OF CHAINS AND GAINS: WHAT REINFORCEMENT LEARNING CHANGES ON REASONING-BASED IMAGE QUALITY ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) has improved the performance of large Vision-Language Models (VLMs) on image quality assessment (IQA). Existing approaches often attribute these gains to improved reasoning before scoring, supported by increasingly plausible Chain-of-Thought (CoT). However, whether these gains exceed sampling effects and how they depend on explicit reasoning remain unclear. We rigorously investigate these questions through sampling-aware evaluation, controlled reasoning interventions, and shortcut experiments. We find that RL-trained models outperform base models under large-budget $\max_g @k$ evaluation, while improvements beyond averaging are concentrated in-distribution. RLVR produces chains that raise base-model performance when transplanted and improves scoring with the chain held fixed. Yet scoring becomes less dependent on the chain during inference, more clearly at 7B than at 32B. Steering the probed quality direction in thinking tokens also has little effect on scoring. Finally, RL can improve scoring accuracy by exploiting textual cues correlated with mean opinion scores (MOSs), which could inflate benchmark performance. Together, these findings show that RLVR raises quality scoring and produces more useful chains, but that better scoring does not imply a stronger reliance on explicit reasoning, while score-based rewards can also encourage reliance on quality-irrelevant shortcuts. We hope this preliminary study can enlighten more and more concrete reasoning schemes in IQA.

1 INTRODUCTION

Image quality assessment (IQA) predicts the perceptual quality of images as judged by human observers (Wang & Bovik, 2006). Large vision-language models (VLMs) have been widely adapted to this task, initially by supervised fine-tuning (You et al., 2024; Wu et al., 2024c; Zhu et al., 2024) and recently by Reinforcement Learning with Verifiable Rewards (RLVR) (Li et al., 2025; Wu et al., 2025; Jiang et al., 2026). RLVR-based IQA methods elicit VLMs to reason with structured chain-of-thoughts (CoTs) (Wei et al., 2022) before scoring[1], and reward the agreement between predictions and mean opinion scores (MOSs) automatically. They achieved state-of-the-art quality prediction and generalisation without literally supervising the reasoning process. It is widely believed that RLVR can enable VLMs to autonomously incentivize in-depth image quality reasoning (Li et al., 2025), such as better awareness of distortion types, degradation levels, and corruption localization, which are evidenced by plausible and comprehensive CoT explanations absent in base models (Guo et al., 2025). Consequently, RLVR is viewed as a route for evolving VLMs for better IQA capabilities (Cai et al., 2025; Jia et al., 2026), potentially advancing closer towards human visual perception.

Despite these empirical gains, RLVR on IQA remains underexamined. The quality reward verifies only the score, so the training objective does not determine whether an improvement of the score is obtained through the chain. This raises a fundamental question:

Does RLVR genuinely improve the quality reasoning of VLMs, with better chains leading to better scores, or can it improve the performance bypass the chain?

[1]We restrict our scope to explicit reasoning which forces models to output readable reasoning.

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

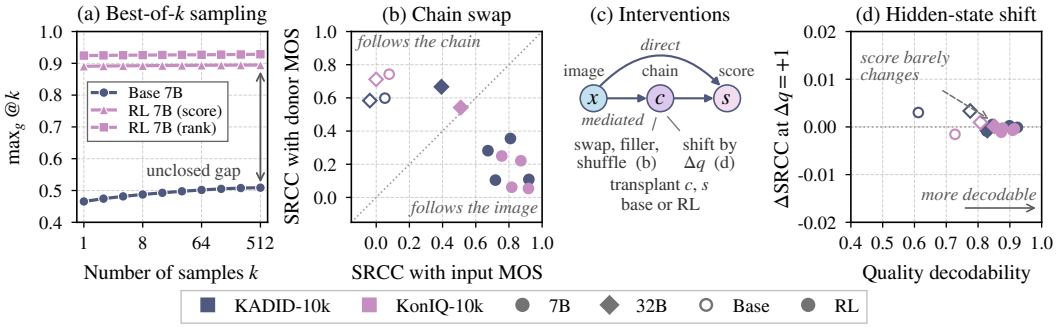

Figure 1: Gains and chains under RLVR for IQA. (a) $\max_g @k$ SRCC on KADID-10k. (b) Scores under a swapped chain, SRCC with the input image against the donor image. (c) The mediated path through the chain, the direct path, and where the interventions act. (d) Quality decodability of the chain against the change in SRCC when its hidden state is steered.

Previous belief rests on two kinds of evidence, higher rank correlation and more plausible chains, while neither establishes that the improved performance is causally attributes to reasoning. RLVR is known to sharpen the output distribution of the base model, which is how it raises single-sample accuracy in mathematics and programming (Yue et al., 2025). RL-trained IQA models do produce more concise chains with less unrelated content (Zhao et al., 2026), which might be what a sharpened output distribution looks like in the text. On the other hand, plausible chains do not causally show that the score is derived from them, since IQA relies densely on vision (Bosse et al., 2018), and language chains can fail to reflect the actual cause of the answer even after outcome-based RL (Turpin et al., 2023; Lanham et al., 2023; Chen et al., 2025; Yu et al., 2026). Since quality reward only verifies agreement with MOS, any correlate of MOS in the training images raises it as readily as quality itself (Skalse et al., 2022; Geirhos et al., 2020; Shao et al., 2025), so higher accuracy with plausible chains cannot tell better reasoning apart from reduced sampling randomness, a path that bypasses the chain, or learned underlying shortcuts.

To investigate this question, we design a set of controlled experiments that address the three alternatives in turn, summarized in Figure 1. To separate the gain from sampling randomness, we note that one prediction per image conflates what a model can assess with how reliably it samples, and therefore adopt $\max_g @k$ (Walder & Karkhanis, 2025), the expected best Spearman rank correlation coefficient (SRCC) among k complete test-set predictions sampled from a model, asking whether a base model reaches its RL-trained counterpart under a larger sampling budget (Figure 1a). If the gain survives this comparison, the next question is whether it is obtained through the chain. Since the image can influence the score through the chain or directly, and a chain that is useful for scoring need not be one that the score depends on, we fix the image and intervene on the chain in three ways (Figure 1c). Swapping in the chain of another image or removing its content reveals whether the score depends on the chain (Figure 1b). Transplanting chains between base and RL-trained models then determines whether the chain nevertheless carries the gain, and shifting the hidden representation of the chain along the direction that encodes quality probes whether the information encoded there is used (Figure 1d). Finally, we train models with textual cues whose correlation with MOS is controlled, which examines whether accuracy can rise without a better assessment of the image. We discovered several perhaps surprising findings that may help to better understand VLM's reasoning in IQA, as follows:

- **RLVR improves quality prediction beyond sampling effects, and the part beyond averaging is concentrated in-distribution.** Under $\max_g @k$, a base model with a large sampling k remains below its RL-trained counterpart, and in-distribution the RL-trained model stays well above the base model even after the sampling noise of the base model is averaged out, whereas out-of-distribution this margin can vanish. This suggests that unlike mathematics or programming tasks, RL *does* improve VLM's image quality scoring capability, but it is confined to in-distribution test sets, whereas out-of-distribution RLVR makes base models more reliable in a single sampling.
- **RLVR produces more useful chains on IQA, yet higher performance comes with a weaker dependence on them during inference.** Through chain intervention experiments on both text and hidden state, we discovered RLVR improves both the usefulness of the chain and the scoring

model itself. However, this joint improvement does not come with a stronger dependence on the chain. The 7B models show less dependence than the 32B model.

- **RLVR can learn shortcuts unrelated to visual quality.** We discover that when a textual cue is related to MOS during training, RL models can gain improvements via predicting this shortcut, even when the cue carries no quality semantics, while models trained with the same text uncorrelated with MOS are unaffected. This indicates that the reward can be easily met by correlate of MOS, so improvements after RLVR may partly reflect such correlates in the data.

In conclusion, we discovered RLVR can actually improve VLM's capability on image quality scoring and produces more useful chains, yet it can indicate a weaker reliance on explicit reasoning during inference than base models. The gain is not carried by the chain alone but also by an improved mapping from image and chain to score. This highlights a gap between using RLVR to achieve accurate and explainable IQA, underscoring more faithful image quality reasoning.

2 Related Works

CoT faithfulness and mechanistic interpretability. A faithful chain of thought (CoT) reflects model's decision process rather than only being plausible (Jacovi & Goldberg, 2020). CoT (Wei et al., 2022) often falls short of it, since models can omit what drives their answers (Turpin et al., 2023), the answer often survives a truncated chain (Lanham et al., 2023; Paul et al., 2024), and it follows a corrupted step less after post-training for reasoning (Jia & Mu, 2026). Post-training weakens the link between chain and answer (Bao et al., 2025), and outcome-based RL raises accuracy without making the chain faithful (Chen et al., 2025) or causally important (Yu et al., 2026; Dutta et al., 2024; Li et al., 2023; Ju et al., 2025; Zhang et al., 2025). These works address tasks in which the chain is essential and correctness is exact. We bring this perspective to the IQA task, where the answer rests on dense visual evidence and has no exact correctness. We further contrast base and RL-trained models to measure how RLVR changes the dependence of the score on the chain.

RLVR for IQA. VLM-based IQA progressed from benchmarking (Wu et al., 2024a) to supervised fine-tuning for opened-question (Wu et al., 2024b; You et al., 2024) and scoring (Wu et al., 2024c; Zhu et al., 2024; You et al., 2025). Recently, reasoning-based methods elicit models to reason before scoring (Li et al., 2025; Wu et al., 2025; Cai et al., 2025; Jia et al., 2026) and obtained significant improvements, which are widely attributed to learned quality reasoning. Zhao et al. (2026) trace the generalization of RL models to the representation formed by textual reasoning, and Li & Nishida (2026) discover that quality derived from the text alone fall well below those from the image. These indicate that the reasoning text can suffice to support the score under essential conditions, indicating the sufficiency side of faithfulness (Lyu et al., 2023), but not that the score must pass through it, *i.e.*, the necessity. It still unclear whether quality scores depend on chains, whether the gains persist at large sampling budgets, or whether the reward is met through shortcuts (Geirhos et al., 2020).

