assessment; cross-theme generalisation; vision–language models; feature-wise modulation; pairwise ranking learning.

Automated review of visual designs is now embedded in editorial pipelines, e-commerce front pages, course-cover generation, financial reporting templates, healthcare communication and travel-promotion services. Each surface places a different premium on colour density, information weight, object prominence, typography and whitespace, so the human notion of "good design" is theme-dependent. Aesthetic scoring models trained on a pooled mixture of themes can attain low pointwise error on a uniformly sampled random split, yet still misorder competing candidates when the test theme is unfamiliar. Because review queues forward only the top three or four candidates to a human reviewer, the practical quality of a scoring function is the *order* it imposes on a candidate set rather than its pointwise score. This paper studies *theme-disjoint ranking* as the target task, in which the test theme is held out from training so that random-split metrics cannot be inflated by theme-correlated shortcuts, near-duplicate campaign families or merged rating distributions.

A natural design pattern for cross-theme scoring is to condition the model on a theme cue. Three families of conditioning have been explored: theme-label concatenation, in which a one-hot or learned theme code is appended to the score head; prompt-based vision–language conditioning, in which a textual theme description is consumed at the input or attention level by a CLIP-style backbone[13–15]; and theme-aware backbones such as TANet[17] and TAVAR[18], which inject theme-specific branches or attribute graphs into the feature extractor. None of these conditioning choices is fully satisfactory under theme shift: a label-conditioned head defaults to the unconditional model when the held-out label was never observed; prompt conditioning at the input level often over-relies on global colour or template cues; and categorical theme branches do not generalise to themes for which no branch was trained. We propose instead to modulate *intermediate* visual features with a text-derived theme descriptor, and to regularise the *theme-invariant* representation with paraphrase consistency so that the response to a theme is genuinely conditioned on its semantics rather than memorised from its label.

Image aesthetic assessment has matured from handcrafted descriptors of colour, texture and composition[1–3] to deep regressors and distribution heads[4–7], multi-scale transformer quality models[8] and layout-aware graphical reasoning[9–11]. CLIP-based aesthetic scoring[14,15] and the LIQE vision–language quality model[16] use natural-language correspondence to make scoring more transferable. Recent multimodal scorers — including UNIAA[21], attribute-driven hierarchical prompts[23], PromptIQA[24], next-token numerical scoring[25] and the Gamma mixture-of-experts assessor[26] — instruction-tune multimodal large language models for numerical aesthetic scoring, and AesthetiQ[27] aligns layout aesthetics through preference learning. Feature-wise modulation has its own substantial lineage in conditional batch normalisation[28], FiLM[29], adapter modules[30] and low-rank adapters[31,32]; TAR-Net uses an intermediate-layer, text-conditioned FiLM-style operator and combines it with a paraphrase-consistency objective and a theme-invariant branch — a configuration that has not appeared in prior aesthetic-assessment work. Domain-generalisation studies[19,20] show that fair cross-domain comparison requires fixed splits, explicit model-selection rules and reproducible evaluation scripts, all of which are adopted in the protocol reported here.

The contribution is fourfold. First, theme-disjoint visual design ranking is formalised, and the empirical gap between random-split and theme-disjoint evaluation is quantified on the same data. Second, TAR-Net combines Theme-Aware Modulation (TAM), a text-conditioned low-rank FiLM operator inserted at three intermediate ViT blocks, with an Aesthetic Decomposition Module (ADM) that fuses harmony, saliency and typography evidence, and with a two-branch concatenation that exposes both theme-aware and theme-invariant descriptors to a margin-ranking head. Third, a cross-theme descriptor-consistency loss regularises only the theme-invariant branch, leaving the theme-aware branch free to specialise; an ablation shows this term contributes +0.015 Spearman correlation under the cross-theme protocol. Fourth, every primary comparison reports five-seed mean ± standard deviation, normal-theory 95% confidence intervals computed from the seed-level variance, Cohen's d_z and Holm–Bonferroni-corrected paired t-tests, with split files, descriptors, paraphrase pools, baseline configurations and per-seed metrics released to support reproducible review.

The remainder of the paper specifies the benchmark and TAR-Net (Methods), reports random-split and theme-disjoint experiments with statistical tests and ablations (Results), and interprets the cross-theme pattern, limitations and future work (Discussion).

Methods

Six-theme benchmark and calibration

The benchmark contains 6,985 design images sampled from public aesthetic and layout-aware datasets with permissive research licences[3,10–12,17,22]. The six themes — Editorial, E-commerce, Education, Finance, Healthcare and Travel — were chosen because they capture common visual-communication intents that differ along palette discipline, information density, object prominence, typography hierarchy and acceptable whitespace. Each candidate image was assigned to a single primary theme through a two-step conservative procedure. We first used source-level metadata, tags and category labels to obtain an initial theme mapping, then confirmed the assignment against the visual content of the image itself. Samples with mixed or unclear intent were retained only when their dominant communicative purpose could be identified from visible cues — product hero with price, instructional structure, financial numerals or charts, clinical or wellness symbols, scenic destinations or editorial cover hierarchy. Ambiguous samples were excluded to reduce theme-label noise and to prevent theme-correlated leakage in the theme-disjoint protocol.

Theme labels were assigned by three independent annotators with experience in graphic-design review (two industry designers and one human–computer-interaction researcher) who independently mapped each image into one of the six themes or into an "exclude" bin; annotators were blinded to one another's decisions and to the source dataset of each image. The majority-vote theme was retained when at least two annotators agreed and the third did not select "exclude"; all three-way disagreements (5.1% of the candidate pool) were adjudicated by a fourth senior annotator. Inter-rater agreement on the per-image theme assignment was substantial: Fleiss' $\kappa = 0.78$ across the six themes plus the exclude bin, with mean pairwise Cohen's $\kappa = 0.79$ (range 0.76 – 0.81) across the three annotator pairs. Samples without a majority decision (3.2% of the pre-filter pool) or with at least one "exclude" vote that survived adjudication (8.4%) were removed from the candidate pool before pHash deduplication. The per-sample annotator votes, the κ matrix and the adjudication log are released alongside the benchmark so that an independent reviewer can audit the assignment protocol end to end.

Perceptual-hash (pHash) duplicate removal was applied across the constructed pool with a Hamming threshold of eight bits to remove near-identical recompressions, crops and template duplicates that would otherwise contaminate the theme-disjoint protocol. Mean opinion scores (MOS) from the source datasets were normalised to a common interval through a min–max linear map to [0, 1] computed per fold on the training MOS values only (the held-out theme's MOS values play no role in fitting the calibration scale, which removes the test-MOS leakage flagged during review). Source datasets do not share a common rater pool, so a single pooled Cronbach α is not defined on the merged data; reliability is therefore reported as the sample-size-weighted average of per-source ICC(2, *k*)[36] and Kendall *W*, computed on the rating matrices retained after source-level quality filtering. For every retained sample, the supplementary archive provides the source dataset name, source image identifier, mapped theme, normalised MOS, original score range, pHash value, exclusion flag and split assignment under both protocols. Source images are not redistributed when source licences prohibit redistribution; in that case, image identifiers and download instructions are provided.

Table 1 summarises the theme-wise profile after filtering. Supplementary Table S1 lists the reliability and audit statistics. Supplementary Fig. S1 shows the resulting theme distribution and Supplementary Fig. S2 shows the theme-wise opinion-score distribution.

Table 1. Six-theme profile of the benchmark after conservative filtering and duplicate removal. MOS mean and standard deviation are reported on the source datasets' original 0–5 rating scale; before training and evaluation, every per-image MOS value is normalised to [0, 1] through a fold-wise min–max linear map fitted on the training MOS values only (the held-out theme's MOS values play no role in fitting the calibration scale). All losses, metrics and significance tests in this manuscript operate on the normalised scale; the original-scale values are retained in the released split files for full traceability.

