CONTENT-ADAPTIVE AESTHETIC ALIGNMENT FOR TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Aesthetic evaluation for text-to-image (T2I) generation must distinguish an inapplicable criterion from a poorly realized one. To assess diverse visual content, we define a taxonomy of six aesthetic criteria—realism, lighting, anatomy, creativity, color harmony, and typography—and construct an aesthetic preference corpus of approximately 240K distinct images and 840K pairwise comparisons from human and GPT-assisted criterion-specific rankings. Based on this corpus, we train **A**pplicability-Aware **W**eighted **A**esthetic **RE**ward (AWARE), which jointly predicts quality and applicability for all six dimensions in one model. We then use AWARE to design an applicability-aware reinforcement-learning algorithm for text-to-image post-training, dynamically selecting and weighting only the criteria relevant to each sample. To evaluate both aesthetic judgment and post-trained generation, we introduce AWARE-RBench for criterion-conditioned preference and applicability evaluation, and AWARE-GenBench for role-conditional T2I aesthetic quality. AWARE improves over the baseline by 60.92% on the aesthetic-model evaluation. Across Z-Image and SenseNova-U1.5, post-trained T2I models achieve average relative gains of 19.01%, 21.72%, and 2.39% on Photography, Artwork, and Infographic, respectively. On general-purpose benchmarks such as Qwen-Image-Bench, the aligned models remain competitive with existing open-source systems. Side-by-side visual comparisons further confirm improvements in fidelity, composition, natural lighting, and text rendering.

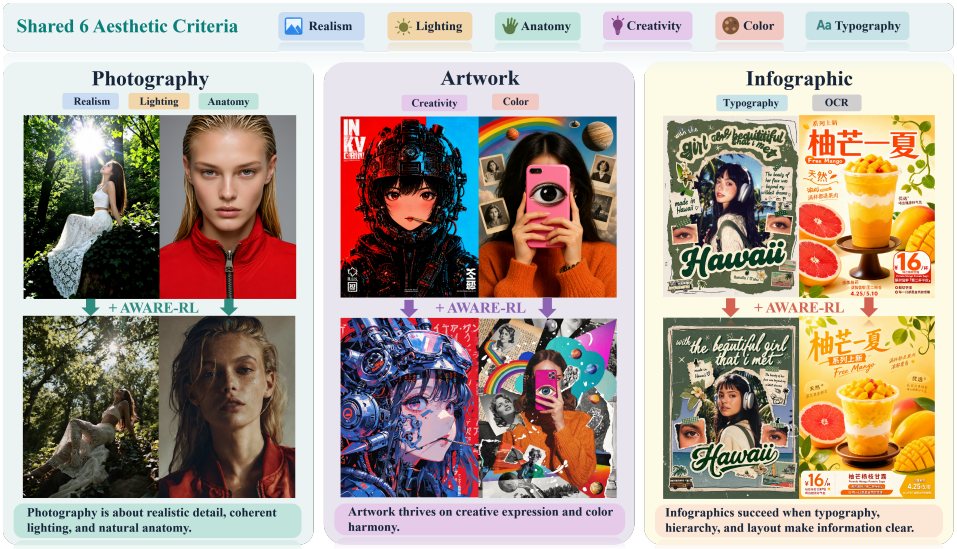

Figure 1: **AWARE-guided aesthetic alignment across image content.** We define six aesthetic criteria for diverse image content, and train AWARE-RL with criterion scores and applicability predictions from AWARE. The top row shows base-model results and the bottom row shows generations after AWARE-RL training. AWARE-RL improves photographic realism and lighting, artwork design and color harmony, and infographic typography.

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.

1 INTRODUCTION

Text-to-image (T2I) systems now generate photographs, artwork, posters, and information graphics in the same workflow. These outputs do not share a single notion of visual quality: photographs require faithful realism and natural lighting, artwork may prioritize creative composition and color harmony, and infographics require readable text and coherent layout. Most existing aesthetic or preference rewards nevertheless apply one global score or a fixed checklist to every image (Xu et al., 2023; Kirstain et al., 2023; Ma et al., 2025; Liu et al., 2026b). This conflates "not applicable" with "low quality": absent typography in a natural photograph should be ignored, whereas malformed typography in a poster should be penalized. It can also reward shortcuts that improve a metric while making an image unsuitable for its intended role.

We address this problem with content-conditional aesthetic judgment. We organize visual content into three broad roles: Photography, Artwork, and Infographic. We describe their quality with six reusable criteria: *realism*, *lighting*, *anatomy*, *creativity*, *color harmony*, and *typography*. The criteria provide a shared vocabulary across content types, while applicability indicates whether a criterion is meaningful for a particular image. This distinction allows a model to separate low quality from an irrelevant criterion. It also enables one reward model to share representations across criteria while selecting only the applicable signals for each image.

To learn these preferences, we construct an aesthetic preference corpus containing approximately 240K distinct images and 840K pairwise comparisons. The corpus combines curated web images with generated outputs and includes real-to-real, real-to-generated, and generated-to-generated comparisons across the three content roles. Human annotators and GPT-assisted judges rank candidate images using criterion-specific instructions and assign applicability labels. We use these annotations to train AWARE, an applicability-aware aesthetic reward model. Given an image and its prompt, AWARE jointly predicts a quality distribution and an applicability probability for all six criteria in one forward pass.

