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.

Diagnosing and mitigating shortcut learning in AI-generated image quality assessment with counterfactual prompt ranking and prompt-decoupled quality modelling

Shengyu Pei

shengyupei@gxmzu.edu.cn

Guangxi Minzu University

Yihua Wu

Guangxi Minzu University

Yule An

Guangxi Minzu University

Sisi Fan

Guangxi Minzu University

Huan Lao

Guangxi Minzu University

Research Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11582191/v1

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

Additional Declarations: No competing interests reported.

Diagnosing and mitigating shortcut learning in AI-generated image quality assessment with counterfactual prompt ranking and prompt-decoupled quality modelling

Shengyu Pei[1,2*], Yihua Wu[1], Yule An[1], Sisi Fan[1], Huan Lao[1,2]

[1*]School of Artificial Intelligence, Guangxi Minzu University, Nanning, 530006, China.
[2]Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Guangxi Minzu University, Nanning, 530006, China.

*Corresponding author(s). E-mail(s): shengyupei@gxmzu.edu.cn;

Abstract

AI-generated image quality assessment (AGIQA) rates the perceptual quality of a generated image and its alignment with the prompt. We examine what two widely used benchmarks, AGIQA-3K and AIGCIQA2023, reward and find three shortcuts. First, generator identity alone explains 57% and 53% of the variance of the quality mean opinion score (MOS), and a pixel-free predictor that returns the training-mean MOS of the image's generator reaches a Spearman rank-order correlation coefficient (SROCC) of 0.63 and 0.69. Second, random image-level splits leak prompts: holding out test prompts lowers the alignment SROCC of the same model from 0.766 to 0.672 and its correspondence SROCC from 0.725 to 0.556. Third, a prompt-conditioned regressor built on contrastive language–image pre-training (CLIP) prefers the true prompt to a random one for only 74% and 58% of images, while its quality outputs change when the prompt is swapped. We propose CPR-IQA, which adds a counterfactual prompt ranking (CPR) hinge constraint to the alignment regressor and predicts quality-type scores from image features alone. Under prompt-disjoint cross-validation (ViT-B/32), CPR-IQA raises alignment-type SROCC over a fine-tuned image-only CLIP baseline from 0.672 to 0.747 and from 0.556 to 0.627, and random-prompt counterfactual accuracy to 93%; on AIGCIQA2023 the SROCC gain comes from prompt conditioning, and CPR changes the counterfactual behaviour. Under a generator × prompt double hold-out, alignment-type SROCC rises from 0.285 to 0.449 and from 0.263 to 0.370. CPR-IQA remains less sensitive than zero-shot CLIPScore to hard prompt substitutions, and the quality shortcut on unseen generators remains unsolved.

Keywords: AI-generated images, Image quality assessment, Text–image alignment, Shortcut learning, Counterfactual evaluation, CLIP

1 Introduction

Text-to-image generators are now in everyday use, yet images generated from the same prompt differ widely in sharpness, structural plausibility, artefacts and semantic fidelity. AI-generated image quality assessment (AGIQA) has therefore become a distinct problem within image quality assessment (IQA) [1–5]. General-purpose IQA continues to advance on natural images with capture and processing distortions [6], for example with region-level assessment of user-specified regions

of interest [7] and with a single framework that handles both full-reference and no-reference evaluation [8]. AGIQA differs from this setting in having at least two targets: the perceptual quality of the image and its correspondence, or alignment, with the prompt that produced it. Databases such as AGIQA-3K [2] and AIGCIQA2023 [3] provide mean opinion scores (MOS) for both. Fine-tuned or zero-shot models built on CLIP [9] have become the default solution [10–13].

Correlation coefficients on these benchmarks are high, but it is less clear what the models learn. We revisit AGIQA from the perspective of shortcut learning [14] and ask three questions. Which parts of the benchmark signal can be predicted without the evidence the task is supposed to rely on? Does the usual evaluation protocol reward memorisation? Do prompt-conditioned models actually use the prompt? The answers point to three issues that published AGIQA evaluations seldom examine.

The first issue is a *generator shortcut*. AGIQA databases are built by running a few generators over a shared prompt set, and the generators differ markedly in average quality. A predictor that ignores the pixels and returns the training-set mean MOS of the image's generator reaches a quality SROCC of 0.626 on AGIQA-3K (0.720 if the generation setting is also used) and 0.694 on AIGCIQA2023 under prompt-disjoint splits. Generator identity accounts for $\eta^2 = 0.57$ and 0.53 of the quality MOS variance, respectively. It is a legitimate correlate of quality, since older generators do produce worse images. The problem lies in the benchmark: when between-generator variance dominates, pooled SROCC largely measures whether a model can tell generators apart. It overstates the model's ability to rank images from the *same* generator, which is what is needed to choose among candidate generations or to assess a new generator. We therefore report within-generator SROCC (wG) and results on generators unseen in training.

The second issue is *prompt leakage*. The common random image-level split puts images of the same prompt in both training and test sets. Alignment MOS depends strongly on the prompt, so a model can score well by memorising which prompts are hard. For a fine-tuned image-only CLIP model, alignment SROCC on AGIQA-3K falls from 0.766 under a random split to 0.672 under a prompt-disjoint split, and correspondence SROCC on AIGCIQA2023 from 0.725 to 0.556.

The third issue is *text under-utilisation with prompt cross-talk*. We probe prompt-conditioned models by replacing the true prompt of a test image with another prompt. An alignment score that uses the text should fall, and a quality score that is meant to be prompt-independent should not move. For a CLIP regressor with a standard image–prompt fusion, the alignment score falls for only 74% of AGIQA-3K images and 58% of AIGCIQA2023 images (50% is chance), while the quality score shifts by 0.11 and 0.12 MOS standard deviations (0.13 for AIGCIQA2023 authenticity). The model uses the text weakly, and the text leaks into outputs that should not depend on it.

We propose CPR-IQA, a deliberately simple model with two components. The first is a *counterfactual prompt ranking* (CPR) constraint. During training, each image is also scored against a prompt from another prompt group in the same prompt-grouped batch, and a hinge loss requires the score under the true prompt to exceed the mismatched score by one MOS standard deviation. CPR is the regression-side counterpart of the in-batch negative ranking used in image–text matching [15–17]; it constrains a MOS regressor rather than an embedding space or a matching classifier. The second component is a *prompt-decoupled quality branch*. Quality-type scores are predicted only from image features, namely the global CLIP embedding and token statistics from four depths of the fine-tuned CLIP image encoder, so they are invariant to the prompt by construction.

We evaluate under two protocols designed to close the shortcuts, prompt-disjoint five-fold cross-validation with three seeds and a generator $\times$ prompt double hold-out in which both the test generator and the test prompts are unseen in training, and we test differences with prompt-cluster paired bootstrap and corrected resampled t-tests. All CLIP configurations share a 25-epoch budget and select checkpoints on validation prompts only. Five no-reference IQA (NR-IQA) methods are re-trained on the same folds.

The contributions are as follows.

1. **A shortcut-diagnostic protocol for AGIQA.** It consists of a pixel-free generator-mean predictor, an η^2 decomposition of MOS variance by generator and by prompt, within-generator SROCC, a paired random-versus-prompt-disjoint leakage test, and

counterfactual prompt accuracy and shift under five substitution schemes, with frozen CLIP-Score as a reference. Applied to two public databases, it documents the three issues above.
2. **Counterfactual prompt ranking (CPR).** An annotation-free hinge constraint that turns in-batch negative ranking into a constraint on a MOS regressor. It adds one pass of a small fusion head during training and nothing at inference. It makes alignment scores respond to the prompt (random-substitution accuracy 93% on both databases with ViT-B/32), and the alignment-type SROCC gain of CPR-IQA over the image-only baseline grows from +0.075 and +0.071 under prompt-disjoint splits to +0.164 and +0.107 when both generator and prompts are unseen.
3. **Prompt-decoupled multi-level quality modelling.** Quality-type outputs are structurally independent of the prompt, which removes prompt cross-talk by construction.
4. **A stricter evaluation with negative results.** It includes same-fold re-training of established NR-IQA models, frozen CLIP probes as strong references, a double hold-out for unseen generators, hard counterfactual substitutions, and an account of what does not work: CPR-IQA remains less sensitive than zero-shot CLIPScore to near-duplicate prompts, and no loss-based remedy we tried improves quality prediction for unseen generators.

We do not claim a new architecture. Our claim is that a model built this way, trained with CPR and evaluated with the proposed diagnostics, gives more trustworthy alignment predictions than the common alternatives, and that the diagnostics expose limitations that pooled correlations hide.

2 Related work

2.1 AGIQA databases and models

AGIQA-1K [1] was the first to collect perceptual-quality MOS for text-to-image outputs. AGIQA-3K [2] extended this to six generators, from AttnGAN to Midjourney, and 300 prompts, and added an alignment MOS. AIGCIQA2023 [3] rates 2400 images from six generators and 100 prompts along three dimensions: quality, authenticity and correspondence. AIGIQA-20K [4] scales the collection further. These four databases share a factorial generator × prompt structure, which is the root of the shortcuts we study. More recent collections broaden the annotation. AIGCIQA2023+ [18] provides human visual preference scores for generated images and is used to train a quality model by instruction tuning. AIGI-IQAD [5] combines a subjective study with an objective quality method. AGHI-QA [19] focuses on AI-generated human images, where global quality scores miss the anatomical and textural distortions that generated humans often show, and pairs a subjective-aligned dataset with a fine-grained metric.

Early AGIQA baselines transferred NR-IQA models designed for natural or synthetically distorted images [2]. These include natural-scene-statistics models [20], CNN regressors such as CNNIQA [21] and DBCNN [22], the content-adaptive HyperIQA [23], transformer models such as MUSIQ [24] and MANIQA [25], and the top-down multi-scale TOPIQ [26]. Several of these aggregate features from multiple depths [23, 25, 26], because low-level artefacts are poorly represented in the final semantic layer. Our multi-level token statistics follow the same intuition inside a CLIP ViT, with simple first- and second-order pooling. The flow of ideas also runs the other way: Sang et al. [27] use generated images as references for NR-IQA training, synthesise distorted versions with a cold diffusion model and label them with a full-reference metric, which addresses the shortage of annotated real-world training data.

Vision–language models have become the dominant backbone. CLIP-IQA [10] scores images zero-shot by contrasting antonym prompts, and its fine-tuned variant CLIP-IQA+ learns the prompt context. LIQE [11] fine-tunes CLIP on a multitask correspondence among scene, distortion type and quality level. Q-Align [28] teaches a large multimodal model to emit discrete rating levels, and Zhu et al. [29] query large multimodal models with two-alternative forced choice (2AFC) prompts and aggregate the pairwise answers into global quality scores by maximum a posteriori estimation. Descriptive IQA, in which a vision–language model describes quality in language, has been extended with a large-scale multi-modal dataset [30]. AGIQA-specific methods feed the generation prompt into the model. IP-IQA [12] fine-tunes CLIP with an image–prompt fusion module and image-to-prompt pre-training; IPCE [13] estimates quality through the correspondence between the

image and templates that combine the prompt with quality descriptors; and MA-AGIQA [31] injects semantic features from a large multimodal model through mixture-of-experts fusion. Xia et al. [32] design task-specific textual prompts for the perceptual and alignment tasks, rather than prompts based solely on the generation prompt, and replace coarse-grained image–text similarity with similarity at multiple granularities. Other recent designs model the consistency among prompt, image and a caption of the image [33], use a cross-modality interactive attention network [34], or apply contrastive learning guided by perceptual features [35]. ELIQ [36] proposes a label-free framework for the quality of evolving AI-generated images, a concern close to our evaluation on unseen generators. IP-IQA, IPCE and MA-AGIQA mostly report results on random image-level splits, and to our knowledge none of them tests whether its alignment output changes when the prompt is replaced, or whether its quality output stays unchanged. CPR is a loss term and is complementary to these architectures.