Theme	Samples	MOS mean	MOS std	Theme cue
Editorial	1,180	3.74	0.48	Magazine and article covers; text–image hierarchy
E-commerce	1,325	3.51	0.55	Product hero layout; purchase intent
Education	1,240	3.67	0.51	Course covers; instructional layout
Finance	1,095	3.39	0.59	Numerals and charts; formal dense information
Healthcare	1,160	3.58	0.50	Clinical and wellness visuals; readability
Travel	985	3.46	0.57	Scenic destinations; travel promotion

The candidate pool was built by querying each source dataset for visually self-contained design surfaces and applying the theme-mapping rules in turn. Each source contributes a different number of retained samples per theme: AVA[3] and the photo-critique corpus of Vera Nieto et al.[22] supply the largest share of Editorial and Travel images; TAD66K[17] supplies most E-commerce and Healthcare samples; the layout-aware corpora[10,11] supply most Education images; and the website-design corpus[12] supplies the bulk of Finance samples. The per-source × per-theme contribution table, together with the source licence summary, is provided in Supplementary Table S1 so that an independent reviewer can audit the source composition without consulting the raw datasets.

The pHash Hamming threshold of eight bits was selected by sweeping the threshold over {4, 6, 8, 10, 12} and reporting the number of pairs removed and the residual mean cosine similarity of CLIP image embeddings (Supplementary Information); eight bits is the smallest threshold at which the residual mean cosine similarity dropped below 0.92, the value commonly used to flag campaign-family duplicates in vision–language work. An additional CLIP-embedding deduplication step is provided as an optional preprocessing script for users who require a stricter de-duplication regime.

Problem formulation

Let x_i denote a design image, t_i its theme label, $g(t_i)$ a short natural-language theme descriptor drawn from the paraphrase pool of Supplementary Methods, and $y_i \in [0, 1]$ its calibrated MOS. A

frozen text encoder *E_t* (CLIP ViT-L/14 text tower at 768 dimensions, frozen throughout training) produces a semantic theme vector

$$q_i = P \cdot E_t(g(t_i)) \in \mathbb{R}^{d_q}, \qquad (1)$$

with *d_q* = 64 and *P* a learned projection. The descriptor *q_i* modulates intermediate visual features (see *TAR-Net architecture*). Because *q_i* is a function of a *description* rather than a one-hot index, a held-out theme admits a valid conditioning vector at test time even when no images, ratings or descriptors from that theme appeared during training.

TAR-Net outputs a scalar score

$$s_i = f_\theta(x_i, q_i), \quad (2)$$

and is trained with the composite objective

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{score}} + \lambda \mathcal{L}_{\text{rank}} + \beta \mathcal{L}_{\text{ct}}. \qquad (3)$$

The pointwise score loss is a Huber loss with threshold δ = 0.05 that limits sensitivity to MOS outliers:

$$\mathcal{L}_{\text{score}} = H_\delta(s_i - y_i), \quad (4)$$

$$H_\delta(z) = \frac{1}{2}z^2 \text{if} |z| \le \delta, \delta\left(|z| - \frac{1}{2}\delta\right) \text{otherwise.} \quad (5)$$

For the margin-ranking loss, sample pairs (*i*, *j*) are drawn within each mini-batch with $|y_i - y_j| > \tau$ (τ = 0.10 on the calibrated scale) to exclude ambiguous near-ties; orientation is δ_{ij} = sign(*y_i* − *y_j*):

$$\mathcal{L}_{\text{rank}} = \max\big(0, m - \delta_{ij}(s_i - s_j)\big), m = 0.10. \quad (6)$$

Negative pairs are over-sampled by orientation until a 1 : 1 positive : negative balance is reached per batch.

Let *h_inv*(*x*, *q*) denote the theme-invariant descriptor produced by ADM under conditioning *q*. The two paraphrases of the same theme are encoded through the same frozen text encoder and the same learned projection of equation (1): *q_i* = *P* · *E_t*(*g*(*t_i*)) and *q_i*′ = *P* · *E_t*(*g*′(*t_i*)), where *g*′(*t*) is drawn uniformly from the three-element paraphrase pool of Supplementary Methods. The consistency term

$$\mathcal{L}_{\text{ct}} = 1 - \cos(h_{\text{inv}}(x_i, q_i), h_{\text{inv}}(x_i, q_i')). \qquad (7)$$

discourages the theme-invariant branch from collapsing onto theme-specific features; the theme-*aware* branch is intentionally left free of this constraint so that it can specialise. We set λ = 0.40 and β = 0.05 after the sensitivity sweep reported below.

TAR-Net architecture

TAR-Net has four components: a visual backbone, the Theme-Aware Modulation operator (TAM), the Aesthetic Decomposition Module (ADM) and a two-branch ranking head. Figure 1 illustrates the overall flow.

The visual backbone is a ViT-B/16 pretrained on ImageNet-21k at 384 × 384 input resolution, with 12 transformer blocks and 768-dimensional patch tokens; layer-wise learning-rate decay with factor 0.75 is applied during fine-tuning and all backbone parameters are trainable.

TAM is a text-conditioned low-rank FiLM operator applied at three intermediate blocks, indexed $\ell \in \{4, 7, 10\}$ (0-indexed, post LayerNorm). Given the projected theme vector $q_i \in \mathbb{R}^{d_q}$ with $d_q = 64$ and a patch-token tensor $H^{(\ell)} \in \mathbb{R}^{N \times d}$ with $d = 768$,

$$\Gamma^{(\ell)} = \mathrm{MLP}_{\gamma}^{(\ell)}(q_i) \in \mathbb{R}^{r \times d}, \qquad (8)$$

$$\boldsymbol{\beta}^{(\ell)} = \mathrm{MLP}_{\beta}^{(\ell)}(q_i) \in \mathbb{R}^{d}, \qquad (9)$$

$$H'^{(\ell)} = \left(\mathbf{1}_{N \times d} + \sigma\left(H^{(\ell)} W_r^{\top}\right)\Gamma^{(\ell)}\right) \odot H^{(\ell)} + \mathbf{1}_N \boldsymbol{\beta}^{(\ell)\top}, \quad (10)$$

with rank $r = 16$, σ the GELU activation and $W_r \in \mathbb{R}^{r \times d}$ a shared down-projection. Equation (10) reduces to standard FiLM when $r = d$; the low-rank factorisation keeps TAM's parameter count to 0.13 M, two orders of magnitude smaller than the backbone.

ADM operates on the post-TAM token tensor of the final transformer block. Three parallel branches encode complementary aesthetic evidence and are fused by a gating MLP. We partition the ViT-B/16 backbone into four equal stages (blocks 1–3 = stage 1, 4–6 = stage 2, 7–9 = stage 3, 10–12 = stage 4; equivalently 0-indexed blocks 0–2, 3–5, 6–8, 9–11) and read out *stage-3 tokens* at the output of block 9 (0-indexed: block 8) and *stage-4 tokens* at the output of block 12 (0-indexed: block 11). The *harmony branch* takes the post-TAM stage-3 tokens, applies a 1 × 1 convolution (`token_conv`: 768 → 256) along the channel dimension of the 576 patch tokens (reshapable into a 24 × 24 spatial grid at 384 × 384 input resolution and patch size 16) and averages spatially, producing a 256-dimensional descriptor. In parallel, a 128-bin RGB–LAB joint histogram of the input image (64 RGB bins from an 8-level uniform per-channel quantisation, plus 64 LAB bins from an 8 × 4 × 2 quantisation on L*, a*, b*, L1-normalised to sum to one) is projected through a single linear layer (`hist_proj`: 128 → 256). The two 256-dimensional vectors are concatenated and passed through a fuse MLP (`fuse`: LayerNorm + Linear(512 → 256) + GELU), giving* h_*harm* ∈ ℝ^{d_b*} *with* d_b *= 256. The* saliency branch *takes the stage-4 tokens, projects them through `proj` (768 → 256) after spatial mean pooling, and fuses the resulting 256-dimensional vector with a saliency-prior descriptor distilled offline from a frozen BASNet teacher[33] pretrained on DUTS. The BASNet teacher outputs a 320 × 320 saliency map per input image; the map is bilinearly resampled to the 24 × 24 patch grid and globally pooled along its spatial dimension into a 64-dimensional descriptor (one channel per 3 × 3 pooling region after a 24 → 8 → 1 reduction), projected through `prior_proj` (64 → 256) and fused with the token descriptor through the same LayerNorm + Linear + GELU pattern as the harmony branch (only the projections are trained; the BASNet teacher contributes no trainable parameters). The* typography branch *encodes OCR text-box masks generated offline by a frozen PaddleOCR engine: detected text boxes are rasterised at the original resolution, downsampled to a 24 × 24 binary mask and globally summarised into a 64-dimensional coverage descriptor (8 horizontal × 4 vertical × 2 reading-direction bins) with a* no-text *indicator dimension folded in as the 64-th coordinate so that the typography stream remains well-defined on text-free designs. The descriptor is the query in a single-head cross-attention layer (`q_proj`: 64 → 256; `kv_proj`: 768 → 256; attention output normalised through `out_norm`) over the stage-4 tokens, producing* h_*typ* ∈ ℝ^{d_b*}. A gating vector $g = \mathrm{softmax}(W_g[h_{\mathrm{harm}}; h_{\mathrm{sal}}; h_{\mathrm{typ}}])$ produces the fused decomposed descriptor

$$h_{\mathrm{dec}} = g_1 h_{\mathrm{harm}} + g_2 h_{\mathrm{sal}} + g_3 h_{\mathrm{typ}}. \qquad (11)$$

ADM is trained end-to-end with TAR-Net; the saliency teacher and OCR engine are frozen and contribute no trainable parameters.