For T2I post-training, we propose applicability-aware reinforcement learning. AWARE predicts criterion-wise quality and applicability, and the RL objective dynamically weights only the criteria that apply to each sample. In our role-specific configurations, Photography emphasizes realism and lighting, Artwork combines creativity and color harmony, and Infographic combines text fidelity with typography. We evaluate the method on SenseNova-U1.5 (Diao et al., 2026a) and Z-Image (Z-Image Team et al., 2025). AWARE-RL improves Z-Image by 10.81%, 16.87%, and 1.36% on Photography, Artwork, and Infographic, respectively. The corresponding gains for SenseNova-U1.5 are 27.20%, 26.57%, and 3.41%.

To evaluate both the learned judgment and its effect on generation, we introduce two benchmarks. AWARE-RBench measures criterion-conditioned preference and per-image applicability. It also includes hard pairs and criterion-conflict pairs, where different criteria can favor different candidates. A content-agnostic reward that always produces one fixed score cannot represent such preference reversals. AWARE improves over the baseline by 60.92% on the aesthetic-model evaluation. AWARE-GenBench evaluates role-conditional T2I quality and tests whether a generator improves the properties that matter for Photography, Artwork, and Infographic content. The aligned models also remain competitive on general-purpose benchmarks, including Qwen-Image-Bench, GenEval, and DPG-Bench. Visual comparisons further show improvements in fidelity, composition, lighting, and text rendering.

Our contributions are: (1) a multi-source aesthetic preference corpus with explicit applicability labels and AWARE, a six-criterion reward model that jointly predicts quality and applicability; (2) AWARE-RBench and AWARE-GenBench for criterion-level judgment and role-conditional generation; and (3) an applicability-aware RL framework across two base generators.

2 RELATED WORK

VLM-as-a-Judge. General-purpose vision-language judges can compare prompt–image pairs and provide flexible rationales (OpenAI, 2023; Chen et al., 2024), but their cost, prompt sensitivity, and version dependence make them difficult to use as stable RL rewards. They also do not explicitly determine whether a criterion is relevant to the image's content role.

Aesthetic Reward Models. HPSv3 (Ma et al., 2025) learns a global human-preference reward, while HPSv3++ (Liu et al., 2026b) separates aesthetic preference from text fidelity. VisionReward (Xu et al., 2024) decomposes evaluation into perceptual questions but uses a fixed checklist and aggregation rule. ArtiMuse (Cao et al., 2025b) and UniPercept (Cao et al., 2025a) provide multi-attribute analysis, yet do not distinguish an irrelevant criterion from poor realization.

Quality and Alignment Evaluation. Q-Align and OneAlign (Wu et al., 2024) improve visual-quality calibration and alignment consistency, but do not condition criterion selection on content. AWARE instead predicts quality and applicability for every criterion, represents irrelevant dimensions as N/A, and exposes a content-adaptive interface for pairwise evaluation and RL. AWARE-RBench and AWARE-GenBench evaluate these capabilities at both judgment and generation time.

3 Criterion Taxonomy and Data Construction

3.1 Role–Criterion Taxonomy

We represent an evaluation instance as (p, x, r), where p is a prompt, x an image, and r its content role. For each criterion k, the annotation contains an applicability label $a_k \in \{0, 1\}$ and a quality judgment only when $a_k = 1$. An inactive criterion is recorded as N/A rather than as a zero-quality example, keeping applicability separate from quality.

Our taxonomy has three roles and six core dimensions. Photography emphasizes realism, lighting, anatomy, and color harmony. Artwork emphasizes creativity and color harmony. Infographic emphasizes text fidelity, typography, layout, and information clarity. The shared core dimensions are independently actionable and can be combined by role. The complete mapping and applicability triggers are given in Appendix A.1.

3.2 Preference Data Construction

To evaluate generalization across visual content, we organize images into three roles: Photography (real portraits and general photography), Artwork (artistic illustration and graphic design), and Infographic (low- or medium-density posters together with high-density information graphics). The corpus combines curated high-quality web images with outputs from 17 widely used image-generation systems spanning high-, medium-, and low-capability tiers. To avoid source-specific shortcuts, examples are paired as real–real, real–generated, and generated–generated comparisons, with varied styles, resolutions, text densities, object counts, and visual difficulty. Role subsets are filtered so that unrelated content does not dominate the supervision for a given role.

The final corpus contains approximately 240K distinct images. Each annotation task contains 5 candidates and is ranked under six criterion-specific rubrics, with a separate overall judgment when collected. Roughly 330K tasks are independently ranked by 5–9 human annotators. GPT-5.6-sol is used only as an auxiliary annotator: it is less reliable for anatomy, while for the other criteria each task is labeled three times and only directionally consistent outputs are retained. Human and GPT annotations cover all three pair types and preserve applicability/N/A decisions before group rankings are expanded into 840K pairwise preference examples.

We apply two quality-control stages. Annotator reliability is screened using about 1K manually verified items and agreement on a 5% repeated subset. Data reliability is checked by converting accepted human and unanimous GPT rankings into pairwise comparisons and retaining each comparison only when its preference direction reaches at least 80% agreement. After quality filtering and distribution balancing, the corpus contains approximately 840K preference pairs over 240K distinct images. For a concrete illustration of the visualization criteria used in our construction, see Appendix Figure 6.