2.2 Vision–language alignment metrics

CLIPScore [37] is the rescaled cosine similarity between CLIP image and text embeddings. Semantic Similarity Distance [38] was proposed as a text–image consistency metric for text-to-image generation. ImageReward [39], PickScore [40] and HPS v2 [41] fine-tune vision–language encoders on human preference comparisons between generations of the same prompt, so their outputs mix alignment with overall preference; Tian et al. [42] review recent advances in aligning text-to-image diffusion models with human preferences. TIFA [43] checks prompt-derived question–answer pairs with a VQA model, and VQAScore [44] uses the probability that an image-to-text model answers "yes" to whether the image shows the prompt. Alignment information can also be read from the generator itself, through attention maps or loss functions. ELBO-T2IAlign [45] questions the assumption that this information reflects perfect text–image alignment, evaluates pixel-level alignment with zero-shot referring image segmentation as a proxy task, and calibrates it with an ELBO-based method. These scoring metrics compare an image with a text by construction and therefore respond to the prompt, though not necessarily as humans do; contrastive encoders are known to behave like bags of words [17, 46]. We use frozen CLIPScore as a reference for counterfactual sensitivity.

2.3 Shortcut learning and negative pairs

Geirhos et al. [14] describe shortcut learning as reliance on decision rules that work on the benchmark distribution but fail to transfer to the intended one. The standard remedies are to change the evaluation so that the shortcut no longer pays off, and to change the training signal so that the intended cue becomes necessary. We follow both routes: prompt-disjoint and double hold-out splits close the evaluation loopholes, and CPR changes the training signal. On the generation side, biased semantic alignment is a related problem, which Chen et al. [47] address at inference time by context-aware mediation in the latent space of text-to-image models. On the training side, CPR belongs to the family of negative-pair objectives. VSE++ [15] uses a max-margin loss over in-batch negatives; ALBEF [16] adds an image–text matching head trained on in-batch negatives, and NegCLIP [17] fine-tunes CLIP with perturbed hard-negative captions. CPR shares the in-batch negative-prompt idea but constrains the ordering of a calibrated regression output, measured in MOS units, while that output is being fitted to MOS. Our evaluation borrows from counterfactually augmented data [48] and Winoground [46]: our attribute-swap and style-swap prompts are minimal, automatically generated edits. The substitution is counterfactual in the interventional sense, with the image fixed and the text changed. We do not claim to recover the counterfactual MOS, which would require human ratings of the edited pairs.

3 Shortcut diagnostics

A database is a set of images x, each with a prompt $p(x)$, a generator $g(x)$ and one MOS $y_k(x)$ per rated dimension k. Quality and authenticity are *quality-type* dimensions; alignment (AGIQA-3K) and correspondence (AIGCIQA2023) are *alignment-type* dimensions.

The generator-mean predictor ignores the pixels and outputs the mean training MOS of the test image's generator,

$$\hat{y}_k^{\text{gen}}(x) = \frac{1}{|\mathcal{D}^{\text{tr}}_{g(x)}|} \sum_{x' \in \mathcal{D}^{\text{tr}}_{g(x)}} y_k(x'), \tag{1}$$

fitted on the training and validation images of each fold. On AGIQA-3K we also report a *generator + setting* variant that conditions on the generation setting. The predictor uses training labels and metadata; it is a supervised reference and not a competitor. SROCC is computed with mid-ranks for ties.

To quantify how much of the MOS is explained by metadata, we decompose the variance of each dimension by a grouping factor G (generator, generator + setting, or prompt):

$$\eta_G^2 = \frac{\sum_g n_g(\bar{y}_g - \bar{y})^2}{\sum_x (y(x) - \bar{y})^2}. \tag{2}$$

The in-sample PLCC of the generator-mean predictor equals η_{gen}, so a large η^2_{gen} allows a pixel-free predictor to score highly. A large η^2_{gen} is not a flaw in itself, but it means that pooled SROCC is dominated by between-generator variance. We use the term *generator shortcut* for a solution that the benchmark rewards and that fails where it matters in practice, namely for ranking outputs of one generator and for assessing unseen generators. η^2_{prompt} plays the same role for prompt leakage.

Within-generator SROCC is computed separately within each generator of a test fold and then averaged,

$$\text{wG} = \frac{1}{|\mathcal{G}_{\text{te}}|} \sum_{g \in \mathcal{G}_{\text{te}}} \text{SROCC}\big(\{\hat{y}(x), y(x)\}_{x \in \mathcal{D}^{\text{te}}_g}\big). \tag{3}$$

wG is invariant to any per-generator offset. Within-generator MOS ranges are narrower than the pooled range, so the attainable wG is lower than the attainable pooled SROCC; we use wG to compare models, not as an absolute target.

For the prompt-leakage test, every model is trained under two five-fold partitions that differ only in the unit of splitting: images (random) or prompts (prompt-disjoint). Leakage is the difference $\Delta_{\text{leak}} = \text{SROCC}_{\text{random}} - \text{SROCC}_{\text{prompt}}$.

Counterfactual prompt sensitivity is measured for a prompt-conditioned model with score $s_k(x, p)$. We replace the true prompt p of each test image with a substitute p', re-encode p' with the same frozen text encoder, and compute the counterfactual accuracy and shift

$$\begin{aligned} \text{CFA}_k &= \frac{1}{|\mathcal{D}^{\text{te}}|} \sum_x \mathbb{1}[s_k(x,p) > s_k(x,p')], \\ \text{CFS}_k &= \frac{1}{\sigma_k |\mathcal{D}^{\text{te}}|} \sum_x |s_k(x,p) - s_k(x,p')|, \end{aligned} \tag{4}$$

where σ_k is the test MOS standard deviation. CFA should approach 1 for alignment-type dimensions (0.5 is chance), and CFS should be 0 for quality-type dimensions. In the (rare) case of a tie, $s_k(x,p) = s_k(x,p')$, the strict inequality in Eq. (4) counts it as a failure, which makes CFA a conservative estimate; ties are negligible at the reported precision because scores are continuous. A positive CFS on a quality-type dimension is *prompt cross-talk*. A CFA well below that of a zero-shot reference on an alignment-type dimension is *text under-utilisation*. We use five substitution schemes:

- **Random:** p' is the prompt of a randomly chosen test image with a different prompt.
- **Attribute swap (AGIQA-3K):** the adjective field of the prompt is replaced with another adjective from the database; object and style are kept.
- **Style swap (AGIQA-3K):** the style field is replaced in the same way.
- **Nearest neighbour:** p' is the most similar different prompt in the whole database by cosine similarity of frozen CLIP text embeddings.
- **Same sub-category (AIGCIQA2023):** p' is drawn from a different prompt in the same content sub-category of the database.

Hard substitutes have no human ratings, and a minimally edited prompt may still be partly satisfied by the image, so the attainable CFA is below 1 and unknown. We therefore report the CFA of frozen CLIPScore with the same backbone on exactly the same substitutions.

4 Method

CPR-IQA (Fig. 1) consists of a CLIP image encoder that is fine-tuned end to end, a frozen

CLIP text encoder, a prompt-decoupled quality branch fed by multi-level token statistics, a prompt-conditioned alignment branch, and the CPR constraint, which is active only during training.

4.1 Encoders

An image is resized to a short side of 224 pixels and centre-cropped, then passed through the CLIP ViT image encoder, which yields the projected global embedding $v \in \mathbb{R}^d$ and the token sequence $Z_l = [z_l^{\text{cls}}; Z_l^{\text{patch}}]$ at every block l. The prompt is encoded by the frozen CLIP text encoder, $t = f_T(p)$. Keeping the text tower frozen prevents it from over-fitting a few hundred prompts and makes re-encoding a substitute prompt exact and cheap.

4.2 Prompt-decoupled multi-level quality branch

Perceptual quality and authenticity are treated as properties of the image; this is a modelling assumption, discussed in Section 7. At four depths $l \in \mathcal{S}$ we pool the class token and the mean and standard deviation of the patch tokens:

$$m_l = \text{GELU}\Big(W_l\,\text{LN}\big([z_l^{\text{cls}}; \mu(Z_l^{\text{patch}}); \sigma(Z_l^{\text{patch}})]\big)\Big) \in \mathbb{R}^{256}, \tag{5}$$

with $\mathcal{S} = \{\text{round}((i+1)L/5) - 1\}_{i=0}^{3}$ for an encoder with L blocks. For each quality-type dimension, the head is a two-layer MLP,

$$s_k(x) = h_q^k\big([v; m_{l_1}; \ldots; m_{l_4}]\big). \tag{6}$$

Intermediate-layer statistics keep texture, noise and local structure that the final CLIP embedding discards [23, 25, 26]. Because h_q^k has no text input, the counterfactual shift of every quality-type dimension is zero by construction.

4.3 Prompt-conditioned alignment branch

Alignment-type dimensions are predicted from a fusion of image and prompt embeddings together with the same multi-level statistics:

$$u = \big[v;\ t;\ \sqrt{d}\,(\hat{v} \odot \hat{t});\ \kappa\cos(v,t);\ m_{l_1}; \ldots; m_{l_4}\big], \qquad s_k(x,p) = h_a^k(u), \tag{7}$$

where $\hat{v}$ and $\hat{t}$ are ℓ_2-normalised and $\kappa = 10$. We also implemented a token-level variant in which prompt tokens cross-attend to patch tokens. It did not improve validation scores in exploratory runs, and the final model uses the global fusion.

4.4 Counterfactual prompt ranking

Training batches contain $P = 8$ prompts with $K = 4$ images each, ordered by prompt group. For image i the negative prompt is that of the image K positions further along in the batch (cyclically), $p_i' = p(x_{(i+K) \bmod |B|})$, which always belongs to a different prompt group. Reusing the image embedding, we evaluate the alignment head once more with p_i' and impose

$$\mathcal{L}_{\text{CPR}}^k = \frac{1}{|B|}\sum_{i \in B} \max\big(0,\ m - [s_k(x_i, p_i) - s_k(x_i, p_i')]\big) \tag{8}$$

for each alignment-type dimension. Scores are regressed on MOS standardised with the training-fold statistics, so $m = 1$ corresponds to one MOS standard deviation. CPR uses no extra annotation. It encodes a prior: a prompt from another group should not describe the image better than the prompt that generated it. The prior can be wrong for near-duplicate prompts, which is why negatives are drawn at random from other groups rather than as nearest neighbours; exploratory runs with nearest-neighbour negatives were worse (Section 6.8).

4.5 Objective

For each dimension the supervised loss is

$$\mathcal{L}_{\text{sup}}^k = \text{MSE}(s_k, y_k) + \big(1 - \text{PLCC}(s_k, y_k)\big) + \mathcal{L}_{\text{rank}}^k, \tag{9}$$

where $\mathcal{L}_{\text{rank}}^k$ is a logistic pairwise-ranking loss over batch pairs weighted by $\tanh(|y_{k,i} - y_{k,j}|/0.3)$. The total objective is $\mathcal{L} = \sum_k \mathcal{L}_{\text{sup}}^k + \lambda \sum_{k \in \mathcal{A}} \mathcal{L}_{\text{CPR}}^k$ with $\lambda = 1$, where $\mathcal{A}$ is the set of alignment-type

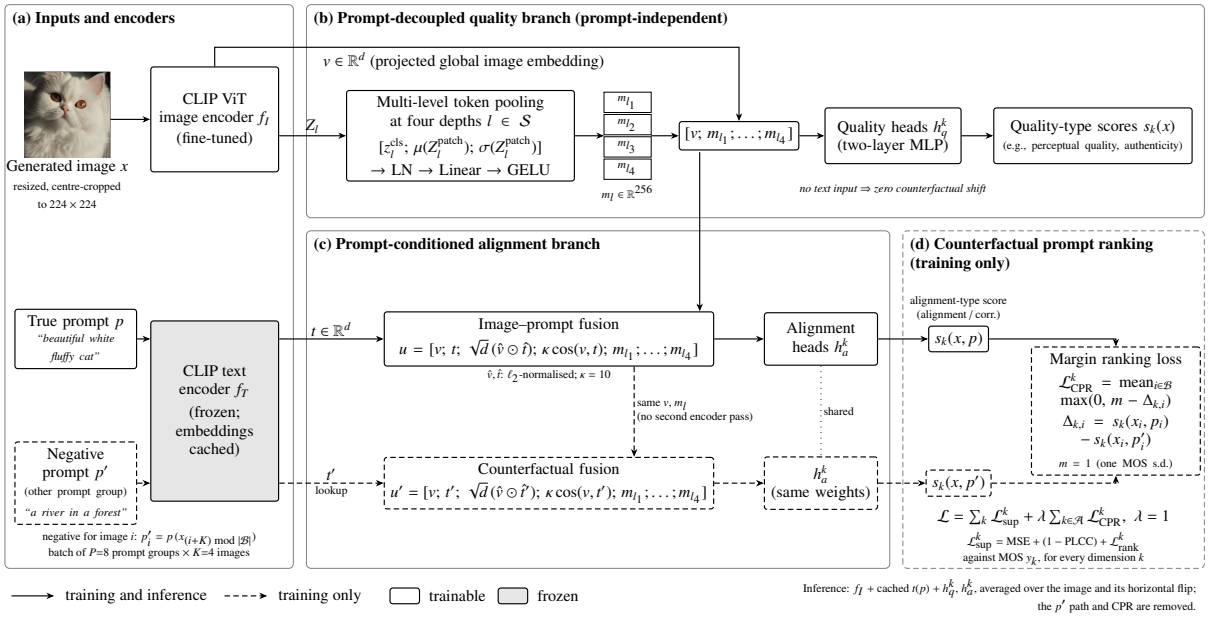

Fig. 1 CPR-IQA. Quality-type heads receive no text input; the alignment head fuses image and prompt embeddings. The counterfactual prompt ranking loss (training only) requires the alignment score under the true prompt to exceed that under a prompt from another group by a margin.

dimensions. At test time predictions are averaged over the image and its horizontal flip.