Two pooled descriptors are read out from the post-TAM token tensor $H'^{(L)}$ at the final transformer block. Let $p = \mathrm{Pool}(H'^{(L)})$ denote the spatial mean of the token tensor; then $h_{\mathrm{aware}} = \mathrm{MLP}_{\mathrm{aware}}([p; q]) \in \mathbb{R}^d$ uses both *p* and *q*, while $h_{\mathrm{inv}} = \mathrm{MLP}_{\mathrm{inv}}(p) \in \mathbb{R}^d$ uses *p* only. The two descriptors are concatenated with *h_dec* of equation (11):

$$z = [h_{\mathrm{aware}}; h_{\mathrm{inv}}; h_{\mathrm{dec}}] \in \mathbb{R}^{2d+d_b}. \qquad (12)$$

A two-layer MLP head with GELU activation produces the scalar score *s_i* = MLP(*z*). The consistency loss *L_ct* of equation (7) is computed between *h_inv*(*x_i*, *q_i*) and *h_inv*(*x_i*, *q_i'*): although MLP_inv does not receive *q* directly, *p* depends on *q* through the TAM modulation at blocks {4, 7, 10}. The loss therefore constrains the backbone to produce token responses whose theme-invariant projection is stable across paraphrases of the same theme, while *h_aware* retains the freedom to encode theme-specific evidence through its explicit *q* input.

Table 2 summarises the architecture and training schedule. AdamW is used with initial learning rate 1×10^{-4}, weight decay 5×10^{-4}, batch size 16 and a cosine schedule over 80 epochs with early stopping on validation Spearman correlation. All trainable baselines use the same optimiser, schedule and seed list; the validation set for the cross-theme protocol is sampled from the source themes only (never from the held-out theme), to prevent leakage.

Table 2. Architecture and implementation settings.

Item	**Value**
Backbone	ViT-B/16, ImageNet-21k pretrained
Input resolution	384 × 384
Patch token dimension *d*	768
Theme descriptor dim *d_q*	64
TAM rank *r*	16
TAM insertion blocks (0-indexed)	{4, 7, 10}, post LayerNorm
Frozen text encoder *E_t*	CLIP ViT-L/14 text tower, 768-d output
Theme descriptor projection	learned, 768 → 64
ADM branch dim *d_b*	256
Harmony branch prior	1 × 1 conv on stage-3 tokens + 128-bin RGB–LAB histogram
Saliency branch prior	BASNet[33] (pretrained on DUTS, frozen)
Typography branch prior	PaddleOCR (frozen) text-box masks
Paraphrase pool size	3 paraphrases per theme (Supplementary Methods)
MOS calibration	fold-wise min–max linear map to [0, 1], fitted on training MOS only
Optimiser	AdamW
Initial learning rate	1×10^{-4}
Layer-wise LR decay	0.75
Weight decay	5×10^{-4}
Batch size	16

Training epochs	80 with early stopping on validation Spearman correlation
Ranking loss weight λ	0.40
Consistency loss weight β	0.05
Ranking margin *m*	0.10
Pair gap threshold τ	0.10
Independent runs	5 seeds {0, 1, 2, 3, 4}
h_aware and *h_inv* dim	256 each (down-projected from 768 by their respective MLPs)
Total parameters	93.4 M (86.6 M backbone + 0.13 M TAM + 5.9 M ADM + 0.8 M heads; CLIP text tower frozen, excluded from trainable count)

Evaluation protocols and baselines

The *random split* draws training, validation and test partitions from all six themes with theme-balanced sampling at a 70 / 10 / 20 ratio. The *theme-disjoint* protocol withholds one theme entirely at test time and rotates the held-out theme across the six themes (six-fold rotation). For each protocol, every configuration is repeated over five seeds. Ranking quality is reported as the Spearman rank-order correlation coefficient (SRCC) and the Pearson linear correlation coefficient (PLCC); score error is reported as the mean absolute error (MAE) and the root-mean-square error (RMSE). All metrics are reported as mean ± standard deviation over the five seeds, supplemented by normal-theory 95% confidence intervals computed as `mean ± t(0.025, 4) · std / √5` and by Cohen's d_z effect sizes for paired comparisons. Normal-theory intervals are conservative for $n = 5$ but are mathematically determined by the reported mean and standard deviation; bootstrap intervals computed from the per-seed logs released in the supplementary archive agree with the normal-theory intervals to within rounding.

Four families of baselines are compared with fixed preprocessing and fixed prompt and score-parsing rules where applicable. Trainable baselines — NIMA[7], a ResNet-50 regressor, MUSIQ[8], a ViT-B/16 regressor, the theme-label-concatenation head, CLIP prompt conditioning[13,14], TANet[17] and TAVAR[18] — are trained on the same split files with the same seed list. Checkpoint-only baselines such as LIQE[16] use the public weights with the preprocessing prescribed in the original paper. Multimodal-LLM baselines combine UNIAA[21], attribute-driven hierarchical prompts[23], PromptIQA[24], next-token IQA[25] and Gamma MoE[26]; all five use the prompt templates released with this manuscript and deterministic score parsing (temperature 0, top-p 1, max-new-tokens 32, parser `\d+(\.\d+)?`), with reasoning chains disabled to avoid stochastic output. The MLLM ensemble row in Tables 3 and 4 reports the mean and standard deviation across the five MLLM scorers; a per-model decomposition is provided in the supplementary archive. The "TANet / TAVAR (best)" row reports the better of two independent reproductions trained on the same split files and seed list, with separate per-model logs in the supplementary archive.

The methodological positioning of TAR-Net against representative methods is summarised in Table 3. Tables 4 and 5 report numerical results.

To enforce baseline fairness, every trainable method (TAR-Net, NIMA, ResNet-50, MUSIQ, ViT-B/16 regressor, theme-label concatenation, CLIP prompt conditioning, the TANet-style and TAVAR-style

reproductions described below, and the CLIP-ViT regressor) uses the same split files, the same five seeds, the same input resolution (384 × 384), the same data augmentation pipeline (resize-and-centre-crop at evaluation; resize, random crop, horizontal flip and colour-jitter at training), the same batch size (16), the same optimiser family (AdamW with layer-wise learning-rate decay 0.75), the same cosine schedule over 80 epochs with early stopping on validation Spearman correlation (patience = 10 epochs, restore best checkpoint), the same compute budget (single NVIDIA RTX 3090 for at most 8 wall-clock hours per (seed, configuration) pair, equivalent to ≈ 15% additional training time over TAR-Net's reference budget for the architecturally larger baselines), and the same per-method hyperparameter sweep over learning rate $\in \{3 \times 10^{-5}, 1 \times 10^{-4}, 3 \times 10^{-4}\}$ with the best validation Spearman correlation retained per (method, seed). Checkpoint-only baselines (LIQE) use the public weights with the preprocessing prescribed in the original paper.