4 AWARE Reward Model

4.1 Model Design

Given a prompt p and an image x, AWARE first encodes the prompt–image pair with a shared vision–language backbone. For every criterion $k \in \mathcal{K}$, a quality head predicts a mean score

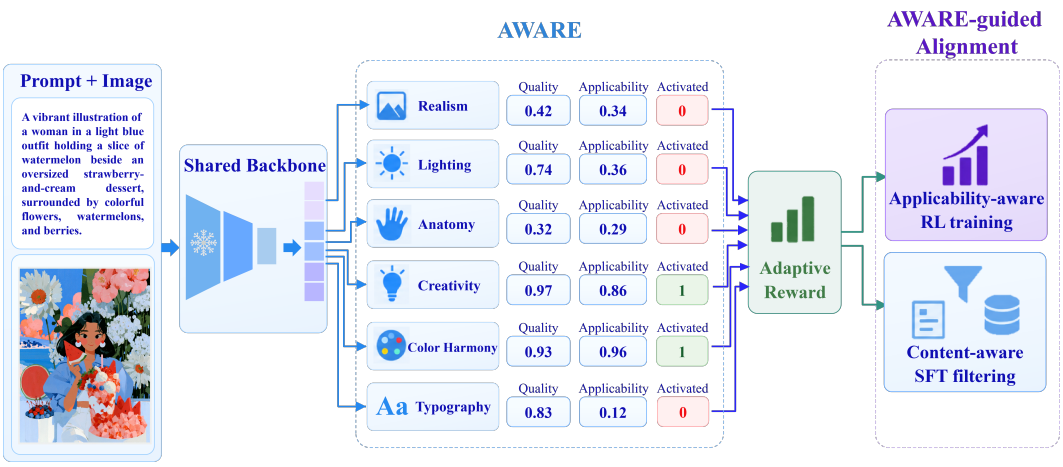

Figure 2: **Overview of AWARE and its downstream alignment interface.** A shared vision-language backbone predicts criterion-wise quality and applicability. Applicability-aware aggregation masks N/A criteria and provides a stable reward interface for applicability-aware reinforcement learning. An optional offline filtering interface is described in Appendix A.9.

$\hat{\mu}_k(p,x)$ and a positive uncertainty $\hat{\sigma}_k(p,x) > 0$. The scale output is mapped to a positive value before it is used as a standard deviation. We interpret the latent criterion quality as $u_k(p,x) \sim \mathcal{N}(\hat{\mu}_k(p,x), \hat{\sigma}_k^2(p,x))$, following the uncertainty-aware output parameterization of HPSv3 (Ma et al., 2025). An applicability head separately predicts $\hat{a}_k(p,x) \in [0,1]$. Thus, one forward pass produces a distribution for each criterion together with its applicability probability. A criterion marked N/A is represented by low applicability rather than by a zero-quality target, preserving the distinction between an image that fails a criterion and an image for which the criterion is irrelevant.

The shared backbone allows evidence useful for one criterion to be reused by the others, while the criterion-specific heads preserve separate signals for typography, color harmony, creativity, realism, lighting, and anatomy. Content roles select combinations of these heads rather than invoking unrelated models: infographic experts emphasize typography and text fidelity, artwork experts combine creativity with color harmony, and realistic-photo experts combine realism, lighting, anatomy, and color harmony. A criterion can therefore be shared across roles while remaining inactive when it is not meaningful for a particular image. Jointly training all criteria also lets them share representations, so supervision from one dimension can benefit the others and increase the effective amount of training data. During RL, a single forward pass returns all criterion rewards, avoiding the cost of running a separate reward model for each dimension.

4.2 Multi-Dimension Pairwise Training

Following the uncertainty-aware pairwise preference training setup of HPSv3 (Ma et al., 2025), we train AWARE from ranked prompt–image pairs instead of isolated scalar ratings. For two prompt–image pairs (p_i, x_i) and (p_j, x_j), the annotation for criterion k provides a binary preference label $y_{ij,k} \in \{0,1\}$ whenever that criterion is applicable to both prompt–image samples. When $y_{ij,k} = 1$, (p_i, x_i) is preferred. Let $m_{ij,k} \in \{0,1\}$ denote this validity mask. AWARE evaluates the two prompt–image pairs independently, producing $\mathcal{N}(\hat{\mu}_{i,k}, \hat{\sigma}_{i,k}^2)$ and $\mathcal{N}(\hat{\mu}_{j,k}, \hat{\sigma}_{j,k}^2)$, where $\hat{\mu}_{i,k} = \hat{\mu}_k(p_i, x_i)$ and $\hat{\mu}_{j,k} = \hat{\mu}_k(p_j, x_j)$ (and likewise for $\hat{\sigma}$). The probability that (p_i, x_i) is preferred under criterion k is

$$P_{ij,k} = \iint \mathrm{sigmoid}(u_i - u_j)\mathcal{N}(u_i \mid \hat{\mu}_{i,k}, \hat{\sigma}_{i,k}^2)\mathcal{N}(u_j \mid \hat{\mu}_{j,k}, \hat{\sigma}_{j,k}^2)\, du_i\, du_j. \tag{1}$$

The uncertainty-aware multi-dimension ranking objective is

$$\mathcal{L}_{\text{rank}} = -\frac{1}{\sum_{i,j,k} m_{ij,k}} \sum_{i,j,k} m_{ij,k} \left[y_{ij,k} \log P_{ij,k} + (1 - y_{ij,k}) \log(1 - P_{ij,k}) \right]. \tag{2}$$

The criterion index is retained throughout training rather than collapsing the annotations into a single holistic label. The predicted $\hat{\sigma}_{i,k}$ values allow uncertain or ambiguous comparisons to contribute less confidently, while the applicability mask excludes criteria that are not meaningful for the pair.