4.6 Training procedure and cost

Algorithm 1 lists one training step. Prompt embeddings are computed once for all prompts of a database, because the text encoder is frozen, so a substitute prompt costs only a lookup. The counterfactual term reuses the image embedding v and the multi-level statistics m_l of the true-prompt pass and evaluates only the alignment head a second time. CPR therefore adds no encoder pass and no annotation, and its cost is that of one extra MLP evaluation per image and alignment-type dimension. At inference CPR-IQA has the same cost as the prompt-conditioned baseline Cond: one image-encoder pass (two with the flip average) and one cached text embedding per prompt.

Algorithm 1 One training step of CPR-IQA

Require: mini-batch $\mathcal{B}$ of $P = 8$ prompt groups with $K = 4$ images each, ordered by group; cached text embeddings $t(p)$; margin $m = 1$; weight $\lambda = 1$

for each image $x_i \in \mathcal{B}$ **do**
  $v_i, \{Z_{l,i}\} \leftarrow f_I(x_i)$; $m_{l,i} \leftarrow$ Eq. (5) for $l \in \mathcal{S}$
  quality-type scores $s_k(x_i) \leftarrow h_q^k([v_i; m_{l_1,i}; \ldots; m_{l_4,i}])$ ▷ no text input
  alignment-type scores $s_k(x_i, p_i) \leftarrow h_a^k(u(v_i, t(p_i), m_{\cdot,i}))$ ▷ Eq. (7)
  $p'_i \leftarrow p\big(x_{(i+K) \bmod |\mathcal{B}|}\big)$; $s_k(x_i, p'_i) \leftarrow h_a^k(u(v_i, t(p'_i), m_{\cdot,i}))$ ▷ reuse $v_i, m_{\cdot,i}$
end for
$\mathcal{L} \leftarrow \sum_k \mathcal{L}_{sup}^k + \lambda \sum_{k \in \mathcal{A}} \frac{1}{|\mathcal{B}|} \sum_i \max\big(0, m - [s_k(x_i, p_i) - s_k(x_i, p'_i)]\big)$
update f_I and all heads with AdamW; update the EMA weights

5 Experimental setup

5.1 Databases

AGIQA-3K [2] contains 2982 images generated from 300 prompts by six generators (AttnGAN, GLIDE, DALL·E 2, Stable Diffusion v1.5, SDXL and Midjourney), some with several inference settings, each with a perceptual-quality MOS and an alignment MOS. One image was missing from the copy we obtained, so 2981 images are used. AIGCIQA2023 [3] contains 2400 images, four for each of 100 prompts from each of six generators (ControlNet, DALL·E, GLIDE, Lafite, Stable Diffusion and UniDiffuser), each with quality, authenticity and correspondence MOS.

5.2 Protocols

The main protocol, P1, is prompt-disjoint cross-validation. Prompts are partitioned into five fixed groups. In each fold, the images of one group form the test set, one ninth of the remaining prompts form the validation set, and the rest are used for training. Each configuration is run with three seeds, which re-draw the validation prompts, initialisation and data order.

P2 is a generator $\times$ prompt double hold-out. Prompts are split into two fixed halves. Each fold tests one generator on one half of the prompts; training uses only the *other* generators on the *other* half, so neither the test generator nor any test prompt is seen in training. Six generators $\times$ two halves give 12 folds, each run with three seeds. Because each test set contains a single generator, SROCC in P2 is a within-generator SROCC.

P3 is random image-level cross-validation with five folds and images assigned independently. It is used only to measure prompt leakage (one seed).

We report SROCC, Pearson's linear correlation coefficient (PLCC, computed on raw predictions without a nonlinear mapping) and wG. For all CLIP configurations, the checkpoint with the highest mean validation SROCC over the rated dimensions is kept; test folds play no role in selection.

5.3 Compared methods

References without fine-tuning: the generator-mean predictors; zero-shot CLIP-IQA [10] and CLIP-Score [37] with ViT-B/32 and ViT-L/14; and frozen-feature ridge probes on the CLIP image embeddings, with the ridge penalty chosen by leave-one-out cross-validation on the training and validation images.

Re-trained NR-IQA models: CNNIQA [21], DBCNN [22], HyperIQA [23], TOPIQ-NR [26] and CLIP-IQA+ [10], re-trained with the pyiqa implementations [49] on exactly the P1 folds (seed 0), one model per dimension, with MSE on standardised MOS, each architecture's customary learning rates, 12 epochs, and validation-prompt checkpoint selection. For alignment-type dimensions these models see the image only.

Controlled CLIP configurations share the same backbone, sampler, 25-epoch budget and selection rule. The seven configurations are (i) **Base**, image-only fine-tuned CLIP with MLP heads, the most common AGIQA baseline; (ii) **+ML**, Base with multi-level token statistics; (iii) **Cond**, a prompt-conditioned baseline in which all heads share the fusion of Eq. (7) without m_l; (iv) **Dec+ML**, prompt-decoupled quality heads with multi-level statistics but no CPR; (v) **CPR+Dec**, CPR with decoupling but without multi-level statistics; (vi) **CPR+ML (shared)**, CPR with multi-level statistics but a shared head for all dimensions; and (vii) **CPR-IQA**, with all three components. The main experiments use ViT-B/32. ViT-L/14 is run for Base, Cond and CPR-IQA.

Not reproduced: ImageReward, PickScore, HPS v2, MA-AGIQA, IP-IQA, IPCE and Q-Align [12, 13, 28, 31, 39–41] could not be run because their weights or large multimodal backbones were not obtainable in our environment.

5.4 Implementation

The encoder is initialised from OpenAI CLIP [9]. We use AdamW [50] with learning rate 1×10^{-5} for the image encoder and 1×10^{-3} for all heads, weight decay 0.05, a one-cycle schedule, gradient clipping at 2, and an exponential moving average of the weights with decay 0.99, which is used for evaluation. Training augmentation consists of random horizontal flips and random resized crops with scale 0.8–1.0. CPR uses $m = 1$ and $\lambda = 1$. All experiments run on one NVIDIA RTX 4090. Training one CPR-IQA fold for 25 epochs takes a median of 102 s on AGIQA-3K and 34 s on AIGCIQA2023 with ViT-B/32, and 133 s and 40 s with ViT-L/14 (measured while the GPU was shared with other jobs).

5.5 Statistical analysis

Tables report fold-mean SROCC: SROCC is computed within each test fold and averaged over folds and seeds, with $\pm$ the standard deviation across seeds. Paired tests use pooled cross-validated predictions. For each seed, the out-of-fold predictions are concatenated into one prediction per image, and the difference between two methods is computed on the pooled predictions and averaged over seeds. Confidence intervals come from a prompt-cluster paired bootstrap [51] with 2000 resamples, in which prompts are resampled and all their images enter or leave together. We report 95% percentile intervals and two-sided p-values. Because the bootstrap holds the trained models fixed, we

also report a Nadeau–Bengio corrected resampled t-test [52] over the 15 fold × seed differences in fold SROCC, together with the number of the 15 differences that favour the second method. Within the main comparison family, bootstrap p-values are Holm-adjusted [53]. Pooled and fold-mean differences can differ, and we never quote one as the other.

6 Results

6.1 Diagnostics

Table 1 gives the variance decomposition. Generator identity explains 57% (AGIQA-3K) and 53% (AIGCIQA2023) of quality variance, and generator + setting explains 65% on AGIQA-3K. Prompt identity explains 12% of alignment variance on AGIQA-3K and 26% of correspondence variance on AIGCIQA2023, where correspondence is the only dimension for which prompt explains more than generator. Fig. 2a–c visualises the generator effect. On AGIQA-3K, the mean quality MOS of the six generators ranges from 0.99 (AttnGAN) to 3.30 (SDXL) on a scale whose overall standard deviation is 1.00, and the distributions of the two weakest generators (AttnGAN and GLIDE) overlap little with those of the others. On AIGCIQA2023 the generator means range from 40.7 (GLIDE) to 57.5 (DALL·E), with an overall standard deviation of 9.3. A model that only recognises the generator therefore already reaches a substantial pooled SROCC (Table 1).

The pixel-free generator-mean predictor reaches a quality SROCC of 0.626 on AGIQA-3K (0.720 with the setting) and 0.694 on AIGCIQA2023 (Table 1). Zero-shot CLIP-IQA with ViT-L/14 reaches 0.658 on AGIQA-3K quality, only slightly above the generator mean without the setting and below the generator + setting mean.

Prompt leakage is large for alignment-type dimensions (Table 2). Moving from random to prompt-disjoint splits lowers the alignment SROCC of Base by 0.094 on AGIQA-3K and the correspondence SROCC by 0.168 on AIGCIQA2023, whereas quality SROCC changes by at most 0.03 and authenticity SROCC by 0.035–0.046. The gap is smaller but still present for CPR-IQA (0.084 and 0.145). Fig. 2d shows the pattern across all five dimensions: the lines for quality are almost flat, those for authenticity fall moderately, and those for alignment and correspondence fall steeply for all three models. Under random splits a model can match a test image's prompt to its training images and reuse their scores, which is not possible for new prompts.

Within-generator SROCC (Fig. 2e) shows how much of the pooled alignment performance is between-generator ranking. For AGIQA-3K alignment, the pooled SROCC of the frozen probe, Base, Cond and CPR-IQA lies between 0.672 and 0.747, but their wG ranges from 0.298 to 0.510. The image-only models (probe and Base) keep less than half of their pooled correlation once the generator is fixed, whereas CPR-IQA keeps about two thirds. On AIGCIQA2023 correspondence, the prompt-conditioned models reach a wG of 0.453–0.459 against 0.375–0.377 for the image-only ones.

Table 3 reports counterfactual sensitivity. The prompt-conditioned Cond model prefers the true prompt to a random substitute for 73.6% of AGIQA-3K images and 57.6% of AIGCIQA2023 images, while its quality outputs shift by 0.11 (AGIQA-3K) and 0.12 (AIGCIQA2023 quality) MOS standard deviations. These are the text under-utilisation and prompt cross-talk described in Section 1.

6.2 Main results under prompt-disjoint splits

Table 4 compares all methods under P1, and Fig. 3 displays the same results by dimension.

On AGIQA-3K, CPR-IQA raises alignment SROCC from 0.672 (Base) and 0.719 (Cond) to 0.747. The fold-mean difference is +0.075 over Base and +0.028 over Cond; on pooled predictions it is +0.071 (95% CI [+0.052, +0.090], Holm-adjusted $p < 0.001$) and +0.025 ([+0.009, +0.041], $p = 0.002$). All 15 fold × seed differences are positive over Base and 13/15 over Cond (corrected t-test $p < 0.001$ and $p = 0.034$). Within-generator alignment SROCC rises from 0.314 (Base) and 0.433 (Cond) to 0.510. Quality SROCC changes little (0.863 → 0.868; pooled $\Delta = +0.005$ [+0.002, +0.009], bootstrap $p < 0.001$; corrected t-test $p = 0.26$, 12/15 folds), and the improvement comes from the multi-level statistics (+ML: 0.867).