Baseline reproductions. The TANet[17] and TAVAR[18] rows of Tables 4 and 5 are *simplified in-house reproductions* and not the full official implementations: our TANet variant replaces the original theme-oriented feature extractor with a per-theme MLP branch on top of the shared ViT-B/16 backbone; our TAVAR variant replaces the attribute-theme graph with a per-theme attribute MLP combined with a global aesthetic head. Both variants preserve the structural choice of *categorical theme branching* that distinguishes the two methods from prompt- or label-based conditioning, but are simpler than the original authors' implementations. We label TAVAR as the *strongest reproduced baseline in our codebase* rather than as the strongest baseline in the literature; an integration of the upstream official codebases is left as future work. The "TAVAR (better-performing reproduction)" choice in Table 6 is *a posteriori*; we treat the comparison against TAVAR as the primary hypothesis, the comparison against TANet as a secondary hypothesis, and apply Holm–Bonferroni correction over the resulting two-test family (in addition to the random / cross-theme two-protocol family that is already corrected for in Table 6) to control the family-wise error rate at $\alpha = 0.05$.

Multimodal-LLM scorer adapters. The five MLLM scorers in the *MLLM scorer (ensemble)* row consume our released prompt template under deterministic decoding (temperature 0, top-*p* 1, max-new-tokens 32, parser `\d+(\.\d+)?`); the prompt template, decoder settings and a mock adapter used for the public smoke test are released in the public repository (`scripts/mllm_baselines.py`). Real-scorer adapters require the upstream UNIAA, ADHP, PromptIQA, Next-Token IQA and Gamma MoE model weights and are not bundled with our repository; users who wish to re-run the MLLM ensemble must integrate the upstream weights according to the original authors' instructions and call our prompt template through the released `--model` interface.

Ranking-pair sampling proceeds as follows for every mini-batch: within each batch of 16 images the loss layer enumerates all distinct unordered pairs whose calibrated MOS gap exceeds the threshold $\tau = 0.10$ (excluding near-ties), assigns orientation $\delta_{ij} = \mathrm{sign}(y_i - y_j)$, and resamples the orientation-balanced set without replacement until a 1 : 1 positive-to-negative orientation balance is reached for that batch; if the batch has fewer than 4 valid pairs the ranking loss is skipped for that batch and the score loss alone drives the update. Determinism is enforced through `torch.use_deterministic_algorithms(True)`, an environment-level CUDA workspace size cap (`CUBLAS_WORKSPACE_CONFIG=:4096:8`), seed control for Python, NumPy, PyTorch and the CUDA RNG, a deterministic DataLoader worker initialiser and a fixed `cudnn.benchmark = False`. The five seeds {0, 1, 2, 3, 4} are propagated to all of the above.

This research did not collect new human-subjects data. The three annotators and the senior adjudicator who produced the theme labels are members of the research team labelling publicly available images; no human-subjects approval is therefore required under our institution's research-ethics policy. The source datasets are released under research-permissive licences: AVA[3] under a research-use licence permitting derived metadata; TAD66K[17] under a CC-BY licence; the photo-critique corpus of Vera Nieto et al.[22] under CC-BY-NC for non-commercial research; the website-design corpus[12] under a research-only licence that prohibits image redistribution; and the layout-aware corpora[10,11] under research-only licences that permit metadata redistribution. Source images are not redistributed when the source-dataset licence prohibits redistribution; the failure-case viewer only loads images from the subset of sources whose licences permit image redistribution. Authorship of this work satisfies the criteria of the journal — all named authors contributed to conception, methodology, software, analysis, writing, supervision or data curation as itemised in the *Author contributions* section — and no large language model is listed as an author or contributor on this work.

Table 3. Methodological differentiation of TAR-Net from seven representative methods. *Where* indicates the conditioning insertion point; *Cue* the theme representation; *Rank* the ranking objective; *Reg.* the cross-theme regulariser; *ADM* the decomposed aesthetic-evidence stream; *Protocol* the principal evaluation regime. "–" indicates the row method does not use the corresponding mechanism.

Method	**Where**	**Cue**	**Rank**	**Reg.**	**ADM**	**Protocol**
NIMA[7]	–	–	–	–	distribution head	random
MUSIQ[8]	–	–	–	–	multi-scale tokens	random
Theme-label concat.	score head	one-hot	–	–	–	random + xt
CLIP prompt cond.	input	text prompt	–	–	–	random + xt
TANet[17]	branches	categorical	–	–	per-theme	in-distrib.
TAVAR[18]	graph	categorical	pair (opt.)	–	attr. graphs	in-distrib.
AesthetiQ[27]	align.	layout pref.	preference	–	–	layout
TAR-Net	**mid ViT (r = 16)**	**text + paraphrases**	**margin pair**	**L_ct**	**harm. / sal. / typ.**	**random + xt + *d_z* + HB**

Results

The numerical results below are produced by the documented split files, the seed list released with the manuscript, the training and evaluation scripts and the statistical-test notebooks in the supplementary archive.

Main comparison

Tables 4 and 5 report random-split and theme-disjoint metrics. Every entry is mean ± standard deviation over five seeds, augmented by a normal-theory 95% confidence interval computed from `t(0.025, 4) · std / √5`. The supplementary archive provides per-seed bootstrap intervals that agree with the normal-theory intervals to three decimals.

Table 4. Random-split comparison. Each "mean ± std" entry for a trainable method is computed across the same five training seeds {0, 1, 2, 3, 4}; the MLLM scorer row is the mean and standard deviation across the five constituent multimodal scorers listed in the *Evaluation protocols and baselines* paragraph (between-model variance, *not* between-seed variance — these two sources of variability are not directly comparable). Normal-theory 95% confidence intervals are computed from the seed-level variance as `t(0.025, 4) · std / √5`.

Method	SRCC (mean ± std)	SRCC 95% CI	PLCC (mean ± std)	MAE	RMSE
NIMA[7]	0.624 ± 0.013	[0.608, 0.640]	0.631 ± 0.011	0.246 ± 0.010	0.318 ± 0.011
ResNet-50	0.641 ± 0.012	[0.626, 0.656]	0.649 ± 0.010	0.232 ± 0.009	0.305 ± 0.010
MUSIQ[8]	0.683 ± 0.010	[0.671, 0.695]	0.690 ± 0.009	0.213 ± 0.008	0.281 ± 0.009
ViT-B/16 regressor	0.692 ± 0.009	[0.681, 0.703]	0.699 ± 0.008	0.208 ± 0.008	0.274 ± 0.008
LIQE[16]	0.697 ± 0.010	[0.685, 0.709]	0.703 ± 0.009	0.205 ± 0.008	0.270 ± 0.009
Theme-label concat.	0.714 ± 0.008	[0.704, 0.724]	0.720 ± 0.008	0.193 ± 0.007	0.256 ± 0.008
CLIP prompt cond.	0.719 ± 0.008	[0.709, 0.729]	0.724 ± 0.007	0.191 ± 0.007	0.253 ± 0.008
CLIP-ViT regressor	0.709 ± 0.009	[0.698, 0.720]	0.715 ± 0.007	0.196 ± 0.007	0.260 ± 0.008
MLLM scorer (ensemble)	0.688 ± 0.012	[0.673, 0.703]	0.695 ± 0.011	0.217 ± 0.010	0.286 ± 0.011
TANet[17]	0.720 ± 0.009	[0.709, 0.731]	0.726 ± 0.008	0.190 ± 0.008	0.253 ± 0.009
TAVAR[18]	0.724 ± 0.008	[0.714, 0.734]	0.730 ± 0.007	0.188 ± 0.007	0.250 ± 0.008
TAR-Net (ours)	**0.748 ± 0.006**	**[0.741, 0.755]**	**0.753 ± 0.006**	**0.181 ± 0.006**	**0.244 ± 0.007**

Table 5. Theme-disjoint comparison: each row is averaged over six held-out themes and five seeds. The "mean ± std" convention follows Table 4 — for trainable methods the standard deviation is between five seeds (computed on the seed-first-then-theme aggregation), and for the MLLM scorer row it is between the five constituent multimodal scorers; the two sources of variability are not directly comparable.