3 Preliminaries

Reinforcement learning with verifiable rewards. Let π_θ be a VLM that, given an image x and a scoring prompt, generates a response $y = (c, s)$, a reasoning chain c followed by a quality score s. The verifier is the MOS m of x. The *score-based* reward (Li et al., 2025) is granted when s falls within a tolerance of m, and the *rank-based* reward (Wu et al., 2025) measures how well the preference between two images agrees with the order of MOSs. Refer to Appendix A for reward details. Training maximizes $J(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\,\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}[r]$ with GRPO (Shao et al., 2024). In this study, the reasoning content is never supervised explicitly to better isolate the effects.

Reasoning paradigms for IQA. RLVR-based IQA methods adopt a *think-then-score* paradigm, where the model generates a reasoning chain before the quality score. Omitting the prompt, the response distribution factorizes as $\pi_\theta(c, s \mid x) = \pi_\theta(c \mid x)\,\pi_\theta(s \mid x, c)$. Therefore, the image could influence the score along either a mediated path $x \to c \to s$, a direct path $x \to s$ that bypasses it, or both simultaneously. To separate them, we use *intervened scoring* to manage on $\tilde{c}$ and see its influence on s, such as swapping and steering (as detailed in Section 5).

Models and benchmarks. We use the generally employed Qwen2.5-VL-Instruct (Bai et al., 2025) at 7B and 32B as base models. For each recipe we train models with the official implementation on synthetic-distorted KADID-10k (Lin et al., 2019), and on authentic-distorted KonIQ-10k (Hosu et al., 2020), respectively. We use the same prompt structure and data throughout the training, and only reward on *quality scoring* regarding quality-relevant supervision. Refer to Appendix A for reward

function details. Following the splits of You et al. (2025), we evaluate all models on the test sets of KADID-10k and KonIQ-10k, and additionally on SPAQ (Fang et al., 2020), AGIQA-3K (Li et al., 2024) and CSIQ (Larson & Chandler, 2010), which are unseen throughout. Appendix B details the datasets used for training and testing.

4 RLVR'S EFFECT ON IQA REASONING CAPACITY OF VLMS

4.1 SETUPS

Evaluation protocols. Every base and RL-trained model is sampled $n = 512$ times per image[2], with temperature 1.0, top-p 0.95 and at most 4,096 generated tokens. All models are tested zero-shot with the same prompt as used in training to avoid any confounding effects. We extract the score enclosed by answer tags. A response without a parsable score is not resampled and counts as an invalid draw within the budget. Surprisingly, although without RLVR, the valid-draw rate can reach 99.5% for base models, and nearly 100% for RL-trained models. Appendix C lists the prompts.

Metrics. Measuring the reasoning ability is challenging, since methods like single-draw greedy decoding or multi-draw averaging can introduce biases (Wang et al., 2023; Holtzman et al., 2020). Generally, pass@k (Chen et al., 2021) is employed to evaluate model's reasoning boundary (Yue et al., 2025), measuring whether at least one of k responses passes a binary verifier. We employ the continuous extension of pass@k, *i.e.*, $\max_g @k$ (Walder & Karkhanis, 2025) to evaluate rank correlation performance in IQA. Each candidate X_j, for $j = 1, \ldots, n$, is a complete test-set prediction with one sampled score per image. Taking m for the vector of MOSs over the test set, we define the reward function as $g(X) = \mathrm{SRCC}(X, m)$, evaluated on the images of X. MOSs are used only to select among complete candidates, never among the scores of an individual image. Thus $\max_g @k$ is the expected best SRCC among k independently sampled candidates $X'_1, \ldots, X'_k$. We compute it from the n observed rewards $g_j = g(X_j)$, sorted as $g_{(1)} \leq \cdots \leq g_{(n)}$, with the unbiased estimation,

$$\max_g @k := \mathbb{E}_{X'_1,\ldots,X'_k}\left[\max_{1\leq j\leq k} g(X'_j)\right] \approx \sum_{i=k}^{n} \frac{\binom{i-1}{k-1}}{\binom{n}{k}}\, g_{(i)}, \tag{1}$$

for $k \in \{1, 2, 4, \ldots, n\}$. The metric is non-decreasing in k and reduces to pass@k for binary rewards, and its estimate equals the mean candidate SRCC at $k = 1$ and the largest observed SRCC at $k = n$. With this measure, we can easily estimate $\max_g @k$ with low variance. Appendix D lists more details.

4.2 IQA REASONING CAPACITY RESULTS

Effect of RLVR: boosted quality scoring capability at all sampling budgets. Figure 2 shows a consistent ordering across sampling budgets, with every RL-trained model above its base model on every test set and at both scales. At $k = 1$, the average-case SRCC of a single draw, RL-trained models lead by 0.03 to 0.46 SRCC, in line with the improvements reported for both recipes (Li et al., 2025; Wu et al., 2025), so RLVR makes a single sampled score considerably more accurate. The improvement persists as k grows to the hundreds. Base models gain more from additional sampling, yet it only closes a part of the gap with RL-trained models. At $k = 512$, the difference still remains positive in every comparison, and even the best of 512 base candidates lies below a single RL draw. Therefore, RLVR does raise the quality scoring capability beyond sampling effect.

Nature of the gain: denoising rather than coverage, better mapping concentrated in-distribution. Interestingly, this discovery looks contrast with the crossover that Yue et al. (2025) report on mathematics, code and visual reasoning tasks, where base models overtake RL-trained ones at large k. This indicates that, in a discrete open answer space, a large budget uncovers rarely sampled solutions, while IQA's SRCC aggregates over every image, the random error of individual scores moves it little from one candidate to the next, which therefore concentrate. RLVR sharpens this distribution by removing noise around the score of each image rather than alternative solutions (Appendix F contains detailed statistics). To investigate how much of the gain is denoising alone, we further evaluate $\mathrm{avg}\,@k$, the SRCC of the per-image average of k scores. As shown in Figure 3, averaging lifts base models more and saturates quickly. This indicates the single-draw gain of RLVR contains a denoising part, which the base model recovers by averaging its own draws, and a mapping part, which it can not. The mapping part is large in-distribution but small or absent out-of-distribution,

[2]We employ vLLM scheme (Kwon et al., 2023) to ensure independent and identically distributed sampling.

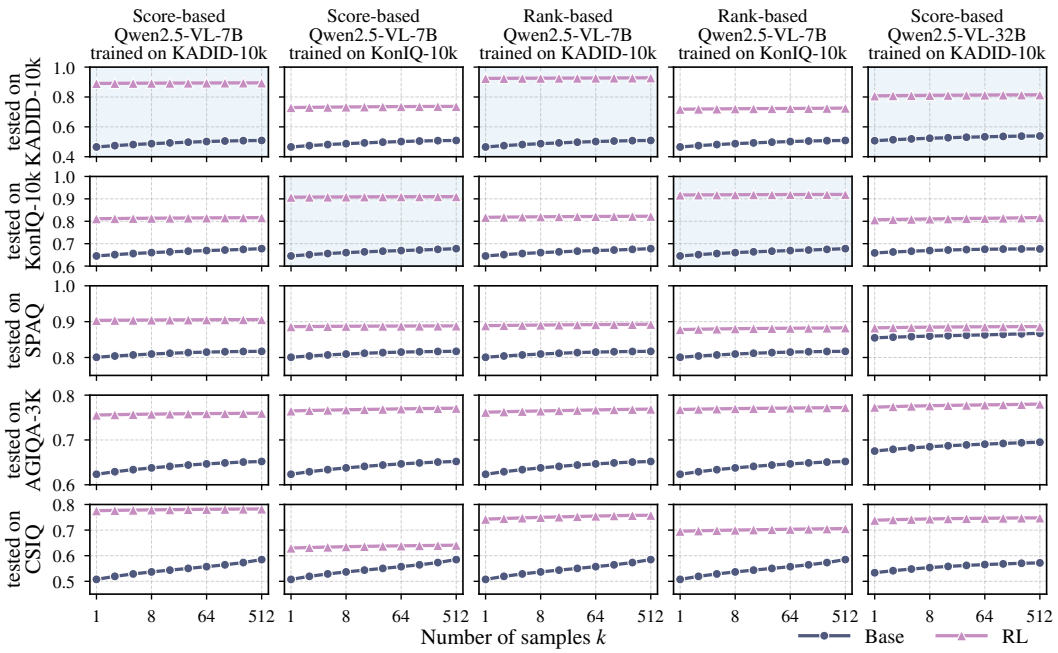

Figure 2: $\max_g @k$ curves for base and RL models. Shaded panels are in-distribution test sets.

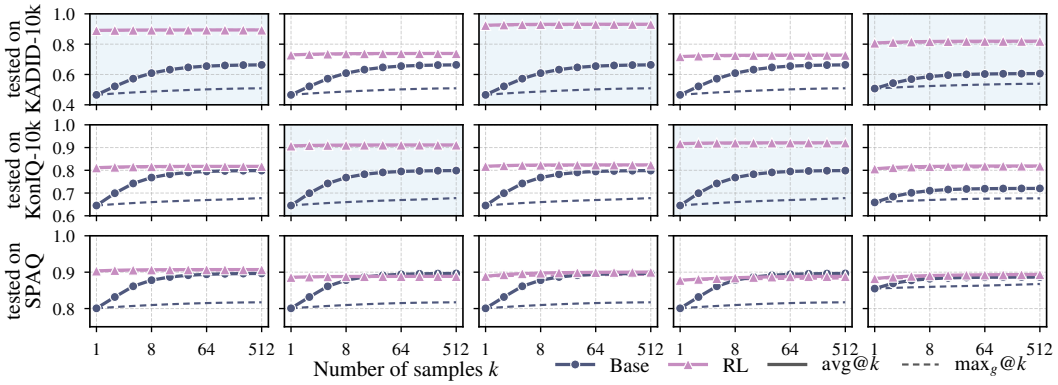

Figure 3: $\mathrm{avg}\,@k$ curves for base and RL models. Columns follow Figure 2. Refer to Appendix E for results on all test sets.

where the averaged base model often reaches or crossovers a single RL draw and the crossover of Yue et al. (2025) reappears. Therefore, RLVR improves the quality scoring in-distribution, and the generalization that Zhao et al. (2026) report thus is inherently contained in the base model.