On AIGCIQA2023, CPR-IQA raises correspondence SROCC from 0.556 (Base) to 0.627, a fold-mean difference of +0.071 (pooled $\Delta = +0.075$ [+0.050, +0.098], $p < 0.001$; 14/15 folds, corrected

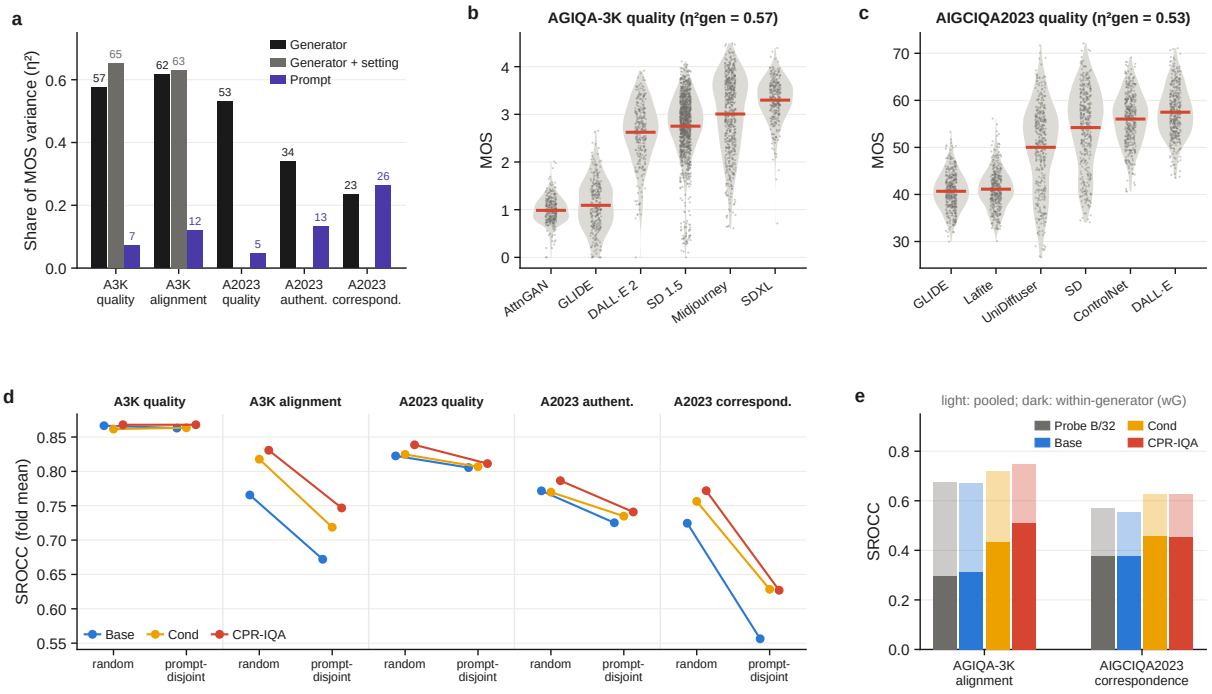

Fig. 2 Shortcut diagnostics. (a) Share of MOS variance explained by generator, generator + setting (AGIQA-3K only) and prompt identity for the five rated dimensions (A3K: AGIQA-3K; A2023: AIGCIQA2023); numbers are percentages. (b, c) Quality MOS by generator (violins, individual images and mean in red), ordered by mean. (d) Fold-mean SROCC under random image-level splits and prompt-disjoint splits for Base, Cond and CPR-IQA (ViT-B/32). (e) Pooled SROCC (light) and within-generator SROCC wG (dark) on the alignment-type dimensions under prompt-disjoint splits, for the frozen ViT-B/32 ridge probe and the three fine-tuned models.

Table 1 Share of MOS variance (η^2) explained by generator, generator + setting and prompt identity, and SROCC of the pixel-free generator-mean predictor under P1 (fold mean).

Dimension	η^2 generator	η^2 generator + setting	η^2 prompt	Generator-mean SROCC	Generator + setting SROCC
AGIQA-3K quality	0.574	0.652	0.073	0.626	0.720
AGIQA-3K alignment	0.617	0.631	0.119	0.587	0.604
AIGCIQA2023 quality	0.530	–	0.048	0.694	–
AIGCIQA2023 authenticity	0.341	–	0.131	0.585	–
AIGCIQA2023 correspondence	0.234	–	0.263	0.472	–

p = 0.019), and authenticity from 0.725 to 0.741, a fold-mean difference of +0.016 (pooled $\Delta = +0.014$ [+0.004, +0.024], Holm-adjusted bootstrap p = 0.04; corrected t-test p = 0.050, 14/15). Its correspondence SROCC, however, is *not* higher than that of Cond (0.628; pooled $\Delta = -0.001$, p = 0.93) or Dec+ML (0.635). On this database, conditioning on the prompt accounts for the whole in-distribution SROCC gain, and CPR changes the counterfactual behaviour (Section 6.5) without changing SROCC.

CPR-IQA is also better than the frozen ridge probes, which are strong references, on every dimension except AIGCIQA2023 authenticity, where the ViT-L/14 probe reaches 0.747 against 0.741. With ViT-B/32 the probes reach 0.676 (AGIQA-3K alignment) and 0.569 (AIGCIQA2023 correspondence), higher than the fine-tuned image-only Base on both. A fine-tuned image-only CLIP is therefore not a strong alignment baseline, whereas the probes are. The five NR-IQA models re-trained on the same folds are weaker than the fine-tuned image-only CLIP baseline on every dimension, although DBCNN comes within 0.001 of Base on AIGCIQA2023 quality (0.804 against 0.805). The best re-trained

Table 2 Prompt leakage: fold-mean SROCC under random (P3, one seed) and prompt-disjoint (P1, three seeds) splits.

Model	Dimension	Random split	Prompt-disjoint	Δ_{leak}
Base	AGIQA-3K quality	0.866	0.863	+0.003
Base	AGIQA-3K alignment	0.766	0.672	+0.094
Base	AIGCIQA2023 quality	0.823	0.805	+0.017
Base	AIGCIQA2023 authenticity	0.772	0.725	+0.046
Base	AIGCIQA2023 correspondence	0.725	0.556	+0.168
Cond	AGIQA-3K quality	0.862	0.863	-0.002
Cond	AGIQA-3K alignment	0.818	0.719	+0.099
Cond	AIGCIQA2023 quality	0.825	0.807	+0.018
Cond	AIGCIQA2023 authenticity	0.770	0.735	+0.035
Cond	AIGCIQA2023 correspondence	0.756	0.628	+0.128
CPR-IQA	AGIQA-3K quality	0.868	0.868	+0.000
CPR-IQA	AGIQA-3K alignment	0.831	0.747	+0.084
CPR-IQA	AIGCIQA2023 quality	0.839	0.811	+0.027
CPR-IQA	AIGCIQA2023 authenticity	0.786	0.741	+0.046
CPR-IQA	AIGCIQA2023 correspondence	0.772	0.627	+0.145

Table 3 Counterfactual accuracy (%) of the alignment-type output under five substitution schemes, and quality shift (CFS_q, MOS s.d.) under random substitution. P1, ViT-B/32, three seeds. CLIPScore: frozen ViT-B/32 on the same substitutions.

Model	A3K random	A3K attribute	A3K style	A3K nearest	A3K CFS_q	A23 random	A23 nearest	A23 same-sub-cat.	A23 CFS_q
Cond	73.6	52.8	55.6	56.8	0.107	57.6	32.9	57.6	0.123
Dec+ML	70.4	50.3	53.2	54.3	0.000	56.7	30.5	56.3	0.000
CPR+Dec	93.3	62.4	66.8	85.4	0.000	98.3	81.9	95.3	0.000
CPR+ML (shared)	93.3	62.6	66.5	84.4	0.113	97.0	76.9	93.3	0.111
CPR-IQA	93.4	62.3	66.4	84.8	0.000	93.2	66.7	88.3	0.000
CLIPScore (frozen ViT-B/32)	–	67.2	70.1	88.8	–	–	97.2	99.3	–

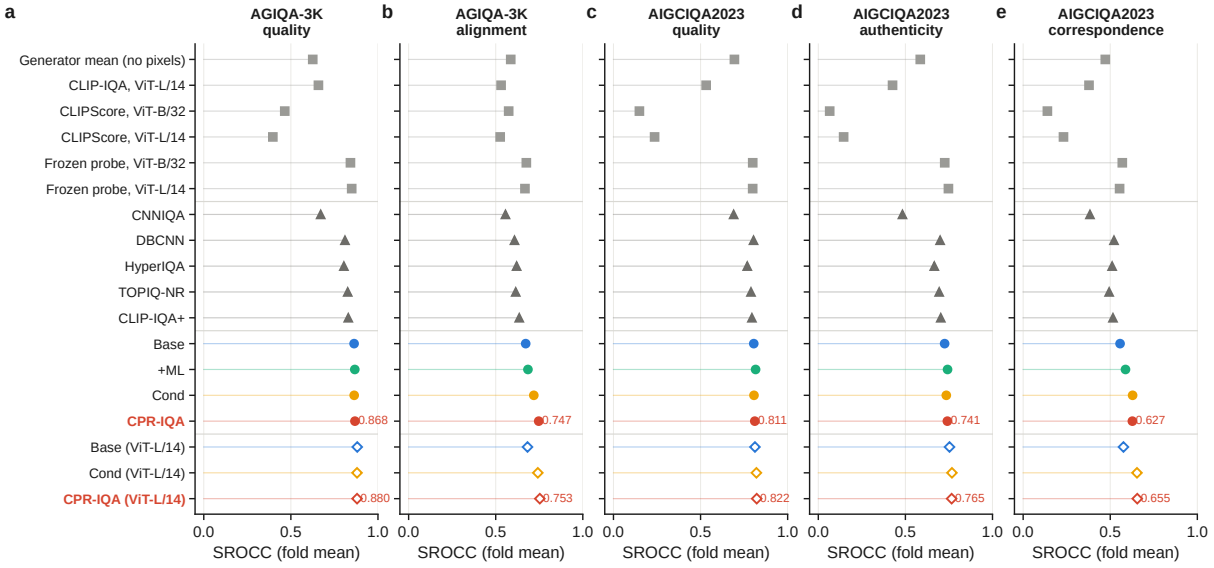

Fig. 3 Main results under prompt-disjoint cross-validation (P1). Fold-mean SROCC per rated dimension for frozen and zero-shot references (grey squares), re-trained NR-IQA models (dark triangles; seed 0), and fine-tuned CLIP models with ViT-B/32 (filled circles, ± s.d. over three seeds) and ViT-L/14 (open diamonds). Values for CPR-IQA are printed.

model per dimension reaches AGIQA-3K quality 0.830 (CLIP-IQA+) against 0.868; AGIQA-3K alignment 0.635 (CLIP-IQA+) against 0.747; AIGCIQA2023 quality 0.804 (DBCNN) against 0.811; AIGCIQA2023 authenticity 0.703 (CLIP-IQA+) against 0.741; and AIGCIQA2023 correspondence 0.521 (DBCNN) against 0.627 for CPR-IQA (Table 4). On alignment-type dimensions the

Table 4 Prompt-disjoint cross-validation (P1): fold-mean SROCC (± s.d. over three seeds; re-trained NR-IQA models one seed). A3K: AGIQA-3K; A23: AIGCIQA2023; wG: within-generator SROCC.