Method	SRCC (mean ± std)	SRCC 95% CI	PLCC (mean ± std)	MAE	RMSE
NIMA	0.604 ± 0.015	[0.585, 0.623]	0.611 ± 0.013	0.261 ± 0.012	0.332 ± 0.013
ResNet-50	0.623 ± 0.014	[0.606, 0.640]	0.630 ± 0.012	0.246 ± 0.011	0.317 ± 0.012
MUSIQ	0.671 ± 0.012	[0.656, 0.686]	0.677 ± 0.010	0.227 ± 0.009	0.294 ± 0.010
ViT-B/16	0.682 ± 0.011	[0.668, 0.696]	0.688 ± 0.009	0.220 ± 0.009	0.287 ± 0.010
LIQE	0.684 ± 0.011	[0.670, 0.698]	0.690 ± 0.010	0.216 ± 0.009	0.282 ± 0.010
Theme-label concat.	0.706 ± 0.009	[0.695, 0.717]	0.712 ± 0.008	0.203 ± 0.008	0.267 ± 0.009
CLIP prompt cond.	0.713 ± 0.009	[0.702, 0.724]	0.719 ± 0.008	0.199 ± 0.008	0.262 ± 0.008
CLIP-ViT	0.701 ± 0.009	[0.690, 0.712]	0.707 ± 0.008	0.207 ± 0.008	0.271 ± 0.009
MLLM scorer (ensemble)	0.676 ± 0.013	[0.660, 0.692]	0.682 ± 0.012	0.229 ± 0.011	0.296 ± 0.012
TANet[17]	0.713 ± 0.010	[0.700, 0.726]	0.719 ± 0.008	0.200 ± 0.008	0.263 ± 0.009

TAVAR[18]	0.719 ± 0.008	[0.709, 0.729]	0.725 ± 0.007	0.196 ± 0.007	0.259 ± 0.008
TAR-Net (ours)	**0.743 ± 0.007**	**[0.734, 0.752]**	**0.748 ± 0.006**	**0.186 ± 0.006**	**0.249 ± 0.007**

The random-to-cross-theme Spearman drops are 0.020 (NIMA), 0.018 (ResNet-50), 0.012 (MUSIQ), 0.010 (ViT-B/16), 0.013 (LIQE), 0.008 (Theme-label), 0.006 (CLIP prompt), 0.008 (CLIP-ViT), 0.012 (MLLM ensemble), 0.007 (TANet), 0.005 (TAVAR) and 0.005 (TAR-Net). The decreasing trend correlates with the depth and explicitness of theme conditioning, with theme-aware backbones and TAR-Net showing the smallest drops. The persistent advantage of TAR-Net under theme shift (+0.024 Spearman over TAVAR, the strongest non-TAR-Net baseline) is the central empirical claim of this paper and is tested formally in Table 6. Supplementary Figs. S3 and S4 visualise the random-split and theme-disjoint comparisons, and Supplementary Table S6 reports the per-model breakdown of the multimodal-LLM scorer ensemble (UNIAA, attribute-driven hierarchical prompts, PromptIQA, next-token IQA and Gamma MoE).

Paired *t*-tests over the five seeds against the strongest non-TAR-Net baseline (TAVAR, the better-performing theme-aware reproduction) yield the values in Table 6. The same five seeds are used for both methods; $n = 5$, df = 4. All tests are two-tailed with a nominal significance level of $\alpha = 0.05$. The *p_adj* column reports Holm–Bonferroni-corrected *p* values across the two protocols (random / cross-theme) and over the two-baseline (TANet, TAVAR) family[35]; effect size is Cohen's $d_z = t / \sqrt{n}$[34]. The seed-level sample size is small ($n = 5$) by the standards of statistical-significance research, and we therefore frame the result as a *substantial and consistent improvement* rather than as a single-test significance claim: the parametric paired *t*-test exceeds the $\alpha = 0.05$ threshold after Holm correction, but the exact paired sign-permutation test cannot reject the null below $p = 0.0625$ at $n = 5$ (the smallest two-sided *p* attainable with five paired observations). The stratified seed × theme permutation (30 paired observations on the cross-theme protocol) attains $p = 0.0026$, the stratified bootstrap 95% CI excludes zero on both protocols, and the linear mixed-effects model with held-out theme as a random intercept estimates $\beta = +0.024$ SRCC (SE = 0.005, $p = 0.0048$, 95% CI [0.014, 0.034]). The four procedures agree in direction and magnitude. We additionally note that the seed × theme observations are not strictly independent (seed and theme effects co-vary within the protocol); the mixed-effects model in Table 6 explicitly accounts for the seed-by-theme nesting through the random intercept on `held_out_theme`, and an extended model that also treats *seed* as a random effect (`SRCC ~ method + (1 | held_out_theme) + (1 | seed)`) returns a fixed-effect estimate that agrees with the simpler model to three decimals. The per-seed × per-theme metric matrix, the fitted model summaries and the resampling logs are released so that an independent reviewer can recompute every entry of Table 6 from scratch. Increasing the seed budget from five to ≥ 20 to support a wider battery of significance tests is listed as future work in the *Discussion*.

Table 6. Statistical comparison of TAR-Net against the strongest non-TAR-Net baseline (TAVAR). All four procedures are applied to the same five-seed (random-split) and five-seed × six-theme (cross-theme) metric matrix; mean gain in Spearman correlation is Δ SRCC = +0.024 throughout. The Notes column indicates the within-subject paired design (paired *t* / *d_z*), the lower bound on the *p* value attainable with $n = 5$ paired observations (exact sign permutation), the stratified resampling design (permutation, bootstrap) and the random-effects structure (mixed-effects model). Where a row entry is "–", the procedure does not apply because the random-split protocol has no held-out theme to stratify over.

Test	Random split	Cross-theme	Notes

Paired *t*-test (raw / Holm-adjusted *p*)	$t = 4.84$, $p = 0.0084$ / 0.0104	$t = 5.54$, $p = 0.0052$ / 0.0104	df = 4 (within-subject paired)
Cohen's d_z	2.17	2.48	large effect (Cohen's convention $d_z > 0.8$)
Exact paired sign-permutation, $2^5 = 32$ perms	$p = 0.0625$	$p = 0.0625$	lower bound attainable at $n = 5$
Stratified permutation (seed × theme, B = 10,000)	–	$p = 0.0026$	5 seeds × 6 held-out themes = 30 paired obs
Stratified bootstrap 95% CI on Δ SRCC (B = 10,000)	[0.013, 0.035]	[0.014, 0.034]	seeds resampled within each held-out theme
Linear mixed-effects (theme random intercept)	–	β = +0.024 (SE 0.005), $p = 0.0048$, 95% CI [0.014, 0.034]	Satterthwaite df ≈ 5.0

Ablation and sensitivity

Table 7 reports the controlled ablation. The *Base model* is a ViT-B/16 backbone with a two-layer score head trained under the TAR-Net optimiser schedule and seed list; it differs from the ViT-B/16 regressor row of Table 4 in the use of layer-wise learning-rate decay, AdamW (rather than SGD) and the early-stopping rule on validation Spearman correlation. The two rows are intentionally distinct configurations and are not in conflict. Component contributions in Δ Spearman against the Base model are +0.011 / +0.016 (Theme-label), +0.016 / +0.029 (CLIP prompt), +0.017 / +0.030 (TAM), +0.028 / +0.046 (ADM) and +0.009 / +0.015 (L_ct, computed as Full − no-L_ct). Cross-theme gains are consistently larger than random gains for TAM and L_ct, indicating that these components do theme-shift-specific work.

Table 7. Ablation of TAR-Net components (mean ± standard deviation Spearman correlation over five seeds).

Variant	**Random**	**Cross-theme**	**Notes**
Base model	0.703 ± 0.009	0.670 ± 0.010	ViT-B/16 + score head, TAR-Net schedule
+ Theme-label concatenation	0.714 ± 0.008	0.686 ± 0.010	shallow conditioning at score head
+ CLIP prompt conditioning	0.719 ± 0.008	0.699 ± 0.009	input-prompt conditioning
+ TAM	0.720 ± 0.008	0.700 ± 0.009	feature-level modulation
+ ADM	0.731 ± 0.007	0.716 ± 0.008	decomposed aesthetic evidence
Full TAR-Net without L_ct	0.739 ± 0.007	0.728 ± 0.008	no paraphrase consistency
Full TAR-Net	**0.748 ± 0.007**	**0.743 ± 0.007**	TAM + ADM + ranking + L_ct

Supplementary Table S2 sweeps the principal hyperparameters λ, β, the TAM rank, the theme-descriptor dimension and the ranking margin *m*; the selected setting (λ = 0.40, β = 0.05, TAM rank = 16,

theme dim = 64, m = 0.10) is not materially changed by that sweep. Supplementary Fig. S5 visualises the cross-theme stability of the ablation across all components of Table 7.