Perplexity analysis. To see whether RL-trained responses remain likely under base models, we computed perplexity (Jelinek et al., 1977) of responses from each model, which measures the model's ability to predict the given response conditioned on the inputs. Lower perplexity indicates that the model has a higher likelihood of generating this response. We randomly sample 16 reasoning for each image, denoted $\mathbf{Y}_{\text{Base}}$, $\mathbf{Y}_{\text{Score}}$ and $\mathbf{Y}_{\text{Rank}}$ by the reward of the model, score every response under the base model of the same scale, and Figure 4 averages the test sets (Appendix F gives details). $\mathrm{PPL}_{\text{Base}}(\mathbf{Y}_{\text{Rank}} \mid x)$ occupies the lower part of the distribution of $\mathbf{Y}_{\text{Base}}$, so rank rewards concentrate sampling on the responses that the base model most readily produces. $\mathrm{PPL}_{\text{Base}}(\mathbf{Y}_{\text{Score}} \mid x)$ shifts upward: at 7B these responses

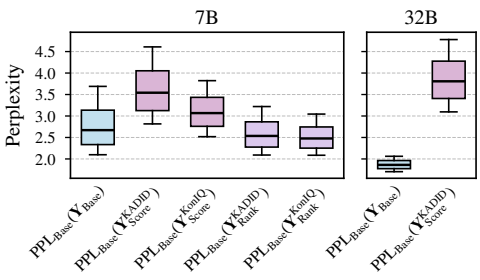

Figure 4: Perplexity of sampled responses under the base model of the same scale.

are less typical of the base model but still partly within its range, whereas at 32B they lie almost entirely above the narrow band of the base model. Score rewards thus move the text further outside the base distribution as scale grows, although neither reward targets the reasoning text, and in both cases the scores change substantially. This raises the question: *whether the gain passes through this altered textual CoT or bypasses it?*

5 Causal Role of CoT in IQA

5.1 Setups

Hypotheses. To answer the question above, we state three hypotheses shown in Figure 5. Under H-1 mediated gain, RLVR improves the quality reasoning expressed in the chain and the score is derived from the chain, so the gain is obtained along $x \to c \to s$. Under H-2 direct gain, RLVR improves the mapping directly from image to score along $x \to s$, and the chain is a by-product of the required format that does not carry the gain. Under H-3 joint gain, both paths are strengthened, so the chain becomes more informative about quality, the score becomes more sensitive to the image, and the gain draws on both.

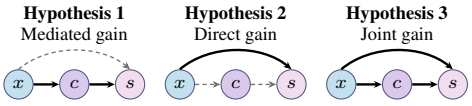

Figure 5: Three hypotheses on source of quality scoring gains, with the image x, the chain c and the score s. Solid black edges carry the gain and dashed grey edges do not.

Methods. In order to probe the hypotheses, we fix the input image, original prompt, and only manage on the chain closed by </think> that precedes the score. Then we let the models to generate the complete answer, from which the score is parsed from the <answer> tags. The first two act on the text of the CoT, and the third on its hidden state, detailed as follows:

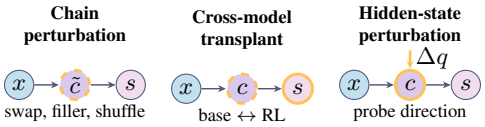

Figure 6: Three kinds of chain interventions. Gold marks where an intervention acts, a dashed outline marks a chain that is replaced.

- **Chain perturbation** replaces the chain with the one that the same model generated for a randomly drawn donor image (*swap*), with filler tokens of the same length (*filler*), or with its own words in random order (*shuffle*). *Swap* describes another image, and thus measures how strongly the score follows the mediated path. *filler* and *shuffle* remove the content of the chain while preserving length and format, so they test whether the content is necessary for scoring.
- **Cross-model chain transplant** tests whether a better CoT carries the gain. We only exchange chains sampled for the same image. Each RL-trained model generates quality scores with the chains of its base model, and vice versa. With the model fixed, the gap between the two chains is the gain carried by CoTs; and with the chain fixed, the difference between the two models is the gain that the chain does not carry. H-1 predicts the former, H-2 the latter, and H-3 both.
- **Hidden-state perturbation** acts on CoT's hidden states to see if we can steer the score. We fit a continuous linear probe of MOS on the hidden state of the last thinking token before </think> for every layer, shift this hidden state along the probe direction by the smallest displacement that changes the probe reading by a dose $\Delta q \in \{\pm 0.25, \pm 0.5, \pm 1\}$ on the $[1, 5]$ scale, and generate one answer per dose. A random direction orthogonal to the probe direction, with the same norm, serves as the orthogonal control. Details for probe training and steering can be found in Appendix G and Appendix H respectively.

We evaluate the models of Section 3 on KADID-10k and KonIQ-10k with the sampling parameters of Section 4.1, and apply the three interventions. The unmodified chains form the reference condition. Appendix H gives the full protocol.

Metrics. Let s and s' be the quality scores of the test images after uniform reasoning and under an intervention, and m their MOSs. Under *swap*, we report two correlations: $\rho = \mathrm{SRCC}(s', m)$ with the MOSs of the input images, which is high when the score follows the image, and $\gamma = \mathrm{SRCC}(s', m_z)$ with the MOSs m_z of the donor images, which is high when the score follows the chain. We further report $\Delta\,\mathrm{SRCC} = \rho - \mathrm{SRCC}(s, m)$ to measure the influence of intervention on each dataset. Perturbation success rate (PSR) (Ju et al., 2025) is adapted to measure the influence on individual image, which measures the fraction of images whose score moves towards the intended direction, *i.e.*, the quality of the donor image under *swap*, and the sign of Δq under *hidden-state perturbation*. *Filler* and *shuffle* define no intended direction, so PSR is not computed.

Table 1: **CoT intervention results.** Ref. denotes SRCC with the original chains. Refer to ***Metrics*** for details of employed measure. Transplant results are reported as model/chain: RL/Base uses the RL-trained model to score with base-model chains, while Base/RL denotes the reverse.

Test set	Model	*Ref.*	**Swap** ($\Delta\text{SRCC} = \rho - \textit{Ref.}$)				**Filler**	**Shuffle**	**Transplant SRCC**	
			ρ	ΔSRCC	γ	*PSR*	ΔSRCC	ΔSRCC	RL/Base	Base/RL
KADID-10k	Base 7B	0.601	0.053	−0.548	0.598	0.707	0.050	0.019	–	–
	Base 32B	0.576	−0.038	−0.614	0.584	0.709	0.017	0.013	–	–
	KADID-10k training set *(in-distribution)*									
	Score 7B	0.893	0.809	−0.084	0.355	0.851	0.000	−0.001	0.879	0.876
	Score 32B	0.819	0.393	−0.426	0.667	0.806	−0.013	−0.006	0.659	0.792
	Rank 7B	0.931	0.920	−0.011	0.108	0.743	−0.001	−0.001	0.909	0.861
	KonIQ-10k training set *(out-of-distribution)*									
	Score 7B	0.739	0.674	−0.065	0.282	0.755	−0.002	0.000	0.718	0.712
	Rank 7B	0.729	0.719	−0.010	0.104	0.724	−0.004	0.001	0.729	0.675
KonIQ-10k	Base 7B	0.761	0.078	−0.683	0.743	0.787	0.018	0.001	–	–
	Base 32B	0.709	−0.001	−0.710	0.712	0.762	−0.005	0.004	–	–
	KonIQ-10k training set *(in-distribution)*									
	Score 7B	0.910	0.871	−0.039	0.221	0.848	0.000	0.000	0.872	0.893
	Rank 7B	0.921	0.917	−0.004	0.054	0.821	0.000	0.000	0.912	0.867
	KADID-10k training set *(out-of-distribution)*									
	Score 7B	0.818	0.757	−0.061	0.250	0.813	−0.004	−0.002	0.818	0.833
	Score 32B	0.819	0.509	−0.310	0.542	0.810	−0.001	−0.001	0.747	0.792
	Rank 7B	0.825	0.817	−0.008	0.061	0.730	−0.002	−0.002	0.829	0.832

5.2 Reasoning Perturbation Results

Chains have less effect on quality scoring performance after RLVR. We first test H-1, under which the score is derived from the chain, with three kinds of perturbations. As shown in Table 1, after *swapping*, all PSRs are non-zero, so every model responds to the swapped chain. Base model's prediction is more correlated with donor's MOS instead of the input image's, indicated by γ is closer to Ref. than ρ. In contrast, 7B RL models show an opposite trend, where quality scores follows the input image more than an irrelevant CoT, while 32B RL models lies between the two. This demonstrates that RLVR moves the score from the mediated to the direct path, almost completely at 7B and partially at 32B. Meanwhile, removing the content of the chain gives a slightly different picture: *filler* tokens of the same length or *shuffling* the chain's own words can barely change model's performance for both the base and RL models. As such, the gain does not only rely on the model to derive the score from an explicit chain, while larger models (32B) can be more reliant. Therefore, H-1 does not explain it. Zhao et al. (2026) does not conflict with us, where scores do attend to the chain without being determined by its content. *It is worth noting that this finding concerns the use of the chain at inference and not its role during training.*

Chains of RL-trained models nevertheless carry quality information for scoring, and scale matters. We then examine whether chains contribute, which H-2 denies. As shown in the transplant columns in Table 1, base models performance can be enhanced by replacing its CoTs with those of its RL-trained counterpart. The chains of the RL-trained models thus carry quality information that the base model can use, in line with their higher quality decodability (Appendix G). Meanwhile, with the chains of the base model fixed, replacing the base model by the RL model generally raises performance. Therefore, both CoT and direct mapping from image carry the gain, supporting H-3. Scale Influences How the Two Factors Combine. At 7B the RL-trained model scores almost as well with the chains of its base model as with its own, so the better chain and the better model largely overlap as routes to the same gain, and the score depends little on the chain. At 32B the chain carries most of the gain and the RL-trained model loses part of it without its own chains, so the chain still feeds the score, as under the swap. What we find interesting is that this scale difference parallels the perplexity in Figure 4, in which the chains distribution of the 32B RL model lifts drastically compared with 7B. This could possibly indicate that the larger the shift of the chain distribution from that of the base model, the more the model relies on the chain.