Method	A3K quality	A3K alignment	A3K align. wG	A23 quality	A23 authenticity	A23 corresp.	A23 corr. wG
References without fine-tuning							
Generator mean (supervised, no pixels)	0.626	0.587	–	0.694	0.585	0.472	–
CLIP-IQA zero-shot, ViT-B/32	0.547	0.393	0.155	0.422	0.312	0.289	0.206
CLIP-IQA zero-shot, ViT-L/14	0.658	0.531	0.190	0.532	0.426	0.378	0.224
CLIPScore, ViT-B/32	0.465	0.575	0.308	0.149	0.065	0.139	0.099
CLIPScore, ViT-L/14	0.397	0.526	0.292	0.236	0.145	0.232	0.130
Frozen ridge probe, ViT-B/32	0.842	0.676	0.298	0.799	0.726	0.569	0.377
Frozen ridge probe, ViT-L/14	0.849	0.668	0.287	0.799	0.747	0.554	0.347
Re-trained NR-IQA models (image only)							
CNNIQA [21]	0.671	0.556	0.176	0.690	0.483	0.384	0.106
DBCNN [22]	0.811	0.608	0.243	0.804	0.699	0.521	0.253
HyperIQA [23]	0.804	0.620	0.217	0.768	0.666	0.511	0.293
TOPIQ-NR [26]	0.826	0.615	0.201	0.789	0.694	0.494	0.264
CLIP-IQA+ [10]	0.830	0.635	0.261	0.795	0.703	0.516	0.285
Fine-tuned CLIP, ViT-B/32 (three seeds)							
Base (image only)	0.863 ± 0.001	0.672 ± 0.005	0.314	0.805 ± 0.004	0.725 ± 0.004	0.556 ± 0.014	0.375
+ML (image only)	0.867 ± 0.001	0.685 ± 0.011	0.328	0.816 ± 0.006	0.742 ± 0.001	0.588 ± 0.014	0.394
Cond (prompt-conditioned)	0.863 ± 0.002	0.719 ± 0.004	0.433	0.807 ± 0.004	0.735 ± 0.008	0.628 ± 0.018	0.459
CPR-IQA	0.868 ± 0.001	0.747 ± 0.005	0.510	0.811 ± 0.006	0.741 ± 0.006	0.627 ± 0.009	0.453
Fine-tuned CLIP, ViT-L/14 (3 seeds)							
Base (image only)	0.881 ± 0.002	0.683 ± 0.006	0.326	0.812 ± 0.002	0.753 ± 0.004	0.576 ± 0.009	0.392
Cond (prompt-conditioned)	0.880 ± 0.002	0.742 ± 0.003	0.462	0.821 ± 0.002	0.767 ± 0.003	0.654 ± 0.007	0.490
CPR-IQA	0.880 ± 0.004	0.753 ± 0.001	0.500	0.822 ± 0.004	0.765 ± 0.002	0.655 ± 0.015	0.488

image-only NR-IQA models are also well below the frozen ViT-B/32 probe, as expected for models that never see the prompt.

With the larger ViT-L/14 backbone (three seeds per configuration), all models are stronger, and the ordering on alignment-type dimensions is preserved: AGIQA-3K alignment 0.683 (Base), 0.742 (Cond) and 0.753 (CPR-IQA); AIGCIQA2023 correspondence 0.576, 0.654 and 0.655. The gap between CPR-IQA and Cond is small at this scale, as with ViT-B/32 on AIGCIQA2023, while the random-prompt CFA of CPR-IQA (91.9% and 77.3%) remains far above that of Cond (62.0% and 54.2%). Quality SROCC is 0.880 and 0.822 for CPR-IQA against 0.881 and 0.812 for Base.

With a shorter 12-epoch budget instead of 25 epochs, the main ordering on alignment-type dimensions is also preserved (AGIQA-3K alignment 0.675, 0.717 and 0.747 for Base, Cond and CPR-IQA), but some close comparisons change. At 12 epochs CPR-IQA is ahead of Cond on AIGCIQA2023 correspondence (0.635 against 0.609), and the order of the configurations within 0.005 of each other on AGIQA-3K alignment differs. Differences of that size should therefore not be interpreted. Table 5 gives both budgets for all configurations. The longer budget matters mainly for the configurations without multi-level statistics on AIGCIQA2023: Base gains 0.023 in quality and 0.032 in correspondence, Cond 0.022 and 0.019, whereas every configuration with multi-level statistics changes by at most 0.008 in quality and 0.009 in authenticity. The multi-level statistics thus also make training converge faster on the smaller database.

6.3 Within-generator and within-prompt ranking

Pooled SROCC mixes three kinds of ranking: between generators, between prompts and between images that share both. Table 6 separates them for the alignment-type dimensions with the within-generator SROCC (wG) and the within-prompt SROCC (wP), the mean SROCC over the test images of each prompt. Within a prompt the images come from different generators (and, on AGIQA-3K, different settings), so wP still rewards generator recognition: the pixel-free generator-mean predictor reaches a wP of 0.616 on AGIQA-3K and 0.548 on AIGCIQA2023. The image-only Base is only slightly above it on AGIQA-3K (0.663) and at the same level on AIGCIQA2023 (0.556), which means that most of its within-prompt ranking can be explained by knowing the generator. Prompt conditioning and CPR move both within-group measures. On AGIQA-3K, CPR-IQA reaches a wP of 0.709 and a wG of 0.510, against 0.690 and 0.433 for Cond. On AIGCIQA2023 the prompt-conditioned configurations are close to each other (wP 0.600–0.615, wG 0.425–0.463), which repeats the pattern of the pooled SROCC on that database. Among the fine-tuned models, PLCC gives the same ordering as SROCC, apart from differences of 0.001. The gap between wG

Table 5 Effect of the training budget under P1 (ViT-B/32, three seeds): fold-mean SROCC after 12 / 25 epochs.

Configuration	A3K quality 12 / 25	A3K alignment 12 / 25	A23 quality 12 / 25	A23 authenticity 12 / 25	A23 correspondence 12 / 25
Base	0.861 / 0.863	0.675 / 0.672	0.782 / 0.805	0.703 / 0.725	0.524 / 0.556
+ML	0.867 / 0.867	0.688 / 0.685	0.810 / 0.816	0.735 / 0.742	0.593 / 0.588
Cond	0.864 / 0.863	0.717 / 0.719	0.785 / 0.807	0.714 / 0.735	0.609 / 0.628
Dec+ML	0.867 / 0.868	0.718 / 0.717	0.808 / 0.815	0.736 / 0.745	0.631 / 0.635
CPR+Dec	0.860 / 0.864	0.731 / 0.742	0.783 / 0.801	0.704 / 0.727	0.610 / 0.604
CPR+ML (shared)	0.865 / 0.868	0.743 / 0.748	0.808 / 0.815	0.735 / 0.741	0.630 / 0.628
CPR-IQA	0.867 / 0.868	0.747 / 0.747	0.806 / 0.811	0.734 / 0.741	0.635 / 0.627

and wP (0.510 against 0.709 for CPR-IQA on AGIQA-3K) shows that ranking the outputs of one generator across prompts is the hardest of the three tasks for every model, including the frozen references.

6.4 Ablation

Table 7 isolates the three components. On AGIQA-3K, CPR is responsible for the alignment gain: every configuration with CPR reaches 0.742–0.748, and every prompt-conditioned configuration without it 0.717–0.719 (Dec+ML → CPR-IQA: pooled $\Delta = +0.027$ [+0.010, +0.043], $p < 0.001$). Multi-level statistics account for the quality gain (Base → +ML: pooled $\Delta = +0.004$, $p = 0.008$). Decoupling does not change SROCC (CPR+ML shared → CPR-IQA: pooled $\Delta = +0.001$ quality, -0.003 alignment, both n.s.); its role is to remove cross-talk (Section 6.5). On AIGCIQA2023 the multi-level statistics help all three dimensions (Base → +ML: +0.011, +0.017 and +0.031 fold-mean SROCC; corrected $p = 0.027$, 0.048 and 0.027). CPR without multi-level statistics (CPR+Dec) is worse than CPR-IQA on all three dimensions (pooled Δ of CPR-IQA over CPR+Dec for quality, authenticity and correspondence: +0.009, +0.011 and +0.022; bootstrap $p = 0.014$, 0.028 and 0.027), and the shared head is slightly better than the decoupled one on AIGCIQA2023 quality (pooled $\Delta = -0.007$ [−0.013, −0.001], $p = 0.022$).

On AGIQA-3K, CPR-IQA also has the highest validation score among the configurations. On AIGCIQA2023, the highest validation score belongs to Dec+ML (0.737 against 0.733 for CPR-IQA), a model whose correspondence output ranks the true prompt above a random one for only 56.7% of images. Selecting by SROCC alone would thus pick a model that barely uses the prompt, which is the practical argument for reporting counterfactual diagnostics alongside correlations.

6.5 Counterfactual sensitivity

Table 3 and Fig. 4a show that CPR makes the alignment output respond to the prompt. Random-substitution CFA rises from 73.6% to 93.4% on AGIQA-3K and from 57.6% to 93.2% on AIGCIQA2023. Decoupling sets the quality shift to exactly zero, whereas the shared-head models shift quality-type outputs by 0.11–0.13 MOS standard deviations with or without CPR (Fig. 4b).

The magnitude of the response is also informative. Fig. 4c, d shows the distribution over test images of the alignment score change when the true prompt is replaced by a random one (seed 0). For Cond and Dec+ML the distribution is centred close to zero (median 0.16 and 0.11 standard deviations on AGIQA-3K, 0.04 on AIGCIQA2023), so a random, unrelated prompt barely changes their alignment score. For CPR-IQA the median change is 1.19 standard deviations on AGIQA-3K, and the score falls for 93% of images in this seed. On AIGCIQA2023 the change is smaller (median 0.29) but positive for 87% of images in this seed. Fig. 4e relates counterfactual accuracy to SROCC. On AGIQA-3K, the configurations with CPR are better on both axes. On AIGCIQA2023, all prompt-conditioned configurations have similar SROCC (0.604–0.635), while CFA separates those with CPR (93–98%) from those without (57–58%), so the two measures carry different information.

Fig. 5 shows four AGIQA-3K test images with their true prompts and the random substitutes used in the counterfactual test. Cond does not separate the true and the substituted prompt (differences between −0.65 and +0.12 MOS units), and in three of the four cases it rates the unrelated prompt higher, for instance 3.52 for "an auto rickshaw driving through paris" against 2.87 for the true prompt describing rats in tuxedos. CPR-IQA lowers the score under the substitute by 1.09 to 1.69 MOS units in all four cases. The examples were selected to illustrate the behaviour; the distributions in Fig. 4c, d give the aggregate picture. Note

Table 6 Alignment-type dimensions under P1: pooled SROCC and PLCC, within-generator SROCC (wG) and within-prompt SROCC (wP), fold means (fine-tuned models: mean over three seeds). wG is undefined for the generator-mean predictor.

Method	A3K SROCC	A3K PLCC	A3K wG	A3K wP	A23 SROCC	A23 PLCC	A23 wG	A23 wP
References without fine-tuning								
Generator mean (no pixels)	0.587	0.786	–	0.616	0.472	0.486	–	0.548
CLIPScore, ViT-B/32	0.575	0.716	0.308	0.587	0.139	0.171	0.099	0.228
Frozen ridge probe, ViT-B/32	0.676	0.806	0.298	0.650	0.569	0.559	0.377	0.568
Frozen ridge probe, ViT-L/14	0.668	0.814	0.287	0.666	0.554	0.553	0.347	0.584
Fine-tuned CLIP, ViT-B/32 (three seeds)								
Base	0.672	0.817	0.314	0.663	0.556	0.546	0.375	0.556
+ML	0.685	0.827	0.328	0.672	0.588	0.585	0.394	0.598
Cond	0.719	0.846	0.433	0.690	0.628	0.623	0.459	0.600
Dec+ML	0.717	0.842	0.419	0.682	0.635	0.629	0.463	0.610
CPR+Dec	0.742	0.863	0.502	0.710	0.604	0.603	0.425	0.603
CPR+ML (shared)	0.748	0.865	0.509	0.714	0.628	0.627	0.450	0.615
CPR-IQA	0.747	0.864	0.510	0.709	0.627	0.624	0.453	0.615
Fine-tuned CLIP, ViT-L/14 (three seeds)								
Base	0.683	0.832	0.326	0.680	0.576	0.571	0.392	0.583
Cond	0.742	0.862	0.462	0.705	0.654	0.650	0.490	0.621
CPR-IQA	0.753	0.869	0.500	0.715	0.655	0.650	0.488	0.628

Table 7 Ablation under P1 (ViT-B/32, three seeds). CFA: random-prompt counterfactual accuracy of the alignment-type output; CFS_q: counterfactual shift of the quality output (MOS s.d.). Val: mean validation SROCC used for checkpoint selection.