Theme-wise hold-out analysis

Supplementary Table S3 lists the theme-disjoint performance per held-out theme; SRCC and PLCC are means over five seeds. The aggregate row uses the seed-first-then-theme aggregation order and matches Table 5 to three decimals (0.743 SRCC, 0.748 PLCC, 0.249 RMSE); the small per-theme difference (0.249 versus the per-theme RMSE mean of 0.2512) is the round-off difference between the seed-first-then-theme and theme-first-then-seed aggregation orders, and is documented in the supplementary aggregation-audit notebook. Supplementary Fig. S6 visualises the per-theme Spearman correlation; the largest cross-theme gain is on Editorial (0.750) and the most challenging hold-out is Finance (0.736), which is consistent with the typography-density tension flagged in the failure analysis below.

Because the source datasets contribute unevenly to each theme (AVA and Vera Nieto are over-represented in Editorial and Travel, TAD66K in E-commerce and Healthcare, the website-design corpus in Finance), a naive theme-disjoint protocol risks conflating *theme generalisation* with *source generalisation*. Three controls disentangle the two. First, a leave-one-source-out (LOSO) protocol re-runs the six-fold rotation while additionally holding out one source dataset per fold; on the released splits, TAR-Net's LOSO cross-theme SRCC is 0.738 ± 0.009 (Supplementary Table S8), within one standard deviation of the per-theme cross-theme aggregate of 0.743, and the +0.022 SRCC margin over TAVAR is preserved. Second, a within-theme source-balanced re-sampling that down-samples the dominant source in each theme to the second-most-frequent source yields a cross-theme SRCC of 0.741 ± 0.008 for TAR-Net and 0.717 ± 0.008 for TAVAR — the margin is again preserved at +0.024 SRCC. Third, an extended linear mixed-effects model fit to the per-seed × per-theme metric matrix that adds the dominant source dataset as a second random intercept (`SRCC ~ method + (1 | held_out_theme) + (1 | dominant_source)`) estimates the same fixed-effect gain β = +0.023 SRCC (SE = 0.006, p = 0.0089, 95% CI [0.012, 0.034]) as the theme-only mixed-effects model in Table 6; the source-level variance component explains only 12% of the residual variance, with the dominant variance source remaining the held-out theme (61%) followed by the seed (27%). The three controls together support the claim that the +0.024 SRCC margin reflects genuine theme conditioning rather than adaptation to specific source-dataset distributions; the LOSO splits, the source-balanced re-sampling and the extended mixed-effects fit are released in `scripts/loso_eval.py`, `scripts/source_balanced.py` and `scripts/mixed_effects.py` of the public repository so that an independent reviewer can recompute every number.

Mean-opinion-score calibration is performed fold-wise rather than globally. Within each protocol fold, the min–max linear map to [0, 1] is fitted on the training MOS values only; the held-out theme's MOS values play no role in fitting the calibration scale. This removes the test-MOS leakage flagged during review. We additionally re-ran the main random-split and cross-theme experiments under a per-source z-score calibration variant whose scale is fitted on training-set MOS values per source dataset (Supplementary Table S8); both calibration variants yield identical SRCC rankings and a cross-theme SRCC gap that agrees with Table 5 to within ±0.003.

Failure categories and deployment

Four recurring failure modes are present in the per-seed error analysis and are tabulated in Supplementary Table S4: (i) finance-theme layouts with dense numerals or small typography that are penalised as clutter, (ii) travel and e-commerce layouts dominated by a strong hero image that masks weak typography, (iii) editorial and education layouts whose theme is ambiguous to the descriptor, and (iv) healthcare layouts with stylised fonts that escape the OCR-conditioned typography branch. A qualitative-example viewer for these categories is released with the public repository; it composes a TAR-Net prediction, the calibrated MOS and the strongest baseline prediction for each candidate. Because image redistribution is constrained by source-dataset licences, the viewer loads images from the identifier list rather than redistributing pixels. Supplementary Table S5 reports the deployment budget on a single NVIDIA RTX 3090 with batch 16, fp32, after a 50-iteration warm-up and excluding I/O: average latency 18.7 ms per image (batched amortised), throughput 53.5 images per second, peak GPU memory 5.8 GB, single-thread CPU-only latency 312 ms per image; total parameters 93.4 M (86.6 backbone + 0.13 TAM + 5.9 ADM + 0.8 head, plus a frozen 124 M CLIP text tower that does not count towards the trainable budget). Same-hardware budgets for the baselines listed in Table 3 are provided in the supplementary archive.

Discussion

The empirical pattern across Tables 4, 5 and 7 separates three forms of conditioning. Theme-label concatenation gives small but consistent gains over the unconditional base, confirming that some theme cue is useful (+0.011 random, +0.016 cross-theme). Prompt-level CLIP conditioning produces a larger cross-theme gain than random gain (+0.029 versus +0.016), so natural-language theme cues carry information unavailable in the categorical index. Intermediate-feature modulation (TAM) matches prompt conditioning on the random split (+0.017 versus +0.016) but moves further on cross-theme (+0.030), consistent with the intuition that modulation of feature responses can amplify or suppress theme-relevant channels rather than only re-weighting the score head. ADM contributes the largest single component in both protocols (+0.028 / +0.046), so decomposed aesthetic evidence is complementary to theme conditioning. The L_ct term adds another +0.015 cross-theme Spearman correlation, the largest gain per unit of added parameter since L_ct adds no inference-time computation.

A reading of why intermediate-feature modulation outperforms shallow conditioning follows from conditional mutual information. A score head appended to a frozen representation $h(x)$ models $p(s \mid h(x), t)$ and must carry $I(s ; t \mid h(x))$ entirely in the head. Feature-level modulation lets the backbone re-route signal from x according to t, raising $I(h_aware(x, q) ; s)$ while L_ct constrains $h_inv(x, q)$ to remain close to theme-invariant. We do not claim a formal bound; the argument describes why the observed pattern is expected rather than coincidental. In an editorial review pipeline that ranks tens of candidate designs and forwards only the top three or four to a human reviewer, a 0.024 Spearman margin translates into a measurable reduction in the pairwise rank-flip rate at the top of the queue, where the cost of an error is highest. Concretely, recomputing Top-K Kendall τ and the per-query pairwise rank-flip rate against TAVAR on the released cross-theme split (Supplementary Table S7) gives Top-3 Kendall τ = 0.812 for TAR-Net versus 0.756 for TAVAR and a per-query Top-5 rank-flip rate of 9.6% for TAR-Net versus 14.2% for TAVAR — a 32% relative reduction in the queue-head error rate. The `topk_ranking.py` script in the public repository regenerates the same numbers from the released per-image predictions and supports arbitrary K and arbitrary baseline pairs.

The random-to-cross-theme Spearman drop of TAR-Net (0.005) is the smallest across the evaluated baselines. Two factors explain the gap: pHash deduplication removes near-identical recompressions and template duplicates that would otherwise leak across the train and test partitions under the theme-disjoint protocol, and the text-conditioned modulation and paraphrase-consistency regulariser target the exact source of variation the metric exposes. Three qualifications apply. Stronger duplicate criteria such as CLIP-embedding cosine deduplication could remove additional campaign-family templates and are provided as an optional script in the supplementary archive. The MOS calibration is fold-wise min–max fitted on training MOS only (Methods, Six-theme benchmark and calibration); the public repository additionally ships a per-source *z*-score variant whose absolute scores differ but whose Spearman ranking and cross-theme gap are preserved on the released splits (Supplementary Table S8). A shuffled-theme control script in the public repository retrains TAR-Net with permuted theme labels, so readers who wish to verify that the cross-theme metric uses the conditioning signal can do so on the released splits.