More useful chains do not mean more dependence on chains. We further test H-1 at the level of the representation. We find out that RLVR raises the quality decodability of the last thinking token at every layer (Appendix G), so if the score were read from this representation, steering it along the probed quality direction would steer the score. However, Figure 7 shows that scoring can hardly be

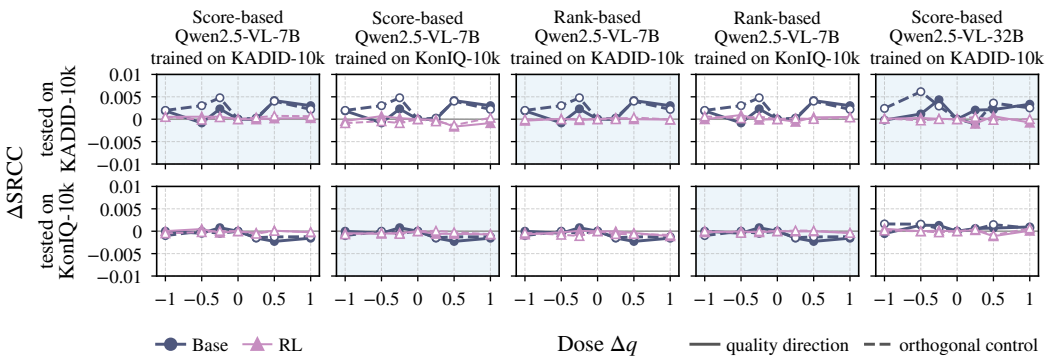

Figure 7: CoT hidden-state perturbation results against Δq. Solid lines shift the hidden state of the last thinking token at every layer. Dashed lines apply the orthogonal control of equal norm. Rows are the test sets whose training splits fitted the probes, and shaded panels are in-distribution test sets.

steered by disturbing CoT's hidden state. Shifting the hidden state by up to one point of the probe reading at every layer changes the SRCC by at most 0.005, no more than the orthogonal control, and base and RL-trained models behave alike. This again shows that the representation is informative about quality, not that the score depends on it, further supporting H-3.

Summary. In conclusion, RLVR improves both chain's usefulness and the scoring model itself, yet the better scores do not come with a stronger dependence on the chain during inference. Since the quality-relevant reward checks the agreement between predictions and MOSs, cues that correlate with MOS can probably be reinforced just as well. The next section therefore tries to answer: *whether RLVR can learn quality-irrelevant shortcuts to a higher scoring performance*.

6 Shortcuts to the Quality Reward

6.1 Setups

Methods. We test whether RLVR learns to exploit prompt cues that are unrelated to image quality. The images and their MOSs stay unchanged, and only a textual cue is appended to the scoring prompt. We use four kinds of cues, each with two variants, noted as P and Q: (i) *A/B identifiers* arbitrarily append an A or B; (ii) *source descriptions* state that the image comes from a good-quality source or a bad one; (iii) *numerical hints* give an unverified preliminary rating of 4.50 or 1.50; and (iv) *irrelevant text length* repeats a neutral sentence eight times or once to change text length. (ii) and (iii) state quality explicitly, whereas (i) and (iv) carry no quality information. For each kind of cue, we train two Qwen2.5-VL-7B-Instruct (Bai et al., 2025) models on KADID-10k (Lin et al., 2019), using GRPO with the score-based and format rewards of Section 3. We designed two training modes, which use the same images, image order and settings and differ only in whether the cue predicts MOS. Comparing the two trained models therefore shows what RLVR learns from a cue that predicts MOS. In *independent* training, which serves as the control, the cues are shuffled across images and are thus unrelated to MOS. In *correlated* training, P is paired with high-MOS and Q with low-MOS images on approximately 90% of training images. At test time, we change the relation between cue and MOS. Both trained models and the base model are evaluated on the KADID-10k test split with 16 draws per image under four cue assignments: *aligned* pairs P with high-MOS images as in correlated training, *independent* shuffles the cues, *reversed* pairs P with low-MOS images, and *removed* omits the cue. If the correlated model outperforms its independent control under alignment but falls below it under reversal, it has learned to exploit the cue, and its advantage under alignment cannot be attributed to a better assessment of image quality. Appendix I gives the full setting details.

Metrics. Let $\rho_{u,v}$ be the SRCC between the mean valid rating and MOS of the correlated or independent model $u \in \{\mathrm{corr}, \mathrm{indep}\}$ under test assignment v. We measure how much the correlated model relies on the cue by the interaction

$$I = \big(\rho_{\mathrm{corr,aligned}} - \rho_{\mathrm{indep,aligned}}\big) - \big(\rho_{\mathrm{corr,reversed}} - \rho_{\mathrm{indep,reversed}}\big). \tag{2}$$

An improved assessment persists under reversal and gives $I \approx 0$, whereas reliance on the cue gives $I > 0$. On individual images, the same-image score shift of a model is $\delta = N^{-1}\sum_i[\bar{s}(x_i, P) - \bar{s}(x_i, Q)]$, where $\bar{s}(x_i, P)$ is its mean valid rating of image x_i under cue P and N the number of

images with a valid rating under both cues. The pixels are fixed, so δ is attributable to the cue alone. Appendix I details the parsing rule and the treatment of invalid ratings.

6.2 Shortcut Results

Correlated training yields cue-dependent accuracy. We first examine whether a cue that predicts MOS during training raises test accuracy, and whether the resulting gain depends on the cue. Table 2 shows that correlated training improves aligned-test SRCC over independent training for all four cues, by 0.026 to 0.069. However, under reversed assignments, every correlated model falls below its independent control. The interactions I are 0.046 for A/B identifiers, 0.903 for source descriptions, 0.892 for numerical hints and 0.300 for text length. The independent controls remain between 0.814 and 0.825 SRCC across all cues and test conditions, and all comparisons retain the full test set, with valid-draw rates of at least 99.98% for the trained models. The advantage under alignment is therefore attributable to the cue and not to a better assessment of the image.

Table 2: **Shortcut results.** Corr. and Indep. denote models trained with correlated and independent cue assignments. Bold denotes the highest SRCC within each cue and test condition.

Cue	Model	Test assignment: SRCC ↑				δ
		Aligned	Independent	Reversed	Removed	
(i)	Base	0.611	0.619	0.618	0.613	−0.013
	Corr.	**0.843**	**0.824**	0.799	**0.821**	0.108
	Indep.	0.818	0.819	**0.819**	0.818	−0.004
(ii)	Base	0.727	0.571	0.407	0.613	0.235
	Corr.	**0.881**	0.550	−0.027	0.781	1.360
	Indep.	0.823	**0.821**	**0.819**	**0.821**	0.006
(iii)	Base	0.823	0.352	−0.315	0.613	0.820
	Corr.	**0.878**	0.521	−0.011	0.760	1.372
	Indep.	0.814	**0.814**	**0.817**	**0.820**	0.007
(iv)	Base	0.591	0.615	0.637	0.613	−0.034
	Corr.	**0.891**	0.771	0.594	**0.827**	0.587
	Indep.	0.822	**0.824**	**0.825**	0.823	−0.009

Cues with and without quality semantics are both exploited. We next examine how strongly the scores follow each cue. For (ii) source descriptions and (iii) numerical hints, which state quality directly, reversing the cues lowers the SRCC of the correlated models from 0.881 and 0.878 to -0.027 and -0.011. On the same image, replacing Q with P raises their mean rating. The independent models shift by less than 0.01 under the same replacement. Cues that carry no quality judgment are exploited as well, though more weakly. An arbitrary A/B identifier lowers the SRCC, and an irrelevant repeated sentence lowers it significantly with $\delta = 0.587$, while removing either cue returns the correlated model to the level of its control. Surface features of the prompt are thus exploited whenever they correlate with MOS, and the more so when they also state quality.

Cue sensitivity of the initial model. Finally, we examine whether this dependence is acquired through RL or is already present before training. As listed in Table 2, base models respond to (ii) source descriptions and (iii) numerical hints and its performance under reversed numerical hints is below the correlated model. The effect of RL therefore depends on RL training. Independent training largely removes this sensitivity, reducing both shifts below 0.01, whereas correlated training strengthens them above 1.3 and gains an advantage whenever the cue predicts MOS.

7 Conclusion

This work separates the gains of RLVR in IQA from sampling effects and examines whether they are obtained through explicit reasoning. RL-trained models remain above their base models at every sampling budget, so the gain is not a sampling effect. Averaging the draws of a base model recovers the denoising part of this gain, whereas the improved image-to-score mapping is concentrated in-distribution and small or absent out-of-distribution. Chain interventions on both text and hidden state show that RLVR yields more useful chains and a better scorer at the same time, so both the chain and the direct path from image to score carry the gain. Yet neither improvement brings a stronger dependence on the chain at inference: the scores of RL-trained models follow the input image more than a swapped chain, barely change when the content of the chain is removed, and can hardly be steered along the quality direction of its hidden state, more so at 7B than at 32B. Shortcut experiments further show that the reward is met as readily by textual cues correlated with MOS as by quality itself, even by cues without quality semantics, and that the advantage reverses with the correlation. Higher accuracy and more useful chains therefore do not establish that scoring relies on explicit quality reasoning, and our interventions concern the use of the chain at inference rather than its role during training. Reliable and explainable IQA calls for evaluating chain dependence and robustness to shortcuts, alongside rewards that connect the reasoning to the final judgment.