Configuration	Prompt	ML	Dec	CPR	Quality	Auth.	Align. / corr.	CFA (%)	CFS_q	Val
AGIQA-3K										
Base	–	–	–	–	0.863 ± 0.001	–	0.672 ± 0.005	–	–	0.782
+ML	–	✓	–	–	0.867 ± 0.001	–	0.685 ± 0.011	–	–	0.790
Cond	✓	–	–	–	0.863 ± 0.002	–	0.719 ± 0.004	73.6	0.107	0.795
Dec+ML	✓	✓	✓	–	0.868 ± 0.004	–	0.717 ± 0.007	70.4	0.000	0.798
CPR+Dec	✓	–	✓	✓	0.864 ± 0.002	–	0.742 ± 0.004	93.3	0.000	0.802
CPR+ML (shared)	✓	✓	–	✓	0.868 ± 0.001	–	0.748 ± 0.002	93.3	0.113	0.810
CPR-IQA	✓	✓	✓	✓	0.868 ± 0.001	–	0.747 ± 0.005	93.4	0.000	0.811
AIGCIQA2023										
Base	–	–	–	–	0.805 ± 0.004	0.725 ± 0.004	0.556 ± 0.014	–	–	0.710
+ML	–	✓	–	–	0.816 ± 0.006	0.742 ± 0.001	0.588 ± 0.014	–	–	0.720
Cond	✓	–	–	–	0.807 ± 0.004	0.735 ± 0.008	0.628 ± 0.018	57.6	0.123	0.719
Dec+ML	✓	✓	✓	–	0.815 ± 0.006	0.745 ± 0.004	0.635 ± 0.008	56.7	0.000	0.737
CPR+Dec	✓	–	✓	✓	0.801 ± 0.001	0.727 ± 0.003	0.604 ± 0.006	98.3	0.000	0.706
CPR+ML (shared)	✓	✓	–	✓	0.815 ± 0.003	0.741 ± 0.005	0.628 ± 0.011	97.0	0.111	0.731
CPR-IQA	✓	✓	✓	✓	0.811 ± 0.006	0.741 ± 0.006	0.627 ± 0.009	93.2	0.000	0.733

that two of the displayed generations contain the model's own text-rendering artefacts (e.g. "CASTE" in Fig. 5a and a misspelled word in Fig. 5c); these are outputs of the generators rather than authorial errors and illustrate the text-utilisation issues discussed in this work.

Hard substitutions are less favourable to CPR-IQA. For nearest-neighbour substitutes CPR-IQA still prefers the true prompt in 84.8% of AGIQA-3K images (Cond: 56.8%), although frozen CLIPScore reaches 88.8%. For attribute and style swaps CPR-IQA reaches 62.3% and 66.4%, against 67.2% and 70.1% for CLIPScore. The gap widens on AIGCIQA2023: 66.7% (nearest neighbour) and 88.3% (same sub-category) for CPR-IQA against 97.2% and 99.3% for CLIPScore. The CPR variant without multi-level statistics is closer to CLIPScore here (81.9% and 95.3%), which suggests that the image-only statistics appended to the alignment input let the head rely less on the prompt. A learned alignment regressor thus fits MOS far better than CLIPScore (correspondence SROCC 0.627 against 0.139) but remains less sensitive to fine-grained prompt differences. CPR with random negatives teaches coarse matching, and closing this gap is left open (Section 7).

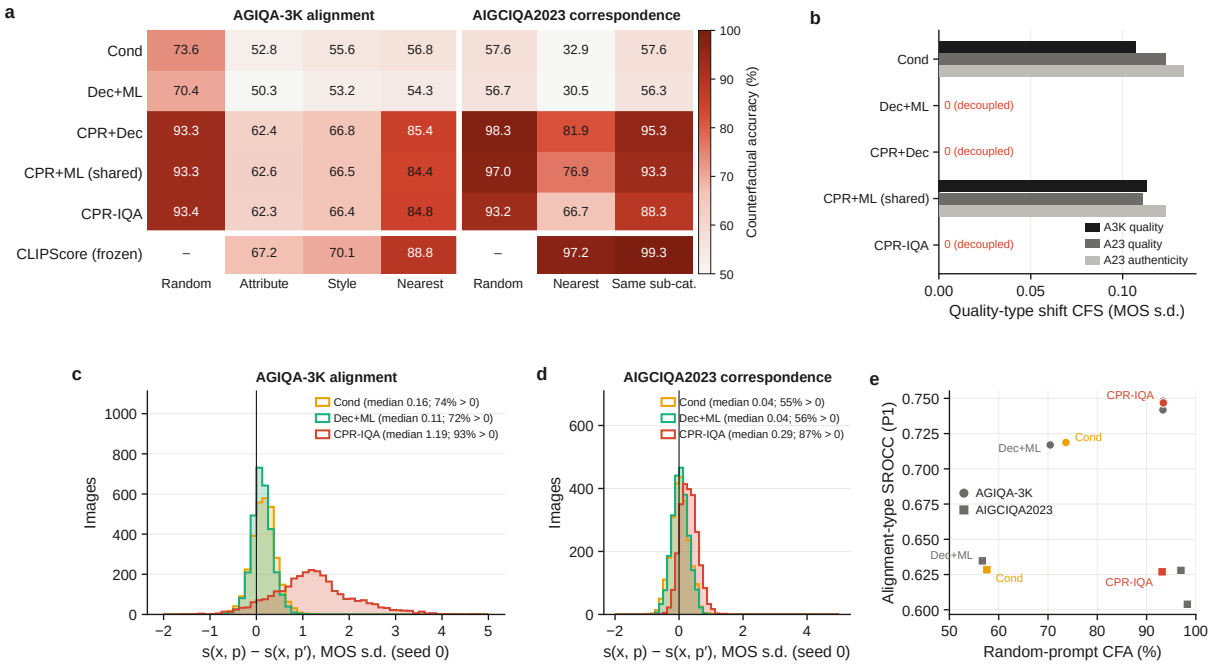

Fig. 4 Counterfactual behaviour (P1, ViT-B/32). (a) Counterfactual accuracy (%) of the alignment-type output under five substitution schemes, three seeds; CLIPScore (frozen ViT-B/32) is evaluated on the same substitutions (not defined for the random scheme). (b) Counterfactual shift of quality-type outputs under random substitution (MOS s.d.); decoupled heads have exactly zero shift. (c, d) Distribution over test images of $s(x, p) - s(x, p')$ for the alignment-type output under random substitution, in MOS standard deviations (seed 0, all five test folds). (e) Random-substitution CFA against alignment-type SROCC for the five prompt-conditioned configurations (circles, AGIQA-3K; squares, AIGCIQA2023).

6.6 Per-generator behaviour under P1

Fig. 6 breaks the P1 results of seed 0 down by generator. On AGIQA-3K, CPR-IQA has the highest within-generator alignment SROCC for all six generators, from +0.04 (SD 1.5: 0.472 → 0.513) to +0.11 (AttnGAN: 0.258 → 0.366) above Cond. On AIGCIQA2023 it is above Cond for all six generators by +0.006 to +0.054 and above Base by +0.05 to +0.17, but Dec+ML is slightly higher for DALL·E (0.380 against 0.354) and Lafite (0.483 against 0.478), in line with the equal pooled correspondence SROCC on this database.

Counterfactual accuracy depends strongly on the generator. The random-substitute test is decisive on four of the six AGIQA-3K generators: CPR-IQA prefers the true prompt for 99–100% of the DALL·E 2, Midjourney, SD 1.5 and SDXL images. It does so for 84% of the GLIDE images, but for only 55% of the AttnGAN images, the same level as Cond (55%). AttnGAN images have the lowest alignment MOS of the database (0.72 on the 0–5 scale against 1.23–3.07 for the other generators), so many of them depict too little of any prompt for a random substitute to fit worse. On AIGCIQA2023, CPR-IQA reaches 80.5–89.8% for every generator, whereas Cond and Dec+ML stay at 52.5–57.5%. A pooled CFA can therefore hide generators for which the alignment output is insensitive to the prompt, and per-generator CFA is the more informative report.

6.7 Double hold-out

Table 8 and Fig. 7 report P2, in which neither the test generator nor the test prompts are seen in training. All scores fall sharply, which confirms how much of the P1 performance rests on generator-level information. CPR-IQA improves alignment SROCC from 0.285 (Base) and 0.360 (Cond) to 0.449 on AGIQA-3K, and correspondence SROCC from 0.263 and 0.319 to 0.370 on AIGCIQA2023. It is better than Base on alignment-type SROCC for all six AGIQA-3K generators and all six AIGCIQA2023 generators, and better than Cond for

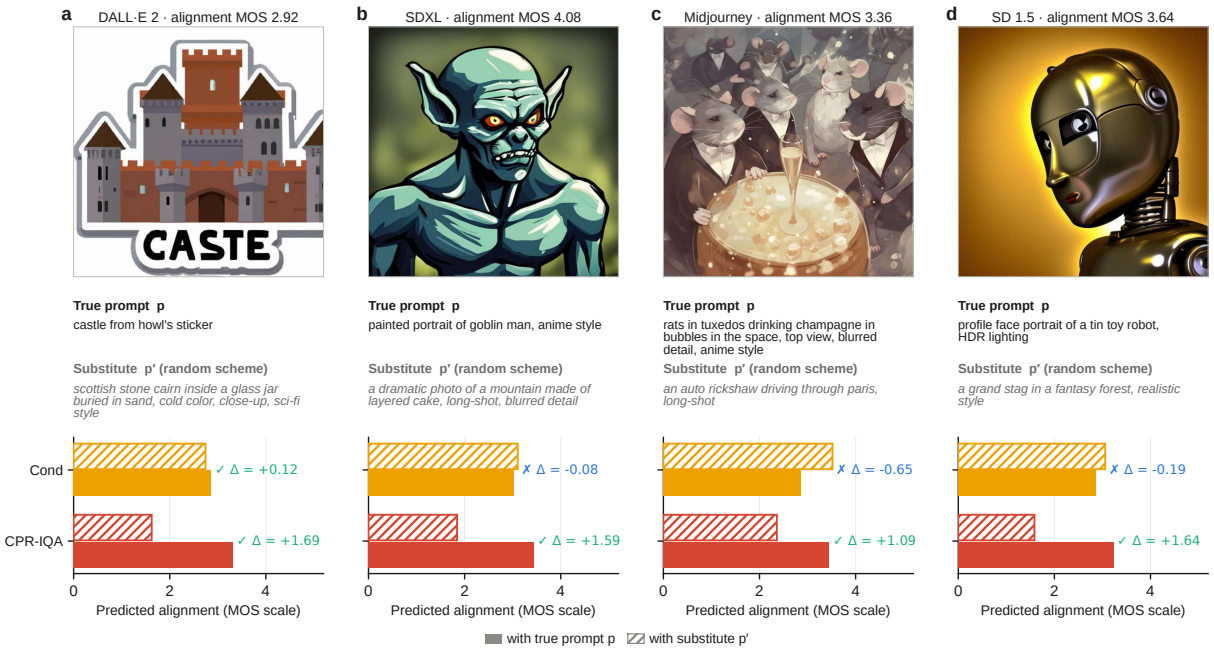

Fig. 5 Counterfactual examples from AGIQA-3K (P1 test folds, seed 0). For each image: generator and alignment MOS (0–5 scale), the true prompt p, a random substitute p′ from the counterfactual test, and the alignment score predicted by Cond and CPR-IQA with p (filled) and p′ (hatched). ✓: the true prompt receives the higher score; ×: the substitute does.

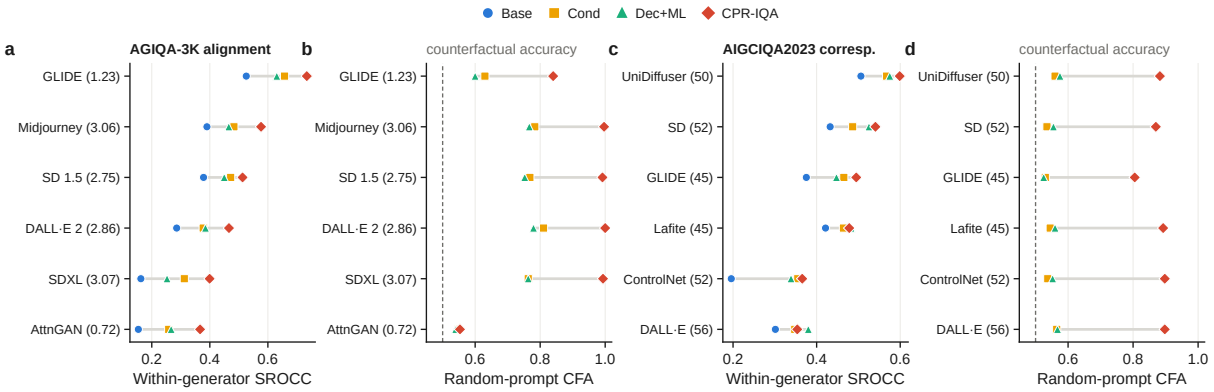

Fig. 6 Per-generator behaviour under P1 (seed 0, five prompt-disjoint folds). (a, c) Within-generator SROCC of the alignment-type output for each generator, mean over folds; mean alignment-type MOS of the generator in parentheses. (b, d) Random-prompt counterfactual accuracy of the same output; the dashed line marks chance. Base has no prompt input and therefore no counterfactual accuracy. Generators are ordered by the within-generator SROCC of CPR-IQA.

six of six and five of six generators, respectively. It keeps a CFA of 88.6% and 90.5% on unseen prompts and generators. The alignment-type gain over Base is largest for SDXL on AGIQA-3K (0.129 → 0.343) and Lafite on AIGCIQA2023 (0.200 → 0.349), smallest for DALL·E on AIGCIQA2023 (0.213 → 0.228), and remains positive for the hardest generators (AttnGAN, ControlNet; Fig. 7).