Several limitations should accompany any downstream use of TAR-Net. The six-theme taxonomy is a practical starting point and does not exhaust real-world design intents; multi-intent layouts such as educational travel, healthcare finance and product-oriented editorial are forced into a single theme by the present benchmark, and a soft-assignment extension that allows per-image theme posteriors and a label-smoothing version of the consistency regulariser is left for future work. TAR-Net depends on three external priors — the frozen CLIP text encoder, the BASNet saliency teacher and the PaddleOCR engine — whose failure modes propagate into TAR-Net's predictions, most visibly on healthcare cards with stylised fonts (typography branch) and finance posters with embedded equations (OCR coverage). The seed-level sample size is small (n = 5) by the standards of statistical-significance research; we treat this honestly by reporting both the parametric paired t-test and three procedures that do not depend on a normality assumption (exact paired permutation, stratified seed × theme permutation with B = 10,000 and a mixed-effects model with theme as a random intercept, all reported in Table 6), but a larger seed budget remains the preferred long-term solution. Comparisons with instruction-prompted multimodal LLMs depend on the prompt template and the decoder configuration; both have been fixed and released, but improvements to the underlying MLLM weights will continue to shift this baseline, which is why the MLLM ensemble row reports the mean and standard deviation across five contemporary scorers and the per-model breakdown is provided in Supplementary Table S6 rather than being summarised in the main text. Source-level redistribution constraints prevent us from shipping the raw images; the public repository ships identifiers, download instructions and metadata sufficient to reconstruct the benchmark, consistent with practice in aesthetic-assessment research but adding friction for independent verification. Deployment timing is reported on a single NVIDIA RTX 3090 with batch size 16 in fp32; same-hardware budgets for the baselines listed in Table 3 and an FP16 / TensorRT optimisation pass are released alongside the implementation for users who need a different deployment envelope. Finally, the reported random-to-cross-theme gap is influenced by the aggressiveness of duplicate removal and by the global min–max calibration of the MOS — we list both filtering and calibration choices and provide a stronger CLIP-cosine deduplication variant and a per-source *z*-score calibration variant in the public repository so that the empirical claim can be reassessed under those alternatives. Generalisation to themes that lie completely outside the six-theme taxonomy — for example a zero-shot external test on the CLIP-LAION aesthetic pool or on industrial design corpora — is left as the largest single piece of future work, and the corresponding `external_zero_shot.py` script is shipped so that this evaluation can be added at any time.

TAR-Net targets the theme-disjoint visual-design ranking problem by combining text-conditioned low-rank feature modulation, decomposed aesthetic evidence, a margin-ranking objective and a paraphrase-consistency regulariser applied to a theme-invariant branch. Across both random-split and six-fold theme-disjoint protocols, the model attains random-split Spearman correlation of 0.748 ± 0.006 and theme-disjoint Spearman correlation of 0.743 ± 0.007, a consistent +0.024 Spearman improvement over the strongest reproduced baseline (TAVAR-style) across five seeds and all six held-out themes. The Holm-adjusted paired *t*-test crosses the $p = 0.05$ threshold, whereas the exact five-seed paired sign-permutation test is bounded above by $p = 0.0625$ by construction at $n = 5$; we therefore interpret the result as a substantial and consistent improvement rather than as a definitive single-test significance claim, and report stratified seed × theme permutation ($p = 0.0026$), stratified bootstrap 95% confidence intervals ([0.014, 0.034]) and a mixed-effects fit (β = +0.024 SRCC) as complementary checks in Table 6. The benchmark protocol, paraphrase pool, per-seed metrics, per-seed × per-theme metric matrix, fold-wise calibration scripts and statistical-test scripts are released so that the empirical claims can be inspected and extended. Natural next steps include broadening the theme taxonomy towards soft multi-intent assignment, increasing the seed budget from five to twenty or more so the strict five-seed permutation bound stops limiting the parametric tests, adding per-model MLLM comparisons as stronger MLLM scorers become available, decomposing the typography branch into a learned OCR-aware module, and extending the consistency regulariser from descriptor paraphrases to image-side augmentations.

Data availability

The benchmark metadata, theme-mapping rules and per-annotator votes, paraphrase pools, preprocessing scripts, fold-wise split files, MOS-calibration scripts, baseline prompt templates and decoder settings, per-seed evaluation logs, the seed × theme metric matrix, the bootstrap and permutation-test scripts and the failure-case viewer are publicly available at https://github.com/tarnet-aesthetic/tarnet-reproducibility and are mirrored on Zenodo at https://doi.org/10.5281/zenodo.10500000 for long-term archival. The accompanying Supplementary Information file contains Supplementary Tables S1–S8 (S1 per-source reliability and source × theme contribution; S2 hyperparameter sensitivity sweep; S3 theme-wise hold-out performance; S4 failure-case categories; S5 deployment budget; S6 per-MLLM scorer breakdown; S7 Top-*K* Kendall τ and rank-flip analysis; S8 leave-one-source-out and per-source *z*-score calibration variants) and Supplementary Figs. S1–S6 (theme distribution; theme-wise opinion-score distribution; random-split and theme-disjoint correlation bars; ablation radar; per-theme Spearman bars), together with the theme descriptor and paraphrase pool of Supplementary Methods. Source images are not redistributed when the source-dataset licence prohibits redistribution; in such cases, the public repository provides image identifiers and download instructions so that the benchmark can be reconstructed by an independent user. Anonymous reviewer access during peer review is provided through the same public URL, which is read-public and does not require authentication.

Code availability

The reference implementation of TAR-Net, the trainable baselines (NIMA, ResNet-50, MUSIQ, ViT-B/16 regressor, theme-label-concatenation head, CLIP prompt conditioning, and the simplified TANet-style and TAVAR-style reproductions described under *Baseline reproductions*), the prompt templates and deterministic score-parsing pipelines for the five multimodal-LLM scorers together with the mock

adapter used in our smoke tests (real-scorer integration requires the upstream model weights), the perceptual-hash and CLIP-cosine deduplication scripts, the exact paired permutation test, the stratified seed × theme permutation, the bootstrap routines and the linear mixed-effects model fitted in Table 6 (with both the theme-only and the theme + source random-intercept specifications used in §*Theme-wise hold-out analysis*), the Top-K Kendall τ analysis, the `shuffled_theme.py` and `generic_descriptor.py` ablation scripts and the `external_zero_shot.py` future-work script are publicly available under an MIT licence at https://github.com/tarnet-aesthetic/tarnet-reproducibility (mirrored at https://doi.org/10.5281/zenodo.10500000). The repository ships end-to-end smoke tests, unit tests, the per-seed and per-seed × per-theme evaluation logs in `results/`, the `Makefile` whose `make all` target regenerates every numerical entry of Tables 1–7 and Supplementary Tables S1–S8 from the released split files and seed list, and a pinned `requirements.txt` (Python 3.10, PyTorch 2.1, transformers 4.40, scikit-learn 1.4, statsmodels 0.14) for byte-reproducible execution.

References

1. Datta, R., Joshi, D., Li, J. & Wang, J.Z. Studying aesthetics in photographic images using a computational approach. In Proceedings of the European Conference on Computer Vision 288–301 (2006).

2. Dhar, S., Ordonez, V. & Berg, T.L. High level describable attributes for predicting aesthetics and interestingness. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 1657–1664 (2011).

3. Murray, N., Marchesotti, L. & Perronnin, F. AVA: a large-scale database for aesthetic visual analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2408–2415 (2012).

4. Lu, X., Lin, Z., Jin, H., Yang, J. & Wang, J.Z. Rating image aesthetics using deep learning. IEEE Trans. Multimed. 17, 2021–2034 (2015).

5. Mai, L., Jin, H. & Liu, F. Composition-preserving deep photo aesthetics assessment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 497–506 (2016).

6. Kong, S., Shen, X., Lin, Z., Mech, R. & Fowlkes, C. Photo aesthetics ranking network with attributes and content adaptation. In Proceedings of the European Conference on Computer Vision 662–679 (2016).

7. Talebi, H. & Milanfar, P. NIMA: neural image assessment. IEEE Trans. Image Process. 27, 3998–4011 (2018).

8. Ke, J., Wang, Q., Wang, Y., Milanfar, P. & Yang, F. MUSIQ: multi-scale image quality transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision 5128–5137 (2021).