AI USE STATEMENT

In this work, we used generative AI tools to refine the experimental design, by providing feedback on procedures proposed by the authors, and to assist in implementing methods. We have not used generative AI tools to propose or refine research hypotheses or to interpret the results. The research questions, methodology, evaluation protocols, and scientific conclusions were developed by the authors, and tasks such as developing theoretical models or formulating and proving mathematical claims are not applicable to this work. Additionally, we used generative AI tools for language editing and for retrieval and discovery (e.g., finding related work); every retrieved reference was checked against the original publication by the authors. All AI-assisted text and code were reviewed and verified by the authors, who take full responsibility for the content and results presented in this work.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate the reproducibility of our results. The model configurations, RLVR training setup, reward formulation, sampling and evaluation protocols, and reasoning intervention procedures are described in the main paper and Appendix. We provide additional implementation details, including training and inference hyperparameters, prompt templates, score extraction procedures, sampling budgets, and experimental settings for the chain transplantation, swapping, removal, representation intervention, and shortcut experiments, in the Appendix. The datasets and evaluation metrics used in our experiments are publicly available and are identified together with the corresponding preprocessing and evaluation procedures. We will also provide the code and configuration files necessary to reproduce the main experiments and reported results.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, Linyi Yang, and Yue Zhang. How likely do LLMs with CoT mimic human reasoning? In *International Conference on Computational Linguistics*, 2025. arXiv:2402.16048.

Sebastian Bosse, Laura Acqualagna, Wojciech Samek, Anne K. Porbadnigk, Gabriel Curio, Benjamin Blankertz, Klaus-Robert Müller, and Thomas Wiegand. Assessing perceived image quality using steady-state visual evoked potentials and spatio-spectral decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(8):1694–1706, 2018. doi: 10.1109/TCSVT.2017.2694807.

Zhuoxuan Cai, Jian Zhang, Xinbin Yuan, Peng-Tao Jiang, Wenxiang Chen, Bowen Tang, Lujian Yao, Qiyuan Wang, Jinwen Chen, and Bo Li. Q-Ponder: A unified training pipeline for reasoning-based visual quality assessment. *CoRR*, abs/2506.05384, 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *CoRR*, abs/2505.05410, 2025.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *Transactions on Machine Learning Research*, 2024. arXiv:2402.18312.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3677–3686, 2020.

Dan Friedman and Adji Bousso Dieng. The Vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, Jan. 2020.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Annual Meeting of the Association for Computational Linguistics*, 2020. arXiv:2004.03685.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, December 1977. ISSN 1520-8524. doi: 10.1121/1.2016299. URL http://dx.doi.org/10.1121/1.2016299.

Renee Jia and Di Mu. From decorative to load-bearing: Task difficulty shapes the causal role of chain-of-thought. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=TiZQnKDIHq.

Ziheng Jia, Jiaying Qian, Zicheng Zhang, Zijian Chen, and Xiongkuo Min. Refine-IQA: Multi-stage reinforcement finetuning for perceptual image quality assessment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(27):22301–22309, 2026. doi: 10.1609/aaai.v40i27.39387.

Yuxuan Jiang, Yixuan Li, Hanwei Zhu, Siyue Teng, Fan Zhang, and David Bull. Q-Tacit: Image quality assessment via latent visual reasoning. *arXiv preprint arXiv:2603.22641*, 2026.

Tianjie Ju, Zhenyu Shao, Bowen Wang, Yujia Chen, Zhuosheng Zhang, Hao Fei, Mong-Li Lee, Wynne Hsu, Sufeng Duan, and Gongshen Liu. Probing then editing response personality of large language models. In *Conference on Language Modeling*, 2025. arXiv:2504.10227.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. arXiv:2309.06180.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702, 2023.

Eric C. Larson and Damon M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, Jan. 2010.

Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. AGIQA-3K: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6833–6846, 2024. doi: 10.1109/TCSVT.2023.3319020.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2306.03341.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-Insight: Understanding image quality via visual reinforcement learning. *Advances in Neural Information Processing Systems*, 38:41337–41362, 2025.

Yuan Li and Shin'ya Nishida. Understanding pure textual reasoning for blind image quality assessment. In *IEEE International Conference on Multimedia and Expo*, 2026. arXiv:2601.02441.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted IQA database. In *International Conference on Quality of Multimedia Experience*, pp. 1–3, 2019.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *International Joint Conference on Natural Language Processing and Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2023. arXiv:2301.13379.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024. arXiv:2402.13950.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in RLVR. *CoRR*, abs/2506.10947, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471, 2022. doi: 10.52202/068431-0687.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2305.04388.

Christian Walder and Deep Tejas Karkhanis. Pass@k policy optimization: Solving harder reinforcement learning problems. *Advances in Neural Information Processing Systems*, 38: 169225–169254, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. arXiv:2203.11171.

Zhou Wang and Alan C. Bovik. *Modern Image Quality Assessment*. Morgan & Claypool, 2006.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. In *International Conference on Learning Representations*, 2024a.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25490–25500. IEEE, 2024b.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning*, pp. 54015–54029, 2024c.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank. *Advances in Neural Information Processing Systems*, 38:97982–98005, 2025.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024.

Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *European Conference on Computer Vision*, pp. 259–276. Springer, 2024.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14483–14494. IEEE, 2025.

Qinan Yu, Alexa Tartaglini, Peter Hase, Carlos Guestrin, and Christopher Potts. Outcome rewards do not guarantee verifiable or causally important reasoning. *arXiv preprint arXiv:2604.22074*, 2026.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, 2025. arXiv:2504.13837.

Yuyi Zhang, Boyu Tang, Tianjie Ju, Sufeng Duan, and Gongshen Liu. Do latent tokens think? a causal and adversarial analysis of chain-of-continuous-thought. *CoRR*, abs/2512.21711, 2025.

Shijie Zhao, Xuanyu Zhang, Weiqi Li, Junlin Li, Li Zhang, Tianfan Xue, and Jian Zhang. Reasoning as representation: Rethinking visual reinforcement learning in image quality assessment. In *International Conference on Learning Representations*, 2026. arXiv:2510.11369.

Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. *Advances in Neural Information Processing Systems*, 37: 32611–32629, 2024.

A Details of Verifiable Rewards

This appendix gives the exact forms of the rewards summarized in Section 3. Both follow the official implementations, and scores and MOSs lie in $[1, 5]$.

Score-based reward. Following Q-Insight (Li et al., 2025), the predicted score s is compared with the MOS m directly on the continuous scale,

$$r^{\text{score}} = \mathbf{1}\big[\,|s - m| < \tau\,\big], \tag{3}$$

with tolerance $\tau = 0.35$.

Rank-based reward. Following VisualQuality-R1 (Wu et al., 2025), the reward verifies the relative quality of image pairs. For the j-th sampled score $s_{i,j}$ of image x_i and another image x_z in a batch of B images, the Thurstone model gives the comparison probability

$$p_{ij,z} = \Phi\left(\frac{s_{i,j} - \mu_z}{\sqrt{\sigma_i^2 + \sigma_z^2}}\right), \tag{4}$$

where G is the number of responses that GRPO samples for each image, μ_i and σ_i^2 are the mean and variance of the G scores sampled for image x_i, and Φ is the standard normal CDF. The reward is the fidelity between $p_{ij,z}$ and the ground-truth preference $p^*_{i,z}$, which equals 1, 0.5 or 0 when m_i is greater than, equal to or less than m_z, averaged over the other images in the batch:

$$r^{\text{rank}}_{i,j} = \frac{1}{B-1} \sum_{z \neq i} \left(\sqrt{p_{ij,z}\, p^*_{i,z}} + \sqrt{(1 - p_{ij,z})(1 - p^*_{i,z})} \right). \tag{5}$$

Format reward. A binary reward $r^{\text{fmt}} \in \{0, 1\}$ checks that the response matches the pattern `<think>⋯</think><answer>⋯</answer>`. The total reward is $r = r^{\text{score}} + r^{\text{fmt}}$ or $r = r^{\text{rank}} + r^{\text{fmt}}$.

B Details of Datasets

Table 3 summarizes the five datasets. All splits are taken from You et al. (2025). KADID-10k is split by reference image, so that the 16 reference images of its test set do not appear among the 65 of its training set, and both parts cover all 25 distortion types at 5 levels. The CSIQ subset contains 30 reference images under 6 distortion types. For a model trained on KADID-10k the test set of KonIQ-10k is also out-of-distribution, and vice versa. The native label scales differ across datasets, and CSIQ reports difference scores for which lower is better. All labels are therefore mapped to MOSs in $[1, 5]$ for which higher is better, and all results are computed on this scale.

Table 3: Datasets used in this work. Training images are used only for RLVR. SPAQ, AGIQA-3K and CSIQ are unseen by every model.

Dataset	Content	Resolution	Train	Test
KADID-10k (Lin et al., 2019)	synthetic distortions	512×384	8,125	2,000
KonIQ-10k (Hosu et al., 2020)	authentic, in the wild	1024×768	7,046	2,010
SPAQ (Fang et al., 2020)	authentic, smartphone photos	1080×1440 to 5824×4368	none	2,224
AGIQA-3K (Li et al., 2024)	AI-generated images	512×512	none	2,982
CSIQ (Larson & Chandler, 2010)	synthetic distortions	512×512	none	866

C Prompt Templates

Same prompts for image quality scoring are used in training and testing. The image is attached to the user message in both cases.

Prompt for training and testing.