In parallel, each image has an applicability target $z_{i,k} \in \{0, 1\}$ for every criterion. We train the applicability heads with a binary cross-entropy objective,

$$\mathcal{L}_{\text{app}} = -\frac{1}{N|\mathcal{K}|} \sum_{i,k} \left[z_{i,k} \log \hat{a}_k(p_i, x_i) + (1 - z_{i,k}) \log \left(1 - \hat{a}_k(p_i, x_i)\right) \right]. \tag{3}$$

The final objective combines preference ranking and applicability prediction,

$$\mathcal{L} = \lambda_{\text{rank}} \mathcal{L}_{\text{rank}} + \lambda_{\text{app}} \mathcal{L}_{\text{app}}. \tag{4}$$

The pairwise preferences and applicability labels are obtained from the human and GPT-assisted, criterion-specific rankings described in Sec. 3. Training all heads jointly keeps the shared visual representation consistent while allowing each dimension to learn its own ordering. At inference time, AWARE processes one prompt–image pair and returns $(\hat{\mu}_k, \hat{\sigma}_k, \hat{a}_k)$ for every criterion. The mean $\hat{\mu}_k$ is used for criterion-level quality ranking, $\hat{\sigma}_k$ records prediction uncertainty, and $\hat{a}_k$ identifies whether the criterion is applicable.

5 AWARE-GUIDED ALIGNMENT

To improve a single generator across diverse image content, we propose applicability-aware RL for Photography, Artwork, and Infographic content. The method combines AWARE's six criterion-wise quality predictions with their applicability scores, so that each update emphasizes the dimensions relevant to the current content. An optional offline SFT filtering interface is described in Appendix A.9.

5.1 APPLICABILITY-AWARE RL

For each prompt p_q, the rollout policy generates G candidates $\mathcal{G}_q = \{x_q^j\}_{j=1}^G$. AWARE predicts criterion-wise quality $s_{q,j,k} = \hat{\mu}_k(p_q, x_q^j)$ and applicability $a_{q,j,k} = \hat{a}_k(p_q, x_q^j)$. Following GDPO (Liu et al., 2026a), we normalize each criterion within the prompt group to account for reward-scale differences:

$$A_{q,j,k}^{\text{GDPO}} = \frac{s_{q,j,k} - \mu_{q,k}}{\sigma_{q,k} + \epsilon}, \tag{5}$$

where $\mu_{q,k}$ and $\sigma_{q,k}$ are the group mean and sample standard deviation. We average applicability within the same group and enable criteria above $\tau_a = 0.5$:

$$\bar{a}_{q,k} = \frac{1}{G} \sum_{j=1}^{G} a_{q,j,k}, \qquad g_{q,k} = \mathbf{1}[\bar{a}_{q,k} > \tau_a]. \tag{6}$$

The group applicability also provides a soft weight for enabled criteria:

$$A_{q,j}^{\text{raw}} = \sum_{k \in \mathcal{K}} w_k \, g_{q,k} \, \bar{a}_{q,k} \, A_{q,j,k}^{\text{GDPO}}, \tag{7}$$

where $\mathcal{K}$ contains the six AWARE criteria and w_k are fixed base weights. Thus, candidates for the same prompt share criterion weights, while different prompts can emphasize different aesthetic dimensions.

We standardize the combined advantages across rollout images from prompt groups with at least one enabled, nonconstant criterion; all other groups receive zero advantage. Appendix A.8 gives the exact normalization and boundary cases. The resulting advantages enter a clipped GRPO objective with reference-policy KL regularization, following Flow-GRPO (Liu et al., 2025). Model-specific training settings are reported in Sec. 7.

6 BENCHMARKS

6.1 AWARE-RBENCH

AWARE-RBench is a held-out benchmark for criterion-conditioned pairwise judgment and per-image applicability across Photography, Artwork, and Infographic content. It covers six aesthetic



a **b** **c** **d**

Figure 3: Criterion-specific preference pairs for AWARE-RBench. Panels **a–c** show hard pairs where the left candidate is preferred under **Lighting**, **Realism**, and **Creativity**, respectively. Panel **d** shows a criterion-conflict pair: the left candidate is preferred for typographic aesthetics, whereas the right candidate is preferred for anatomical correctness.

criteria: realism, lighting, anatomy, creativity, color harmony, and typography, together with a separate overall aesthetic preference label.

Design and scale. The benchmark contains 3,500 images and 1,750 annotated image pairs, including real-real, real-generated, and generated-generated comparisons. Each image receives an applicability label for every criterion, indicating whether that criterion is meaningful for the image. Thus, an N/A label denotes irrelevance rather than low quality. Representative applicability decisions are provided in Appendix Figure 5. For each criterion applicable to both images, the corresponding pair also receives a preference label. We also construct hard pairs of visually similar candidates or controlled perturbations that require fine-grained judgment. For real-real pairs, professional artists curate semantically similar images that differ primarily in the target criterion, as illustrated by the lighting example in Figure 3**a**. For real-generated pairs, we use GPT-Image-2 to locally edit only the visual aspect associated with the target criterion. In Figure 3**b**, for example, the skin details of a real portrait are modified to produce an AI-like appearance, creating a controlled realism pair. For generated pairs, GPT-Image-2 generates two similar subjects from distinct prompts designed to differ along the target criterion. As shown in Figure 3**c**, the left image is more creative than the right. We also include criterion-conflict pairs whose annotated preference reverses when the active criterion changes. Figure 3**d** illustrates this case, with stronger typographic aesthetics on the left and more anatomically correct figures on the right. Together, these cases preserve applicability and criterion-specific judgments rather than reducing every comparison to one global preference.