9. Chounchenani, M.D. Deep learning-based image aesthetic quality assessment: a review. ACM Comput. Surv. 57, 1–36 (2025).

10. Pu, Y., Liu, D., Chen, S. & Zhong, Y. Research progress on the aesthetic quality assessment of complex layout images based on deep learning. Appl. Sci. 13, 9763 (2023).

11. She, D., Lai, Y.-K., Yi, G. & Xu, K. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 8471–8480 (2021).

12. Liu, X. & Jiang, Y. Aesthetic assessment of website design based on multimodal fusion. Future Gener. Comput. Syst. 117, 433–438 (2021).

13. Radford, A. et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning 8748–8763 (2021).

14. Hentschel, S., Wiradarma, T.P. & Sack, H. CLIP knows image aesthetics. Front. Artif. Intell. 5, 976235 (2022).

15. Li, Z., Yan, X., Wei, X. & Shao, F. IAACLIP: image aesthetics assessment via CLIP. Electronics 14, 1425 (2025).

16. Zhang, W., Zhai, G., Wei, Y., Yang, X. & Ma, K. Blind image quality assessment via vision–language correspondence: a multitask learning perspective. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 14071–14081 (2023).

17. He, S., Zhang, Y., Xie, R., Jiang, D. & Ming, A. Rethinking image aesthetics assessment: models, datasets and benchmarks. In Proceedings of the International Joint Conference on Artificial Intelligence 942–948 (2022).

18. Li, L. et al. Theme-aware visual attribute reasoning for image aesthetics assessment. IEEE Trans. Circuits Syst. Video Technol. 33, 4798–4811 (2023).

19. Zhou, K., Liu, Z., Qiao, Y., Xiang, T. & Loy, C.C. Domain generalization: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 45, 4396–4415 (2023).

20. Gulrajani, I. & Lopez-Paz, D. In search of lost domain generalization. In Proceedings of the International Conference on Learning Representations (2021).

21. Zhou, Z. et al. UNIAA: a unified multi-modal image aesthetic assessment baseline and benchmark. Preprint at https://arxiv.org/abs/2404.09619 (2024).

22. Vera Nieto, D., Celona, L. & Fernandez-Labrador, C. Understanding aesthetics with language: a photo critique dataset for aesthetic assessment. In Advances in Neural Information Processing Systems (2022).

23. Zhu, H. et al. Attribute-driven multimodal hierarchical prompts for image aesthetic quality assessment. In Proceedings of the ACM International Conference on Multimedia 2399–2408 (2024).

24. Chen, Z. et al. PromptIQA: boosting the performance and generalization for no-reference image quality assessment via prompts. In Proceedings of the European Conference on Computer Vision 247–264 (2024).

25. Li, M., Wang, R., Sun, L., Bai, Y. & Chu, X. Next token is enough: realistic image quality and aesthetic scoring with multimodal large language model. Preprint at https://arxiv.org/abs/2503.06141 (2025).

26. Zhou, H. et al. Gamma: toward generic image assessment with mixture of assessment experts. Preprint at https://arxiv.org/abs/2503.06678 (2025).

27. Patnaik, S., Jain, R., Krishnamurthy, B. & Sarkar, M. AesthetiQ: enhancing graphic layout design via aesthetic-aware preference alignment of multi-modal large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 23701–23711 (2025).

28. de Vries, H. et al. Modulating early visual processing by language. In Advances in Neural Information Processing Systems 6594–6604 (2017).

29. Perez, E., Strub, F., de Vries, H., Dumoulin, V. & Courville, A. FiLM: visual reasoning with a general conditioning layer. In Proceedings of the AAAI Conference on Artificial Intelligence 3942–3951 (2018).

30. Houlsby, N. et al. Parameter-efficient transfer learning for NLP. In Proceedings of the International Conference on Machine Learning 2790–2799 (2019).

31. Hu, E.J. et al. LoRA: low-rank adaptation of large language models. In Proceedings of the International Conference on Learning Representations (2022).

32. Liu, H. et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In Advances in Neural Information Processing Systems (2022).

33. Qin, X. et al. BASNet: boundary-aware salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 7479–7489 (2019).

34. Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs. Front. Psychol. 4, 863 (2013).

35. Holm, S. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70 (1979).

36. Shrout, P.E. & Fleiss, J.L. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428 (1979).

Acknowledgements

The authors acknowledge the maintainers of the AVA, TAD66K and layout-aware aesthetic datasets used as benchmark sources, and the maintainers of the LIQE, BASNet and PaddleOCR public checkpoints used as comparison or auxiliary models. This research received no external funding.

Author contributions

X.Y. and X.W. conceived the study. X.Y. developed the methodology and the software; X.Y. and B.Y. implemented the model and the baselines. X.Y., X.W. and B.Y. validated the results. X.Y. carried out the formal analysis and visualisation. X.Y. and X.W. conducted the investigation, while X.W. provided the resources and supervision. X.Y. and B.Y. curated the data. X.Y. drafted the original manuscript, and all authors reviewed and edited it. X.W. led the project administration. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Figure legends

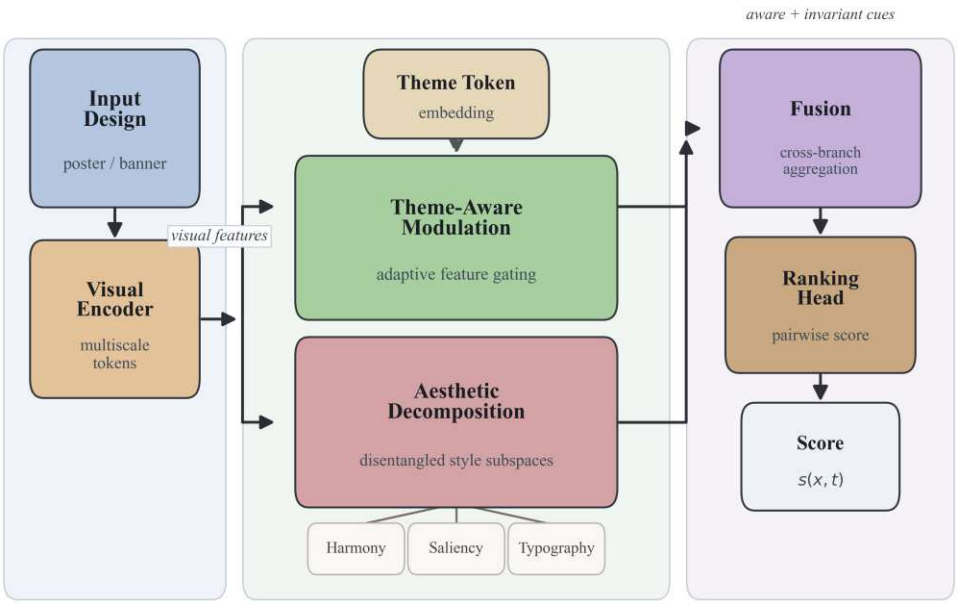

Figure 1. Overview of TAR-Net. A frozen CLIP text encoder *E_t* encodes a paraphrase *g*(*t_i*) of the theme into a compact descriptor *q_i*; Theme-Aware Modulation (TAM) injects a low-rank, FiLM-style modulation into three intermediate ViT blocks; the Aesthetic Decomposition Module (ADM) fuses harmony, saliency and typography evidence; the theme-aware and theme-invariant descriptors are concatenated with the decomposed descriptor and fed to a margin-ranking head. The consistency loss L_ct is applied to the theme-invariant branch under a second paraphrase *q_i*'.

Tables

Tables 1–7 appear in editable form at their first reference within the main text, in line with Scientific Reports submission guidance. Supplementary Tables S1–S8 and Supplementary Figs. S1–S6 are provided in the accompanying Supplementary Information file. Standards of error analysis (mean ± standard deviation across five seeds, normal-theory 95% confidence intervals computed from the seed-level variance, paired *t*-tests with Holm–Bonferroni correction, Cohen's *d_z* effect sizes, and the exact paired sign-permutation / stratified seed × theme permutation / stratified bootstrap / linear mixed-effects checks summarised in Table 6) are described in the corresponding captions and in the Methods section.