> **System.** `A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>`
>
> **User.** `What is your overall rating on the quality of this picture? The rating should be a float between 1 and 5, rounded to two decimal places, with 1 representing very poor quality and 5 representing excellent quality. Return the final answer in JSON format with the following keys: "rating": The score.`

D Estimation Details

Candidate pool. For each image, the 512 draws are put in a random order, which is drawn independently for each image, shared across models and independent of the MOSs. Taking the score at the same position of this order from each of the N test images gives one candidate, hence $n = 512$ candidates per model and test set. A budget of k candidates corresponds to k draws per image, or Nk single-image generations. All models use the complete test sets in Table 3. The candidate pool is fixed across budgets and bootstrap rounds, and the estimator in Equation 1 averages over all k-subsets without Monte Carlo subset sampling.

Confidence intervals. The intervals in Table 4 are pointwise 95% percentile intervals over 1,000 image-bootstrap rounds. Each round resamples N image indices with replacement, recomputes the SRCC of all 512 candidates on those indices, and then reevaluates the estimator in Equation 1. The same resampled indices are used for every model on a dataset. Intervals for RL–base differences are obtained by subtracting the paired bootstrap estimates before taking percentiles. These intervals

describe image-resampling uncertainty conditional on the observed candidate pool; they do not include all uncertainty from fresh model generations. Reselecting the best candidates can shift the bootstrap distribution, so percentile intervals need not be centered on, or contain, the original estimate. Intervals are pointwise rather than simultaneous across models and test sets.

Unparsable responses. A raw draw is valid when it yields a finite rating in $[1, 5]$, parsed from a rating field or a bare numerical answer. An invalid draw occupies the budget and is neither resampled nor imputed. Each candidate's SRCC uses only its valid images, so the effective image set can vary between candidates. Tied scores and MOSs receive average ranks. A candidate with fewer than two valid images or constant scores or MOSs is assigned reward zero, and no candidate in our experiments required this convention.

E Score Averaging

$\max_g @k$ selects among complete candidates with the MOSs. This appendix uses the same budget without any MOS, by averaging the scores of each image. For a set of k candidates, the valid scores that the candidates assign to an image are averaged, and $\mathrm{avg}\ @k$ is the SRCC between these averages and the MOSs, in expectation over sets of k candidates. The computation shares the candidate pool, the image sets, the treatment of unparsable scores and the paired bootstrap indices of Appendix D, so it coincides with $\max_g @k$ at $k = 1$, which we verify numerically for the estimate and for every bootstrap round. An image without a valid score among the k draws is omitted from the SRCC of that set. At $k = 1$ all 512 candidates are used and at $k = 512$ there is a single set. For other k the expectation is estimated over 128 sets formed from disjoint partitions of the candidates, shared by all models and test sets. The standard error of this estimate is at most 0.002.

Figure 8 shows the curves. The averaged base model at $k = 64$ is within 0.009 SRCC of its value at $k = 512$ on every test set. RL-trained models gain little from averaging, at most 0.041, because their scores are already concentrated. At equal budgets the averaged base model overtakes the averaged RL-trained model in 6 of the 25 comparisons, on SPAQ, AGIQA-3K and CSIQ.

Table 4 splits the gain of a single RL draw over a single base draw, RL@1 − Base@1, into two parts that sum to it exactly. Denoising is Base avg@512 − Base@1, the part that the base model recovers by averaging its own draws. Mapping is RL@1 − Base avg@512, the margin of a single RL draw over the averaged base model. Intervals are paired pointwise 95% bootstrap intervals as in Appendix D. The mapping part exceeds 0.20 in-distribution on KADID-10k and 0.10 in-distribution on KonIQ-10k and from KADID-10k to CSIQ. In the other out-of-distribution comparisons between KADID-10k, KonIQ-10k and CSIQ it ranges from -0.012 to 0.086. On SPAQ and AGIQA-3K it ranges from -0.019 to 0.019, and it is negative in eight of these ten comparisons.

F Sampling Distribution Statistics

This appendix reports the statistics of the sampled responses that Section 4.2 summarizes. They reuse the draws of Section 4.1 without further generation and cover KADID-10k, KonIQ-10k, SPAQ, AGIQA-3K and CSIQ. In all tables an RL-trained model is named by its reward type and training set.

Score and chain statistics. For each image we collect the valid responses among its 512 draws. The score statistics are the Shannon entropy, in bits, of the ratings rounded to two decimals, and the standard deviation of the ratings. The chain statistics are the semantic diversity and the length of the reasoning chains. Diversity is the Vendi Score (Friedman & Dieng, 2023), the exponential of the von Neumann entropy of the cosine similarity matrix divided by the number of chains, computed on bge-base-en-v1.5 embeddings (Xiao et al., 2024) of 128 chains drawn at random per image so that values are comparable across models. It equals 1 when all chains are identical and 128 when they are mutually orthogonal. Length counts words after lowercasing and removing punctuation. Every entry of Tables 5 and 6 is the median over the images of a test set.

RL lowers the entropy and the standard deviation of the scores on every test set and at both scales. The standard deviation shrinks by a factor of 3.7 to 12 at 7B and of 1.4 to 4.4 at 32B. The chains become shorter, most visibly at 32B, where the median length falls from above 200 words to below

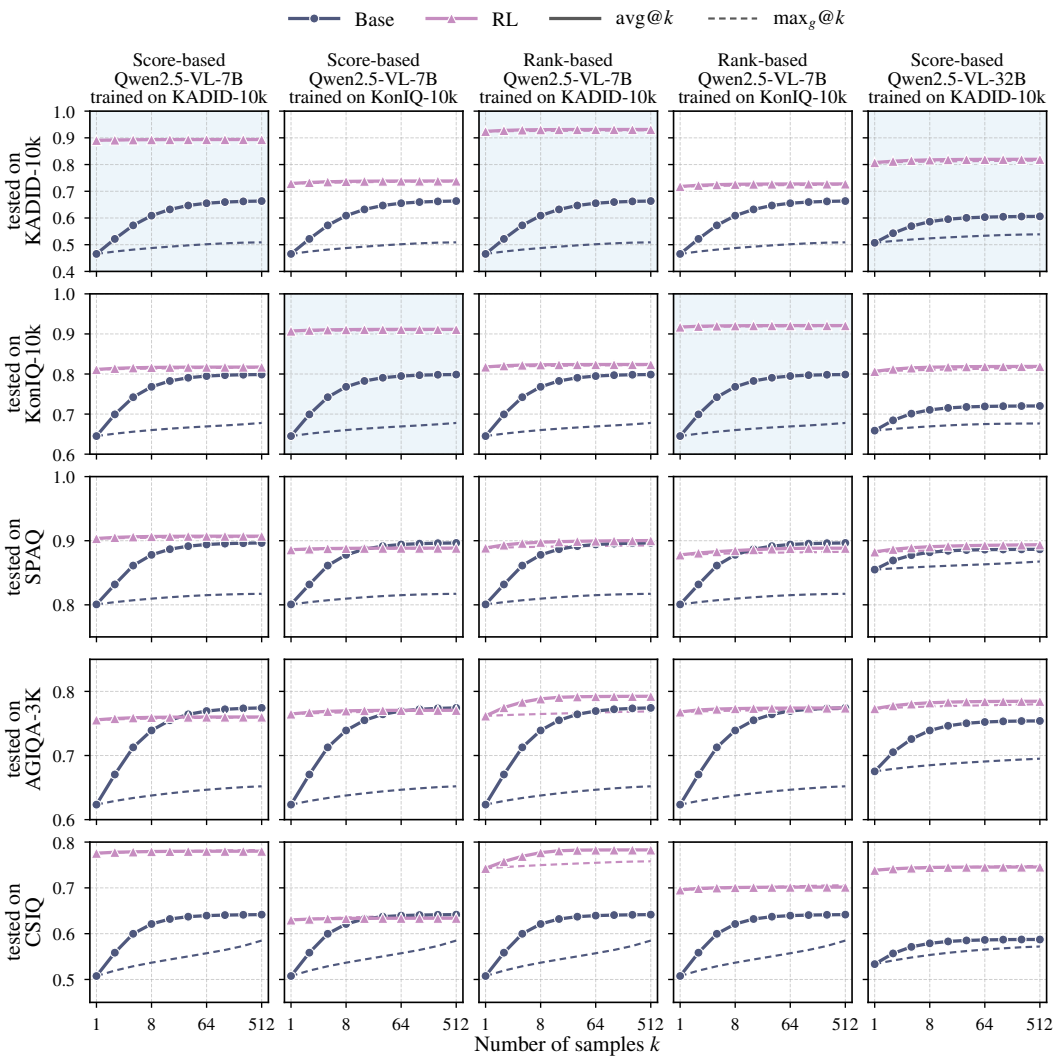

Figure 8: avg @k (solid lines with markers) of base models and their RL-trained counterparts, in the layout of Figure 2. The scores of k draws are averaged for each image before the SRCC is computed, and no MOS is used. Dashed lines repeat $\max_g$ @k from Figure 2. Shaded panels are in-distribution test sets. Averaging lifts the base models far more than selection and saturates near $k = 64$. On SPAQ and AGIQA-3K the averaged 7B base model reaches the RL-trained models.

60. Chain diversity decreases at 7B by 18% to 32%, less than the effective number of distinct scores 2^H, which decreases by 59% to 76%. At 32B chain diversity increases, although the large change in length confounds this comparison.

Perplexity under the base model. Following Yue et al. (2025), we measure how likely the responses of each model are under the base model of its own scale, the quantity $\mathrm{PPL}_{\mathrm{Base}}(y \mid x)$ of Section 4.2, so the 7B base model scores the responses of the 7B models and the 32B base model scores those of the 32B models. For a random 20% of the images of a test set, the same images for every model, we take the first 16 parsable responses of each model, serialize the chain and the rating in the tag format of Appendix C, and compute the per-token perplexity of the response conditioned on the image and the prompt of Appendix C. Every model thus contributes 16 responses for each sampled image of KADID-10k, KonIQ-10k, AGIQA-3K and CSIQ. Since the responses are serialized again, perplexities are comparable across response sources under a fixed scorer, and we do not compare them across scorers. Figure 4 pools the responses of the four test sets. Table 7 reports each test set

Table 4: Decomposition of the single-draw gain of RL into denoising and mapping. Base@1 and RL@1 are SRCCs of single draws, which equal $\max_g @1$, and Base avg is $\text{avg} @512$ of the base model, its SRCC after averaging 512 scores per image. Each RL-trained model is compared with the base model of its scale. † marks in-distribution test sets.