Evaluation metrics. Pairwise preference accuracy measures agreement with the annotated preference, reported separately for each criterion and for overall aesthetic preference. Applicability is treated as binary classification: after thresholding the predicted probability at 0.5, accuracy is the primary metric reported per criterion.

6.2 AWARE-GenBench

Existing T2I benchmarks mainly test prompt adherence, compositional correctness, or text rendering, leaving dedicated evaluation of role-dependent aesthetic quality underdeveloped. AWARE-GenBench fills this gap with prompt-specific aesthetic requirements for Photography, Artwork, and Infographic content. It contains 600 prompts, evenly divided across the three roles, with 4,800 questions and fixed scoring rubrics.

The same questions and rubrics are applied to every generator. Each image is independently judged twice by `gpt-5.6-sol`. Answers are scored by rubric level, and the benchmark reports the percentage that reaches the highest level. Photography focuses on realism and lighting, with anatomy questions for portrait prompts. Artwork focuses on creativity and color harmony. Infographic focuses on text fidelity, typography, layout, and information organization. The exact aggregation formula and treatment of unassessable answers are given in Appendix A.5. Representative prompts appear in Appendix A.12 (Table 9), and generation comparisons are reported in Section 7.

7 EXPERIMENTS

7.1 IMPLEMENTATION DETAILS

Table 1: **Pairwise preference accuracy (%) on the AWARE-RBench.** Green and light green cell backgrounds denote the best and second-best results within each model group, respectively. Human is included as a reference and is not part of the ranking. Random choice is 50%.

Models	Realism	Lighting	Anatomy	Creativity	Color Harmony	Typography	**Overall**
Human	89.66	88.41	96.00	97.14	88.46	100.00	91.71
Gemini-Flash-3.7	85.19	80.93	42.33	87.41	61.15	99.34	77.64
Grok 4.6	75.50	73.64	30.67	86.72	70.86	97.37	73.66
Claude-Opus-5.0	88.03	78.88	52.76	86.35	78.06	98.03	80.97
GPT-5.6-Sol	88.32	85.42	37.42	89.30	68.35	98.03	80.51
GPT-6-Astra	97.15	85.42	71.17	89.30	80.94	92.11	86.91
HPSv3 (Ma et al., 2025)	45.30	65.79	33.13	50.92	43.17	84.87	54.40
HPSv3++ (Liu et al., 2026b)	35.61	76.26	26.38	31.73	39.93	75.00	50.69
ImageReward (Xu et al., 2023)	34.76	63.18	46.63	44.65	42.09	70.39	50.34
PickScore (Kirstain et al., 2023)	33.90	58.13	34.97	50.55	53.24	73.68	50.51
ArtiMuse (Cao et al., 2025b)	74.07	51.21	67.48	73.80	68.35	67.76	64.97
UniPercept (Cao et al., 2025a)	45.30	49.35	35.58	64.21	64.39	63.82	53.20
AWARE (Ours)	91.17	87.66	87.73	88.93	75.18	98.68	87.54

Aesthetic Reward Model. We use Qwen2-VL-7B (Wang et al., 2024) as the vision-language backbone and initialize AWARE from HPSv3 (Ma et al., 2025). We replicate the HPSv3 reward head into six criterion-specific score-and-uncertainty heads and jointly train them with the applicability head using the masked pairwise ranking and applicability-classification objectives described in Sec. 4. We freeze the vision encoder and pretrained language-model parameters, and optimize the LoRA adapters, visual merger, and reward heads. See Appendix A.2 for full reward-model training details.

RL Training Details. We select SenseNova-U1.5 (Diao et al., 2026a) and Z-Image (Z-Image Team et al., 2025) as complementary open-source bases, strongest on Infographic and Photography content, respectively. We apply online RL to these two base generators, using FlowGRPO (Liu et al., 2025) and Flow-Factory (Ping et al., 2026), respectively. Role-specific reward combinations follow Sec. 5, with an additional OCR reward for Infographic prompts. Rollout, optimization, and OCR implementation details are provided in Appendix A.2.

7.2 AESTHETIC REWARD MODEL EVALUATION

We compare AWARE with proprietary multimodal judges and existing aesthetic reward models using the AWARE-RBench metrics defined in Sec. 6.1.

Evaluation protocol. Proprietary models use medium thinking during inference and the same concise prompt. Under the Hard protocol, each pair is presented in three orders (AB, BA, and a random order) and is counted as correct only if all three predictions agree with the ground truth.

Pairwise preference evaluation. Table 1 reports AWARE-RBench accuracy by realism, lighting, anatomy, creativity, color harmony, and typography, together with averaged overall score. The full comparison with established aesthetic benchmarks is provided in Appendix A.6. AWARE is initialized from HPSv3 Ma et al. (2025), their comparison most directly reflects the improvement introduced by our criterion-specific, applicability-aware training. HPSv3 obtains only 54.40% overall accuracy on AWARE-RBench, merely 4.40 percentage points above random choice. In contrast, AWARE reaches 87.54%, improving over HPSv3 by 60.92%. AWARE also surpasses the recent strong baseline ArtiMuse Cao et al. (2025b) by 22.57 points.

On previous benchmarks such as HPDv3, AWARE trails HPSv3, while even proprietary judges achieve only moderate accuracy. Our visual audit suggests that many legacy pairs are generated with older image models and force a choice between two visibly weak candidates, making them low-margin tests. AWARE-RBench instead uses curated, quality-controlled real and generated images to provide more informative comparisons.