For quality, the gains are small (0.587 → 0.610 on AGIQA-3K; 0.358 → 0.374 on AIGCIQA2023), and on AIGCIQA2023 quality prediction for unseen generators remains poor for all models (Lafite: 0.02–0.07). The frozen references separate

the two kinds of dimension. For alignment-type dimensions CPR-IQA is well ahead of CLIPScore and the frozen probes (AGIQA-3K alignment: 0.449 against 0.307 for CLIPScore ViT-B/32 and 0.287 for the ViT-L/14 probe; AIGCIQA2023 correspondence: 0.370 against 0.302 for the best probe). For quality-type dimensions, a frozen ViT-L/14 ridge probe comes close to the fine-tuned models on unseen generators and exceeds them on authenticity (AGIQA-3K quality 0.584; AIGCIQA2023 quality 0.368 and authenticity 0.463, against 0.610, 0.374 and 0.434 for CPR-IQA). End-to-end fine-tuning on a few generators therefore buys little quality generalisation to new generators.

6.8 Negative and exploratory results

Several further ideas did not help. In exploratory single-seed runs with a 12-epoch budget, evaluated on test folds and reported only for transparency, within-generator and within-prompt ranking losses and adversarial removal of generator information changed AGIQA-3K quality SROCC by -0.006 to $+0.002$ (Base 0.863), so none reduced the generator shortcut. Nearest-neighbour negatives in CPR lowered alignment SROCC to 0.696 against 0.737 with random negatives, and a 50/50 mix gave 0.721. A graded margin, margins of 0.5 or 2, a CPR weight of 2 and an explicit prompt-invariance penalty on the shared quality output all stayed within 0.73–0.74. Token-level cross-attention did not improve validation scores over the global fusion.

7 Discussion and limitations

Pooled SROCC hides all three shortcuts, and random splits add prompt leakage on top. On AIGCIQA2023, the configuration with the best validation SROCC barely uses the prompt, and the prompt-conditioned baseline has the same correspondence SROCC as CPR-IQA while being far less sensitive to prompt substitutions. Counterfactual accuracy and shift, within-generator SROCC and held-out generators expose differences that the usual metrics do not. The per-generator analysis adds a caveat to the counterfactual test itself: for a generator whose images rarely depict their prompt (AttnGAN), a random substitute fits about as poorly as the true prompt, and no model separates the two. Counterfactual accuracy should therefore be reported per generator, or at least checked for generators with very low alignment MOS.

For practitioners, the results suggest a reporting checklist for AGIQA models: (i) prompt-disjoint rather than random splits; (ii) within-generator and within-prompt SROCC next to the pooled value, compared with the generator-mean predictor; (iii) counterfactual accuracy for alignment outputs and counterfactual shift for quality outputs, per generator; and (iv) at least one held-out-generator evaluation. Items (ii) and (iii) are computed from the predictions of a standard evaluation, with one extra forward pass of the alignment head per substitute prompt for (iii); items (i) and (iv) change only how the data are split. The same diagnostics apply to the newer prompt-aware AGIQA models discussed in Section 2.1 [32, 33], whose alignment outputs could be checked for prompt sensitivity in the same way.

The study has three groups of limitations.

1. **Scope and generalisation.** Only two medium-sized databases are used, AGIQA-3K and AIGCIQA2023; AIGIQA-20K [4] and the newer databases [5, 18, 19] are not included, cross-database transfer between the two was not tested, and only OpenAI CLIP ViT-B/32 and ViT-L/14 at 224 pixels were fine-tuned, where short-side resizing with a centre crop can hide boundary and fine-detail artefacts. Within this scope the generator shortcut is not solved: quality prediction for unseen generators remains poor, loss-based remedies did not help, data with more within-generator and within-prompt variation is likely to be more effective than further loss design, and label-free adaptation to evolving generators [36] was not evaluated.
2. **Missing comparisons and the counterfactual test.** Preference-trained reward models and prompt-conditioned AGIQA architectures could not be run, so CPR was not tested as a plug-in to those architectures. The re-trained NR-IQA models used a 12-epoch budget and one seed, so their comparison with the 25-epoch, three-seed CLIP configurations is not budget-matched. Hard substitutions have no human ratings, so CFA is interpreted relative to CLIPScore: CPR-IQA remains below CLIPScore under nearest-neighbour, attribute and style substitutions, and fine-grained compositional

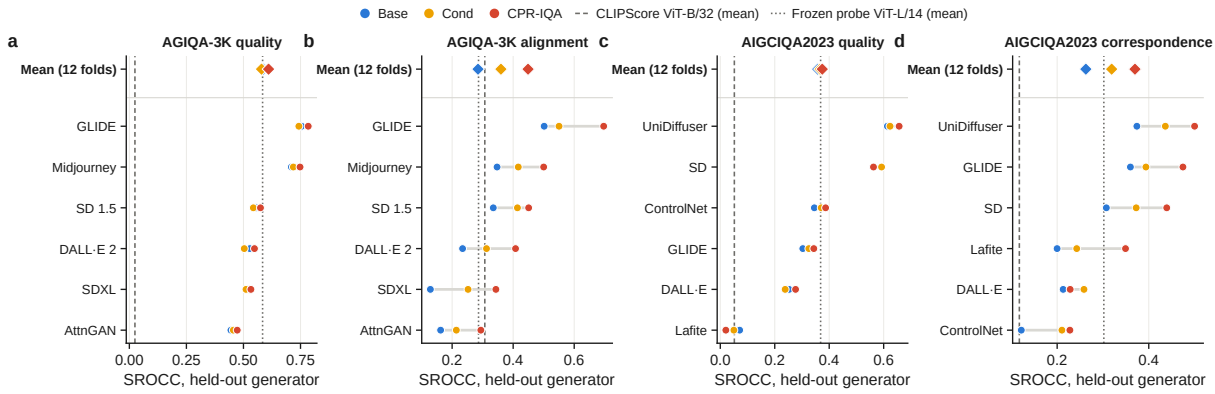

Fig. 7 Generator × prompt double hold-out (P2). SROCC for each held-out generator (mean over two prompt halves and three seeds) for Base, Cond and CPR-IQA, and the mean over the 12 folds (diamonds). Vertical lines: 12-fold mean of CLIPScore ViT-B/32 (dashed) and of the frozen ViT-L/14 ridge probe (dotted). Generators are ordered by the SROCC of CPR-IQA.

Table 8 Generator × prompt double hold-out (P2): SROCC per held-out generator (mean over two prompt halves and three seeds) and mean over the 12 folds.

Method	A3K quality	A3K alignment	A23 quality	A23 authenticity	A23 corresp.	CFA A3K / A23 (%)
CLIP-IQA zero-shot, ViT-L/14	0.328	0.189	0.281	0.240	0.226	–
CLIPScore, ViT-B/32	0.024	0.307	0.052	0.032	0.117	–
Frozen ridge probe, ViT-B/32	0.569	0.274	0.332	0.424	0.261	–
Frozen ridge probe, ViT-L/14	0.584	0.287	0.368	0.463	0.302	–
Base	0.587 ± 0.009	0.285 ± 0.008	0.358 ± 0.016	0.431 ± 0.015	0.263 ± 0.025	– / –
Cond	0.579 ± 0.009	0.360 ± 0.008	0.367 ± 0.006	0.422 ± 0.024	0.319 ± 0.038	68.7 / 56.0
CPR-IQA	0.610 ± 0.002	0.449 ± 0.014	0.374 ± 0.014	0.434 ± 0.018	0.370 ± 0.036	88.6 / 90.5

Table 8b. Per-generator SROCC under P2 (mean over two prompt halves and three seeds).

Held-out generator	Quality: Base / Cond / CPR-IQA	Alignment-type: Base / Cond / CPR-IQA
AGIQA-3K: AttnGAN	0.445 / 0.456 / 0.473	0.162 / 0.214 / 0.294
AGIQA-3K: DALL·E 2	0.528 / 0.504 / 0.548	0.234 / 0.313 / 0.408
AGIQA-3K: GLIDE	0.755 / 0.743 / 0.784	0.502 / 0.550 / 0.697
AGIQA-3K: Midjourney	0.710 / 0.719 / 0.748	0.347 / 0.417 / 0.500
AGIQA-3K: SD 1.5	0.556 / 0.543 / 0.575	0.335 / 0.414 / 0.451
AGIQA-3K: SDXL	0.527 / 0.511 / 0.533	0.129 / 0.252 / 0.343
AIGCIQA2023: ControlNet	0.345 / 0.371 / 0.387	0.122 / 0.210 / 0.228
AIGCIQA2023: DALL·E	0.253 / 0.238 / 0.276	0.213 / 0.259 / 0.228
AIGCIQA2023: GLIDE	0.303 / 0.325 / 0.343	0.360 / 0.394 / 0.475
AIGCIQA2023: Lafite	0.071 / 0.050 / 0.020	0.200 / 0.243 / 0.349
AIGCIQA2023: UniDiffuser	0.614 / 0.623 / 0.657	0.374 / 0.436 / 0.500
AIGCIQA2023: Stable Diffusion	0.563 / 0.592 / 0.563	0.307 / 0.373 / 0.439

sensitivity (counting, relations, attribute binding [17, 46]) remains open. Nearest-neighbour and same-sub-category substitutes are drawn from the whole database and can therefore be training prompts.

3. **Assumptions, protocol dependence and exploratory runs.** Treating quality and authenticity as prompt-independent is a modelling assumption; if raters saw the prompt, quality MOS may carry prompt information that a decoupled head cannot use, and on AIGCIQA2023 quality the shared head was slightly better. On that database CPR does not change correspondence SROCC under P1 while greatly changing CFA, so its value is visible mainly under P2 and in the counterfactual diagnostics. Early exploratory runs (Section 6.8) were evaluated on test folds; they are reported for transparency, and the final configuration was fixed before the three-seed runs reported in Tables 4–8.

8 Conclusion

On the two databases we study, AGIQA benchmarks built from a generator × prompt design reward three shortcuts. Generator identity explains more than half of the quality variance, random splits leak prompts into alignment scores, and prompt-conditioned regressors use the prompt weakly while leaking it into quality outputs. Prompt-disjoint evaluation removes the second, and CPR-IQA addresses the third with a counterfactual prompt ranking constraint and a prompt-independent quality branch. Under prompt-disjoint cross-validation it improves fold-mean alignment SROCC over the image-only CLIP baseline by 0.075 on AGIQA-3K and correspondence SROCC by 0.071 on AIGCIQA2023, and under a double hold-out of generators and prompts it improves alignment-type SROCC from 0.285 to 0.449 and from 0.263 to 0.370. Over the prompt-conditioned baseline the P1 gain is +0.028 on AGIQA-3K and absent on AIGCIQA2023, where CPR changes counterfactual behaviour rather than SROCC. With ViT-B/32 its alignment scores respond to random prompt substitutions for 93% of images (91.9% and 77.3% with ViT-L/14), and its quality scores do not respond at all. It remains less sensitive than zero-shot CLIPScore to near-duplicate prompts, and quality prediction for unseen generators remains unsolved. We recommend that AGIQA studies report prompt-disjoint results, within-generator SROCC with the η^2 of the benchmark, results on unseen generators and prompts, and counterfactual prompt accuracy and shift with a zero-shot reference. Split indices, code and per-fold predictions will be released on publication (see Data availability).