Test set	RL-trained model	Base@1	Base avg	RL@1	Denoising	Mapping	95% interval
KADID-10k	Score, KADID-10k, 7B†	0.465	0.663	0.891	0.198	0.227	[0.207, 0.249]
	Score, KonIQ-10k, 7B	0.465	0.663	0.729	0.198	0.066	[0.051, 0.080]
	Rank, KADID-10k, 7B†	0.465	0.663	0.924	0.198	0.261	[0.240, 0.285]
	Rank, KonIQ-10k, 7B	0.465	0.663	0.718	0.198	0.054	[0.035, 0.073]
	Score, KADID-10k, 32B†	0.507	0.606	0.808	0.099	0.202	[0.182, 0.221]
KonIQ-10k	Score, KADID-10k, 7B	0.645	0.799	0.811	0.154	0.013	[−0.005, 0.028]
	Score, KonIQ-10k, 7B†	0.645	0.799	0.907	0.154	0.109	[0.096, 0.122]
	Rank, KADID-10k, 7B	0.645	0.799	0.818	0.154	0.019	[0.005, 0.034]
	Rank, KonIQ-10k, 7B†	0.645	0.799	0.917	0.154	0.119	[0.105, 0.133]
	Score, KADID-10k, 32B	0.659	0.720	0.807	0.062	0.086	[0.073, 0.101]
SPAQ	Score, KADID-10k, 7B	0.801	0.897	0.903	0.096	0.007	[0.003, 0.012]
	Score, KonIQ-10k, 7B	0.801	0.897	0.886	0.096	−0.011	[−0.015, −0.006]
	Rank, KADID-10k, 7B	0.801	0.897	0.889	0.096	−0.008	[−0.013, −0.003]
	Rank, KonIQ-10k, 7B	0.801	0.897	0.878	0.096	−0.019	[−0.024, −0.013]
	Score, KADID-10k, 32B	0.855	0.887	0.883	0.032	−0.004	[−0.010, 0.002]
AGIQA-3K	Score, KADID-10k, 7B	0.623	0.774	0.756	0.151	−0.019	[−0.031, −0.007]
	Score, KonIQ-10k, 7B	0.623	0.774	0.765	0.151	−0.009	[−0.019, −0.000]
	Rank, KADID-10k, 7B	0.623	0.774	0.762	0.151	−0.012	[−0.026, 0.000]
	Rank, KonIQ-10k, 7B	0.623	0.774	0.768	0.151	−0.006	[−0.018, 0.005]
	Score, KADID-10k, 32B	0.675	0.754	0.773	0.079	0.019	[0.008, 0.030]
CSIQ	Score, KADID-10k, 7B	0.507	0.642	0.776	0.134	0.134	[0.098, 0.172]
	Score, KonIQ-10k, 7B	0.507	0.642	0.630	0.134	−0.012	[−0.033, 0.009]
	Rank, KADID-10k, 7B	0.507	0.642	0.743	0.134	0.101	[0.058, 0.144]
	Rank, KonIQ-10k, 7B	0.507	0.642	0.696	0.134	0.054	[0.031, 0.077]
	Score, KADID-10k, 32B	0.534	0.587	0.738	0.054	0.151	[0.117, 0.185]

Table 5: Statistics of the scores sampled for an image, reported as medians over the images of each test set. H is the Shannon entropy in bits of the ratings rounded to two decimals, and SD is the standard deviation of the ratings on the $[1, 5]$ scale.

	KADID-10k		KonIQ-10k		AGIQA-3K		CSIQ	
Model	H	SD	H	SD	H	SD	H	SD
Base 7B	3.32	0.282	3.57	0.384	3.30	0.372	3.84	0.545
Score, KADID-10k, 7B	1.67	0.070	1.60	0.073	1.80	0.090	1.77	0.070
Score, KonIQ-10k, 7B	1.69	0.040	1.69	0.037	2.01	0.046	2.00	0.044
Rank, KADID-10k, 7B	1.89	0.075	1.90	0.082	1.88	0.100	1.92	0.093
Rank, KonIQ-10k, 7B	1.99	0.045	2.02	0.056	2.01	0.065	2.14	0.091
Base 32B	1.79	0.151	1.94	0.296	1.85	0.184	2.51	0.403
Score, KADID-10k, 32B	1.01	0.106	0.90	0.103	1.12	0.133	0.89	0.092

under the 7B base model for every source, so its 32B rows describe the 32B responses under the smaller scorer and are descriptive only, whereas the 32B panel of Figure 4 uses the 32B base model. The 10th to 90th percentile range of the responses of the 7B base model itself is 2.05 to 3.58 on KADID-10k, 2.09 to 3.57 on KonIQ-10k, 2.13 to 3.83 on AGIQA-3K and 2.16 to 3.67 on CSIQ. The medians of the rank-reward models lie below the median of the base model on every test set. The medians of the score-reward models lie above it and below its 90th percentile. Under the 32B base model, the 10th to 90th percentile range of its own responses is 1.70 to 2.06 when the four test sets are pooled, and the 10th percentile of the score-reward 32B model, 3.10 when pooled, lies above the 90th percentile of the 32B base model on every test set.

Table 6: Statistics of the reasoning chains sampled for an image, reported as medians over the images of each test set. V is the Vendi Score of 128 chains and L is the chain length in words.

	KADID-10k		KonIQ-10k		AGIQA-3K		CSIQ	
Model	V	L	V	L	V	L	V	L
Base 7B	2.62	90	2.68	90	2.88	97	2.87	90
Score, KADID-10k, 7B	2.14	76	2.11	77	2.28	78	2.24	72
Score, KonIQ-10k, 7B	2.14	66	2.16	69	2.35	65	2.31	66
Rank, KADID-10k, 7B	1.81	77	1.85	80	2.10	75	2.02	75
Rank, KonIQ-10k, 7B	1.78	74	1.86	77	2.08	72	2.05	75
Base 32B	1.51	211	1.56	217	1.71	202	1.64	215
Score, KADID-10k, 32B	1.87	56	1.87	57	2.01	56	1.89	54

Table 7: Median per-token perplexity of sampled responses under the 7B base model, conditioned on the image and the prompt.

Response source	KADID-10k	KonIQ-10k	AGIQA-3K	CSIQ
Base 7B	2.59	2.63	2.75	2.72
Score, KADID-10k, 7B	3.56	3.30	3.76	3.48
Score, KonIQ-10k, 7B	2.94	3.02	3.16	3.13
Rank, KADID-10k, 7B	2.47	2.43	2.65	2.58
Rank, KonIQ-10k, 7B	2.39	2.41	2.56	2.56
Base 32B	2.23	2.30	2.41	2.35
Score, KADID-10k, 32B	3.13	3.06	3.36	3.19

G Layer-wise Quality Decodability

We measure how well image quality can be read from the reasoning representation at each transformer layer. For each frozen model, we take the hidden state of the last thinking token before the closing </think> tag and fit a separate linear ridge probe to predict MOS at every layer. We call the resulting test-set SRCC *quality decodability*: it measures information available to a linear readout, rather than the sensitivity of the model's final rating to an intervention.

Probe fitting and evaluation. Each probe is fitted on the training split of KADID-10k or KonIQ-10k and evaluated on the test split of the same dataset (Table 3). All seven models are evaluated on the full test splits. Features are standardized using training-set statistics. For each layer, ridge regularization is selected by SRCC on a single 20% image split within the training set, followed by refitting on the full training set. Test MOSs are used only for evaluation. We report point estimates from a single fit per layer, without uncertainty bands. The control probes, fitted after shuffling training MOSs, have lower test SRCC than their quality probes at every measured layer.

Depth and model scale. Figure 9 shows higher test SRCC for the RL-trained representation than for its base counterpart at every layer on both datasets. The separation is larger at 7B; at 32B, the layer-wise RL–base gap ranges from 0.039 to 0.135 on KADID-10k and from 0.013 to 0.058 on KonIQ-10k. Thus, quality is linearly accessible throughout the representation of the last thinking token, with a depth profile that depends on both the model and dataset. These observational readouts do not establish that the final score uses the decoded information. In particular, the final RMSNorm representation remains informative for a probe even though changing that CoT token after the last attention block cannot affect subsequent answer tokens.

H Details of Chain Interventions

This appendix details the interventions of Section 5.

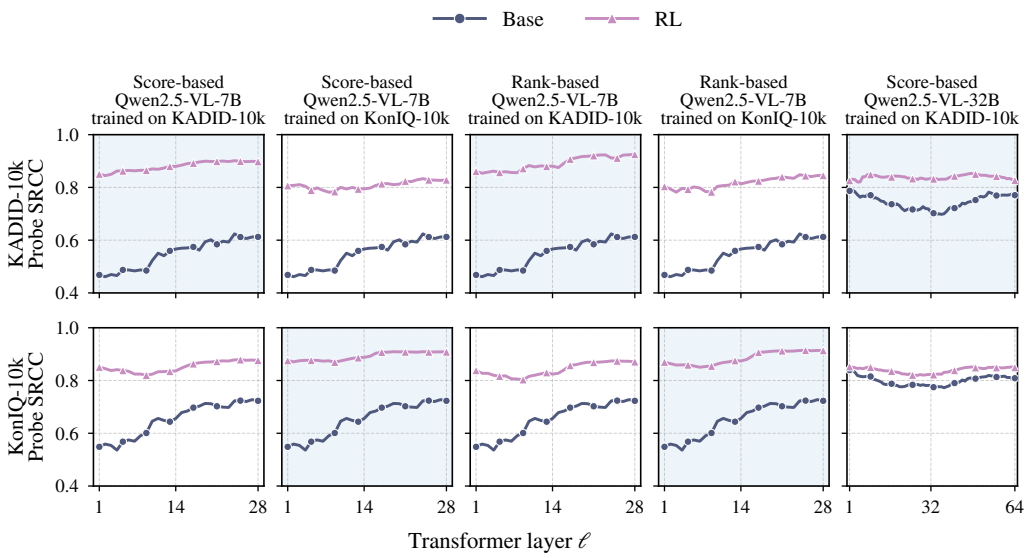

Figure 9: Layer-wise quality decodability at the last thinking token. Each curve reports the held-out SRCC between a layer's linear-probe predictions and MOS, for a base model or its RL-trained counterpart of the same scale. Rows specify the probe dataset; columns and colors follow Figure 2. Shaded panels indicate that the dataset matches the VLM's RL training set. In every panel, both probes are fitted on the training split of the row's dataset and evaluated on its disjoint test split. The horizontal axis uses actual layer indices: 1–28 for 7B and 1–64 for 32B. The final point is the final RMSNorm output. Lines connect all measured layers without smoothing; markers are subsampled for legibility. All panels share the same vertical range.