Table 2: **AWARE-GenBench results and expert specialization.** Each block compares a base generator, three role-specific experts, and unified AWARE-RL. Blue denotes AWARE-RL's relative percentage gains over the corresponding base model. Green and light green mark the best and second-best scores per column within each block.

Models	Photography	Artwork	Infographic
Z-Image (Z-Image Team et al., 2025)	60.69	61.13	39.79
Photography expert	69.34	68.94	40.96
Artwork expert	62.13	71.97	39.25
Infographics expert	59.97	68.66	43.38
Z-Image + AWARE-RL	67.25 (+10.81%)	71.44 (+16.87%)	40.33 (+1.36%)
SenseNova-U1.5-SFT (Diao et al., 2026a)	46.88	57.13	86.71
Photography expert	68.25	66.56	88.79
Artwork expert	43.91	77.53	83.29
Infographics expert	56.84	68.25	90.42
SenseNova-U1.5-SFT + AWARE-RL	59.63 (+27.20%)	72.31 (+26.57%)	89.67 (+3.41%)

Applicability evaluation. The overall mean applicability accuracy is 97.29%. By content type, it is 97.27% for Photography, 94.61% for Artwork, and 100.00% for Infographic. These results show that AWARE accurately identifies applicable criteria across content types. The full criterion-wise breakdown is provided in Appendix A.10.

7.3 RL Model Evaluation and Ablation

Comparison with base generators. Table 2 compares each base generator with role-specific experts and unified AWARE-RL. With applicability-aware RL training, both generators improves its aesthetic quality across all three image contents, yielding overall gains of 5.80 points for Z-Image and 10.30 points for SenseNova-U1.5-SFT. Figure 4 complements these results with examples of more natural photographic lighting and texture, richer artistic composition, and improved integration of typography with visual content. Refer to Appendix(Figure 9) for more visualization.

Applicability and expert specialization. Table 2 compares unified AWARE-RL with experts trained on carefully designed, role-specific data. Each expert is trained exclusively on data from its target content type and does not use an applicability score, providing a specialization-only reference. Each expert performs best in its target role, while unified AWARE-RL improves all three roles on both generators and ranks second in five of the six comparisons. Unified Z-Image trails the Artwork expert by only 0.53 points, and unified SenseNova-U1.5-SFT trails the Infographic expert by 0.75 points. These results demonstrate that predicted applicability enables a single generator to capture substantial benefits of specialization across heterogeneous content, providing clear evidence of its effectiveness. Z-Image's Infographic gain is more modest. We directly reuse the SenseNova-U1.5 alignment hyperparameters for Z-Image, which may be suboptimal given their different infographic capabilities. Model-specific tuning of the applicability threshold τ_a may further improve Infographic performance.

Comparison with SOTA. Beyond AWARE-GenBench (AGB), we evaluate our aligned models on Qwen-Image-Bench (Li et al., 2026), GenEval (Ghosh et al., 2023), and DPG-Bench (Hu et al., 2024). AWARE-RL achieves the best overall AGB score among the evaluated open-source models while remaining competitive on these general-purpose benchmarks, supporting its overall generation performance. Full results are provided in Appendix A.7 (Table 5).

Single versus multiple rewards. Figure 8 in Appendix A.11 illustrates the complementary roles of aesthetic rewards. Adding Realism to Lighting encourages natural illumination and realistic texture. For text-rich images, OCR targets text correctness, while Typography additionally encourages coherent lettering and integration with the scene. The quantitative results in Appendix A.11 (Table 7) show a substantial Photography gain from combining rewards, while the OCR–Typography