Statements and Declarations

Funding. This work was supported by the Guangxi Minzu University Research Project (No.2022KJQD020), the Key Research and Development Program of Guangxi (No.Guike AB23075120, No.Guike AB24010044), Natural Science Foundation of Guangxi Youth Science Fund Project (No.2025GXNSFBA069234).

Competing interests. The authors declare that they have no competing interests.

Ethics approval and consent to participate. Not applicable. This study uses only publicly available image databases and their published subjective scores; no new human or animal experiments were conducted.

Consent for publication. Not applicable.

Data availability. AGIQA-3K and AIGCIQA2023 are publicly available; the copies used here were obtained from the Hugging Face mirrors strawhat/agiqa-3k and Int-MeGroup/AIGCIQA2023, and one AGIQA-3K image was missing from the former (midjourney_normal_294.jpg). Fold definitions and per-fold predictions will be made available on publication.

Code availability. The code will be made available on publication.

Author contributions. **Shengyu Pei:** Conceptualization, Methodology, Software, Writing – original draft, Supervision. **Yihua Wu:** Investigation, Validation, Data curation. **Yule An:** Validation, Data curation. **Sisi Fan:** Formal analysis, Writing – review & editing. **Huan Lao:** Visualization, Investigation.

References

[1] Zhang, Z., Li, C., Sun, W., Liu, X., Min, X., Zhai, G.: A perceptual quality assessment exploration for AIGC images. In: Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW), pp. 440–445 (2023). https://doi.org/10.1109/ICMEW59549.2023.00082

[2] Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., Zhai, G., Lin, W.: AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. IEEE Trans. Circuits Syst. Video Technol. **34**, 6833–6846 (2024) https://doi.org/10.1109/tcsvt.2023.3319020

[3] Wang, J., Duan, H., Liu, J., Chen, S., Min, X., Zhai, G.: AIGCIQA2023: A large-scale image quality assessment database for AI generated images: from the perspectives of quality, authenticity and correspondence. In: Artificial Intelligence (CICAI 2023). Lect. Notes Comput. Sci., vol. 14474, pp. 46–57. Springer,

Singapore (2024). https://doi.org/10.1007/978-981-99-9119-8_5

[4] Li, C., Kou, T., Gao, Y., Cao, Y., Sun, W., Zhang, Z., Zhou, Y., Zhang, Z., Zhang, W., Wu, H., *et al.*: AIGIQA-20K: A large database for AI-generated image quality assessment. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp. 6327–6336 (2024). https://doi.org/10.1109/CVPRW63382.2024.00636

[5] Chen, J., Shao, F., Chen, H., Wang, X., Guo, H., Jiang, Q.: Quality Evaluation of AI-Generated Images: Subjective Study and Objective Methodology. IEEE Trans. Multimedia **28**, 57–70 (2026) https://doi.org/10.1109/tmm.2025.3618562

[6] Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Trans. Image Process. **29**, 4041–4056 (2020) https://doi.org/10.1109/TIP.2020.2967829

[7] Chen, Z., Wang, J., Wang, W., Xu, S., Guo, J., Wang, S., Yuan, C., Li, B., Hu, W.: Towards Region-Level No-Reference Image Quality Assessment. IEEE Trans. Image Process. (2026) https://doi.org/10.1109/tip.2026.3734467 . Early access

[8] Yun, Y.K., Lin, W.: You Only Train Once: A Unified Framework for Both Full-Reference and No-Reference Image Quality Assessment. IEEE Trans. Image Process. **35**, 4467–4479 (2026) https://doi.org/10.1109/tip.2026.3661408

[9] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: Proc. 38th Int. Conf. Mach. Learn. (ICML). Proc. Mach. Learn. Res., vol. 139, pp. 8748–8763 (2021)

[10] Wang, J., Chan, K.C.K., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: Proc. AAAI Conf. Artif. Intell., vol. 37, pp. 2555–2563 (2023). https://doi.org/10.1609/aaai.v37i2.25353

[11] Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 14071–14081 (2023). https://doi.org/10.1109/CVPR52729.2023.01352

[12] Qu, B., Li, H., Gao, W.: Bringing textual prompt to AI-generated image quality assessment. In: Proc. IEEE Int. Conf. Multimedia Expo (ICME), pp. 1–6 (2024). https://doi.org/10.1109/ICME57554.2024.10688254

[13] Peng, F., Fu, H., Ming, A., Wang, C., Ma, H., He, S., Dou, Z., Chen, S.: AIGC image quality assessment via image-prompt correspondence. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp. 6432–6441 (2024). https://doi.org/10.1109/CVPRW63382.2024.00644

[14] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nat. Mach. Intell. **2**(11), 665–673 (2020) https://doi.org/10.1038/s42256-020-00257-z

[15] Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: Improving visual-semantic embeddings with hard negatives. In: Proc. Br. Mach. Vis. Conf. (BMVC) (2018)

[16] Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, pp. 9694–9705 (2021)

[17] Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: Int. Conf. Learn. Represent. (ICLR) (2023)

[18] Wang, J., Duan, H., Zhai, G., Min, X.: Quality Assessment for AI Generated Images With

Instruction Tuning. IEEE Trans. Multimedia **28**, 2404–2419 (2026) https://doi.org/10.1109/tmm.2026.3651009

[19] Li, Y., Wu, S., Sun, W., Zhang, Z., Zhu, Y., Zhang, Z., Duan, H., Min, X., Zhai, G.: AGHI-QA: A Subjective-Aligned Dataset and Metric for AI-Generated Human Images. IEEE Trans. Circuits Syst. Video Technol. **36**, 4636–4649 (2026) https://doi.org/10.1109/tcsvt.2025.3639620

[20] Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. **21**(12), 4695–4708 (2012) https://doi.org/10.1109/TIP.2012.2214050

[21] Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1733–1740 (2014). https://doi.org/10.1109/CVPR.2014.224

[22] Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans. Circuits Syst. Video Technol. **30**(1), 36–47 (2020) https://doi.org/10.1109/TCSVT.2018.2886771

[23] Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3664–3673 (2020). https://doi.org/10.1109/CVPR42600.2020.00372

[24] Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale image quality transformer. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 5128–5137 (2021). https://doi.org/10.1109/ICCV48922.2021.00510

[25] Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: MANIQA: Multi-dimension attention network for no-reference image quality assessment. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp. 1190–1199 (2022). https://doi.org/10.1109/CVPRW56347.2022.00126

[26] Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.: TOPIQ: A top-down approach from semantics to distortions for image quality assessment. IEEE Trans. Image Process. **33**, 2404–2418 (2024) https://doi.org/10.1109/TIP.2024.3378466

[27] Sang, Q., Li, Q., Liu, L., Deng, Z., Wu, X., Bovik, A.C.: No-Reference Image Quality Assessment Leveraging GenAI Images. IEEE Trans. Image Process. **34**, 6204–6214 (2025) https://doi.org/10.1109/tip.2025.3610238

[28] Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., Yan, Q., Min, X., Zhai, G., Lin, W.: Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In: Proc. 41st Int. Conf. Mach. Learn. (ICML). Proc. Mach. Learn. Res., vol. 235, pp. 54015–54029 (2024)

[29] Zhu, H., Sui, X., Chen, B., Liu, X., Chen, P., Fang, Y., Wang, S.: 2AFC Prompting of Large Multimodal Models for Image Quality Assessment. IEEE Trans. Circuits Syst. Video Technol. **34**, 12873–12878 (2024) https://doi.org/10.1109/tcsvt.2024.3434999

[30] You, Z., Gu, J., Cai, X., Li, Z., Zhu, K., Dong, C., Xue, T.: Enhancing Descriptive Image Quality Assessment With a Large-Scale Multi-Modal Dataset. IEEE Trans. Image Process. **34**, 8201–8215 (2025) https://doi.org/10.1109/tip.2025.3639998

[31] Wang, P., Sun, W., Zhang, Z., Jia, J., Jiang, Y., Zhang, Z., Min, X., Zhai, G.: Large multi-modality model assisted AI-generated image quality assessment. In: Proc. 32nd ACM Int. Conf. Multimedia (MM), pp. 7803–7812 (2024). https://doi.org/10.1145/3664647.3681471

[32] Xia, J., He, L., Deng, C., Li, L., Gao, X.: AI-Generated Image Quality Assessment Based on Task-Specific Prompt and Multi-Granularity Similarity. IEEE Trans. Image

Process. **34**, 8216–8228 (2025) https://doi.org/10.1109/tip.2025.3639984

[33] Sun, W., Chen, C., Liao, L., Lin, W.: Prompt-Image-Caption Consistency for AI-Generated Image Quality Assessment. IEEE Trans. Multimedia **28**, 5592–5602 (2026) https://doi.org/10.1109/tmm.2026.3668530

[34] Zhou, T., Tan, S., Li, L., Zhao, B., Jiang, Q., Yue, G.: Cross-Modality Interactive Attention Network for AI-generated image quality assessment. Pattern Recognit. **167**, 111693 (2025) https://doi.org/10.1016/j.patcog.2025.111693

[35] Yang, Y., Zhang, Y., Yang, X., Zhang, W.: PeC-AGIQA: Perceptual Feature-guided Contrastive Learning for AI-Generated Image Quality Assessment. IEEE Trans. Circuits Syst. Video Technol. (2026) https://doi.org/10.1109/tcsvt.2026.3718652 . Early access

[36] Li, X., Xu, Z., Tang, M., Cai, Z., Wu, S., Min, X., Chen, Y., Zhai, G.: ELIQ: A Label-Free Framework for Quality Assessment of Evolving AI-Generated Images. IEEE Trans. Circuits Syst. Video Technol. (2026) https://doi.org/10.1109/tcsvt.2026.3734768 . Early access

[37] Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Proc. 2021 Conf. Empir. Methods Nat. Lang. Process. (EMNLP), pp. 7514–7528 (2021). https://doi.org/10.18653/v1/2021.emnlp-main.595

[38] Tan, Z., Yang, X., Ye, Z., Wang, Q., Yan, Y., Nguyen, A., Huang, K.: Semantic Similarity Distance: Towards better text-image consistency metric in text-to-image generation. Pattern Recognit. **144**, 109883 (2023) https://doi.org/10.1016/j.patcog.2023.109883

[39] Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: ImageReward: Learning and evaluating human preferences for text-to-image generation. In: Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 36, pp. 15903–15935 (2023)

[40] Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In: Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 36, pp. 36652–36663 (2023)

[41] Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)

[42] Tian, Y., Li, Z.: Recent advances in human preference alignment of text-to-image diffusion models. Neurocomputing **700**, 134466 (2026) https://doi.org/10.1016/j.neucom.2026.134466

[43] Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 20349–20360 (2023). https://doi.org/10.1109/ICCV51070.2023.01866

[44] Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: Computer Vision – ECCV 2024. Lect. Notes Comput. Sci., vol. 15067, pp. 366–384. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72673-6_20

[45] Zhou, Q., Zhang, Z., Wang, J., Li, X., Zhang, J., Yu, Q., Sheng, L., Xu, D.: ELBO-T2IAlign: A Generic ELBO-Based Method for Calibrating Pixel-Level Text-Image Alignment in Diffusion Models. IEEE Trans. Image Process. **35**, 7319–7331 (2026) https://doi.org/10.1109/tip.2026.3709461

[46] Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 5228–5238 (2022). https://doi.org/10.1109/CVPR52688.2022.00517

[47] Chen, J., Zeng, H., Huang, C., Huang, Q.,

Li, Z., Huang, X.: Context-aware latent space mediation for inference-time unbiased semantic alignment in text-to-image models. Expert Syst. Appl. **330**, 133019 (2026) https://doi.org/10.1016/j.eswa.2026.133019

[48] Kaushik, D., Hovy, E., Lipton, Z.C.: Learning the difference that makes a difference with counterfactually-augmented data. In: Int. Conf. Learn. Represent. (ICLR) (2020)

[49] Chen, C., Mo, J.: IQA-PyTorch: PyTorch toolbox for image quality assessment. GitHub repository. https://github.com/chaofengc/IQA-PyTorch (accessed 30 September 2026) (2022)

[50] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Int. Conf. Learn. Represent. (ICLR) (2019)

[51] Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993). https://doi.org/10.1007/978-1-4899-4541-9

[52] Nadeau, C., Bengio, Y.: Inference for the generalization error. Mach. Learn. **52**(3), 239–281 (2003) https://doi.org/10.1023/A:1024068626366

[53] Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**(2), 65–70 (1979)