Intervened scoring. Each model keeps the prompt it uses in Section 4.1. The assistant turn is prefilled with the opening <think> tag, the fixed chain, the closing </think> tag and the opening <answer> tag, and the model continues autoregressively with temperature 1.0, top-p 0.95, no top-k filtering and at most 128 new tokens. A continuation is valid when it yields a complete finite rating in $[1, 5]$, parsed from a rating field or a bare number, and a truncated answer is never read as a score. An invalid continuation occupies its slot and is neither resampled nor imputed. The five chains of an image are its first five responses with a parsable answer among the draws of Section 4.1. Under the reference, swap, filler and shuffle conditions and in the cross-model transplant, each of the five chains receives eight answers, and the score of an image under a condition is the mean of the valid ratings among these 40 continuations. The reference condition scores each model with its own unmodified chains. The continuations of an image are generated with matched random states across conditions.

Chain perturbation. Each image draws one donor independently and with replacement from the other images of the same test set. The assignment is shared by the seven models, and the SRCC between the MOSs of the images and of their donors is 0.001 on KADID-10k and -0.011 on KonIQ-10k. Under the swap, the p-th chain of an image is replaced by the p-th chain that the same model sampled for the donor. The filler repeats a period followed by a space as many times as the original chain has tokens. The shuffle permutes the whitespace-separated words of the original chain, which preserves the words and removes their order.

Cross-model chain transplant. The five pairs are the 7B base model with each of the four RL-trained 7B models, and the 32B base model with the RL-trained 32B model. The two cells in which a model scores with its own chains are the reference condition above. In the two crossed cells, the scorer receives the five chains that the other model of the pair sampled for the same image, so the four cells share their images and their two sets of chains. Every scorer keeps its own prompt and answer format, and only the text between the think tags is replaced.

Hidden-state perturbation. The probes are those of Appendix G, fitted on the training split of the dataset under test. For each image the model samples one chain of at most 2,048 tokens, the original answer is discarded and the prefix ends with the opening <answer> tag. Let h_ℓ be the hidden state of the last thinking token at layer ℓ and $q_\ell = w_\ell^\top \big((h_\ell - \bar{h}_\ell)/\eta_\ell\big) + b_\ell$ the reading of its standardized probe, where $\bar{h}_\ell$ and η_ℓ are the coordinate-wise mean and standard deviation of h_ℓ over the training split, with gradient $v_\ell = w_\ell/\eta_\ell$ in the space of the hidden state. Along the probe direction we add $\Delta h_\ell = \Delta q\, v_\ell/\|v_\ell\|^2$, the displacement of smallest norm that changes the reading by Δq. The orthogonal control adds $\Delta h_\ell = \Delta q\, r_\ell/\|v_\ell\|$, where r_ℓ is a random unit vector orthogonal to v_ℓ, which matches the norm and keeps the sign of the dose. Displacements are applied once, when the prefix is encoded, at layers 1 to 27 of the 7B models and 1 to 63 of the 32B models. The output of the final normalization is excluded, since no later token attends to it. The shift at a layer is relative to the reading that reaches this layer, so the shifts of earlier layers propagate. At the nominal dose of ± 1, the reading at the last perturbed layer therefore moves by 2.9 to 3.8 points for the 7B models and by 5.5 to 6.6 points for the 32B models on average over the test images, whereas the orthogonal control moves it by less than 0.01. Between 89% and 96% of the test images of every model receive an identical rating in all 13 conditions. Figure 10 shows the resulting readings for one test image. Displacements are calibrated after rounding to bfloat16 so that the realized change of each reading matches the dose within a fixed tolerance. One answer is generated per direction and dose with matched random states, which gives 13 scores per image, one without perturbation, six along the probe direction and six under the orthogonal control.

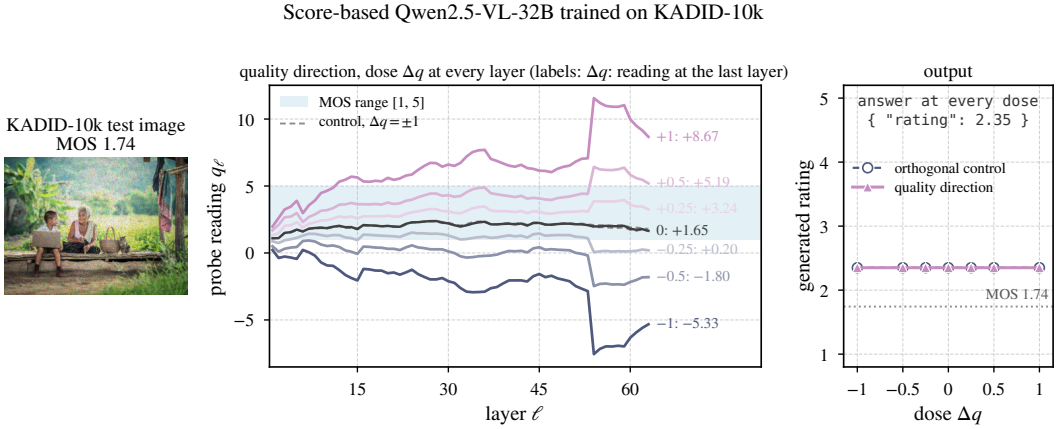

Figure 10: One KADID-10k test image under hidden-state perturbation of the score-based Qwen2.5-VL-32B trained on KADID-10k. Left, the image and its MOS. Middle, the probe reading of the last thinking token after injection at every perturbed layer, one curve per dose along the probe direction, with the reading at the last layer printed at the end of each curve. Dashed grey curves are the orthogonal control at $\Delta q = \pm 1$ and the shaded band is the MOS range. The local doses accumulate across layers, so the reading at layer 63 moves from 1.65 to -5.33 and to 8.67. Right, the rating generated under every dose along the probe direction and under the orthogonal control, which is 2.35 in all 13 conditions, and the answer text, which is repeated verbatim. Among the test images whose rating is identical in all 13 conditions, 94% of the test set for this model, this image has the largest excursion of the final reading.

Statistics. All comparisons use the images with a valid score in every compared condition. Under the reference, swap, filler and shuffle conditions, this set is the full test set of KADID-10k and of KonIQ-10k for all seven models.

I Details of Shortcut Experiments

This appendix details the controlled experiments in Section 6.1.

Textual cues. All conditions use the scoring prompt in Appendix C with the added instruction "Rate the image itself." Each experiment then appends one of two cue variants, P and Q. *A/B identifiers* append the arbitrary label `source_id: A` or `source_id: B`. *Source descriptions* describe a curated exhibition archive with polished, carefully prepared entries, or an informal draft archive with rough, hastily prepared entries. *Numerical hints* provide an external preliminary rating of 4.50 or 1.50, explicitly described as unverified for the current image. *Irrelevant text length* repeats the neutral sentence "Inventory note: This file is included in the current archive batch." eight times or once. In each case, the first variant is P and the second is Q. The identifiers and the repeated sentence carry no quality judgment, whereas the descriptions and numerical hints do. The two identifiers differ in one letter and the two descriptions have equal token lengths, as do the two numerical hints, so text length is manipulated only in the final experiment.

Correlated and independent training. All four experiments share the training split of KADID-10k and the same P/Q assignments. We divide images into high/low quality groups at the training MOS median, with ties assigned to the high group. In correlated training, P accompanies high quality and Q low quality on approximately 90% of images. Independent training permutes these assignments while preserving the P/Q frequencies, giving approximately 49.4% agreement. Both models are trained with GRPO using the score-based and format rewards of Section 3, eight rollouts per image exposure, learning rate 10^{-6}, KL coefficient 0.001 and training seed 42. Each model is trained for a single pass over the training split in 508 updates of 16 images, with three images repeated to fill the last update.

Test conditions. We evaluate the initial model and both final models on the test split of KADID-10k, whose reference images are unseen during training. Each image is scored with cue P, cue Q and no cue, with 16 complete think-and-answer generations per variant. A fixed assignment then defines four conditions: *aligned* assigns P at or above the training MOS median and Q below it; *independent* permutes these assignments; *reversed* swaps P and Q for every image; and *removed* omits the cue while retaining the shared instruction. Test MOSs define the evaluation assignments, and the model receives the resulting cue alongside the image and prompt. Training and evaluation use temperature 1.0, top-p 0.95, no top-k filtering and a 1,024-token output limit.

Metrics. We average the valid complete ratings among the 16 draws of each image and report SRCC against MOS. Parsing accepts a complete tagged answer or a standalone JSON rating or number in $[1, 5]$; failures occupy the budget without resampling or imputation. Comparisons within an experiment use the common valid-image intersection across models and conditions, with equal image weights and average ranks for tied scores. The interaction I of Equation 2 and the score shift δ of Section 6.1 are computed from these mean ratings.