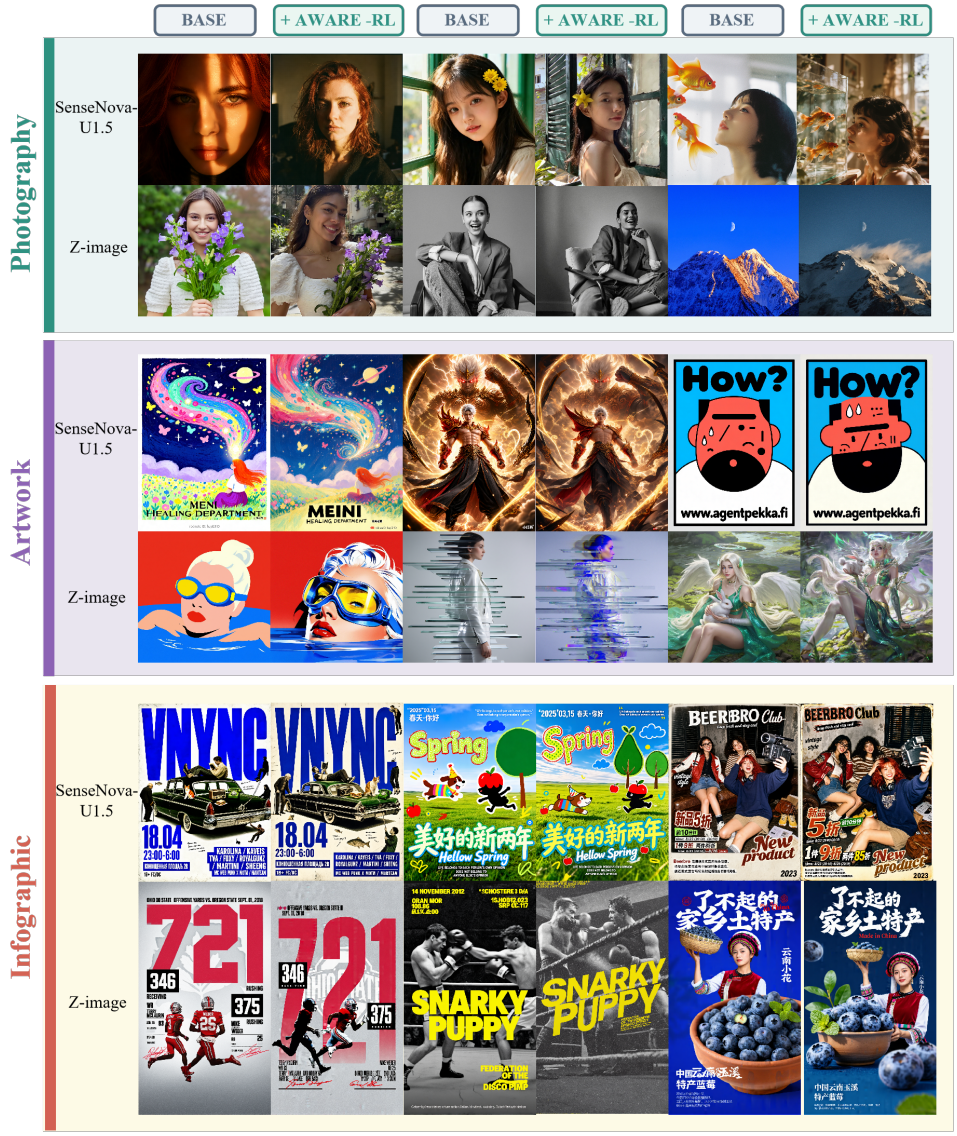

Figure 4: **Qualitative comparison before and after AWARE-guided RL training.** Qualitative comparison across Photography, Artwork, and Infographic content shows that AWARE-RL consistently improves visual quality over the corresponding base models for both Z-Image and SenseNova-U1.5.

combination yields a slightly lower aggregate Infographic score. These results highlight the value of complementary criteria and the need to balance them for each task.

8 Conclusion

We presented AWARE, an aesthetic reward model trained on criterion-specific preferences and explicit applicability labels to distinguish low quality from irrelevant criteria. AWARE-RBench and AWARE-GenBench evaluate criterion-level judgment and role-conditional generation, while experiments on SenseNova-U1.5 and Z-Image demonstrate the utility of AWARE-guided RL reward optimization. Beyond T2I generation, the same approach could support training for image editing and interleaved text–image generation by conditioning aesthetic criteria on editing instructions and multimodal context.

AI USE STATEMENT

Generative AI tools were used to assist with identifying and summarizing related work, as well as polishing and editing the manuscript. In addition, generative image models, including GPT-Image-2, were used to generate portions of the experimental data, particularly controllable data designed for our experiments. GPT-based vision-language models were also used to assist with image annotation and the assessment of aesthetic quality. These tools served as research aids rather than autonomous decision-makers: the authors specified the data-generation and evaluation protocols, reviewed and filtered AI-generated samples and annotations, and made all final decisions regarding the experimental design, interpretation, and presentation of results. AI-assisted literature summaries, manuscript revisions, generated data, annotations, and aesthetic assessments were subject to author review and quality control using primary sources, task-specific criteria, manual inspection, and reproducible evaluation procedures, as appropriate. The authors take full responsibility for the complete content of this work, including all text, claims, data, annotations, results, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

Our work focuses on aesthetic image analysis and generation, while recognizing that the visual data and outputs may contain human faces. The training corpus consists of copyrighted images used under applicable licenses, permissions, or other lawful bases. All data and results intended for external release—including datasets, figures, examples, and model outputs—undergo manual review for potential privacy, copyright, consent, and misuse concerns. Samples that do not pass this review are excluded, replaced, or withheld from release. We do not seek to identify individuals, infer sensitive attributes, or enable surveillance; our objective is to study and improve visual aesthetics while respecting the privacy and dignity of depicted persons. If related models or datasets are released in the future, we will consider incorporating provenance-tracking and imperceptible watermarking mechanisms, such as SynthID-like techniques, to help trace the origin and transformation history of facial imagery and generated outputs. These measures will be implemented in accordance with applicable laws, licenses, and institutional policies.

REPRODUCIBILITY STATEMENT

The paper organizes the information needed to reproduce AWARE and its alignment experiments across the main text and Appendix A. Section 3 describes the task formulation, role–criterion taxonomy, and construction of the preference corpus. Section 4 specifies the reward-model architecture and multi-dimension training objective. Sections 5–7 describe the alignment procedure, benchmark protocols, and ablations. Appendix A provides the reward-model and RL training configurations, detailed taxonomy, data sources and filtering rules, annotation and split procedures, AWARE-GenBench scoring rule, evaluation protocol, and representative prompts and outputs. The code, pretrained AWARE weights, benchmark prompts and rubrics, and releasable portions of the annotated corpus will accompany the paper as reproducibility artifacts. The release will also include configuration files and data manifests needed to regenerate the reported tables and figures.

REFERENCES

ByteDance Seed. Seedream 5.0 Pro. Official model page, 2026. URL https://seed.bytedance.com/en/seedream5_0_pro.

Shuo Cao, Jiayang Li, Xiaohui Li, Yuandong Pu, Kaiwen Zhu, Yuanting Gao, Siqi Luo, Yi Xin, Qi Qin, Yu Zhou, Xiangyu Chen, Wenlong Zhang, Bin Fu, Yu Qiao, and Yihao Liu. UniPercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture. *arXiv preprint arXiv:2512.21675*, 2025a.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajuin Yao, and Yihao Liu. ArtiMuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*, 2025b.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, Yue Zhu, Yuwei Niu, Zhengqi Bai, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Bo Yang, Chen Feng, Chengguang Lv, Guangjia Liu, Guanlin Wang, Hanyu Zhang, Haojia Yu, Hongcan Xiao, Hongli Wang, Huan Wu, Huaping Zhong, Jian Fang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jing Zuo, Jingcheng Ni, Junxiang Xu, Linjun Dai, Mutian Xu, Peishen Yan, Penghao Wu, Ruijie Mao, Ruisi Wang, Shihao Bai, Shuang Yang, Shuya Yang, Shuyan Zheng, Silei Wu, Siying Li, Tao Chu, Tianbo Zhong, Tongxi Zhou, Weichao Luo, Weichen Fan, Wenhao Jia, Wenjie Gao, Xiangli Kong, Yan Li, Yang Yong, Zimo Wen, Zixuan Qian, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin. SenseNova-U1.5: Towards native unified visual intelligence. *arXiv preprint arXiv:2609.11929*, 2026a. URL https://arxiv.org/abs/2609.11929.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, Yan Li, Yubo Wang, Zhijie Cao, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Yuwei Niu, Yue Zhu, Bo Liu, Chengguang Lv, Haojia Yu, Haozhe Xie, Hongli Wang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jingcheng Ni, Junxiang Xu, Kaihuan Liang, Lianqiang Shi, Linjun Dai, Linyan Wang, Oscar Qian, Peng Gao, Pengfei Liu, Qingping Sun, Rui Shen, Ruisi Wang, Shengnan Ma, Shuang Yang, Siyi Xie, Siying Li, Tianbo Zhong, Xiangli Kong, Xuanke Shi, Yang Gao, Yongqiang Yao, Yves Wang, Zhengqi Bai, Zhengyu Lin, Zixin Yin, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin. SenseNova-U1: Unifying multimodal understanding and generation with NEO-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026b. URL https://arxiv.org/abs/2605.12500.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. *arXiv preprint arXiv:2310.11513*, 2023.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. URL https://arxiv.org/abs/2403.05135.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, Ziyi He, Wei Wang, Dalin Li, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyue Chen, Yuxiang Chen, Yan Shu, Yanran Zhang, Yilei Chen, Yixian Xu, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, Hongzhu Shi, Yi Wang, Bing Zhao, Hu Wei, Lin Qu, and Chenfei Wu. Qwen-Image-Bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, 2026. URL https://arxiv.org/abs/2605.28091.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. *arXiv preprint arXiv:2505.05470*, 2025.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization. *arXiv preprint arXiv:2601.05242*, 2026a. URL https://arxiv.org/abs/2601.05242.

Yijun Liu, Jie Huang, Zeyue Xue, Yuming Li, Ruizhe He, Haoran Li, Shijia Ge, and Siming Fu. HPSv3++: Scaling reward models across the full spectrum of diffusion model capabilities. *arXiv preprint arXiv:2606.14657*, 2026b.

Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. *arXiv preprint arXiv:2508.03789*, 2025.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. GPT-Image-2. Official model documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-image-2.

Bowen Ping, Chengyou Jia, Minnan Luo, Hangwei Qian, and Ivor Tsang. Flow-Factory: A unified framework for reinforcement learning in flow-matching models. *arXiv preprint arXiv:2602.12529*, 2026. URL https://arxiv.org/abs/2602.12529.

Qwen Team. Qwen-Image-2.1: Compact, Efficient, and Unified Image Creation. Official model announcement, 2026a. URL https://qwen.ai/blog?id=qwen-image-2.1.

Qwen Team. Qwen-Image-3.0: Rich Content, Authentic Details, Deep Knowledge. Official model announcement, 2026b. URL https://qwen.ai/blog?id=qwen-image-3.0.

Naina Raisinghani. Nano Banana 2: Combining Pro capabilities with lightning-fast speed. Google official announcement, 2026. URL https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/.

Feng Wang and Zihao Yu. Coefficients-preserving sampling for reinforcement learning with flow matching. *arXiv preprint arXiv:2509.05952*, 2025. URL https://arxiv.org/abs/2509.05952.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Ming Ding, Xiaohan Zhang, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Shijie Huang, Zhaohui Hou, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025. URL https://arxiv.org/abs/2511.22699.

A Additional Details

A.1 Role-Conditioned Taxonomy

The main paper summarizes the taxonomy by listing the three content roles and six core dimensions. The table below gives the full mapping used for data collection and evaluation. A criterion can be N/A even within a role when its applicability trigger is absent.

Table 3: Role-conditioned taxonomy used for data collection and evaluation. A criterion can be N/A even within a role when its applicability trigger is absent.

Role	Primary criteria	Applicability and evaluation focus
Photography	realism, lighting, anatomy, color harmony, composition, mood	Realism and lighting are active for photographic scenes; anatomy is active when a person or animal is visible; composition, mood, and color harmony assess the image as a photograph.
Artwork	creativity, color harmony, composition, mood, style coherence	Creativity and style coherence capture originality and medium consistency; composition, mood, and color harmony assess the visual organization and emotional effect.
Infographic	text fidelity, typography, layout, information clarity, color harmony, creativity	Text fidelity and typography are active when rendered text is present; layout and information clarity measure hierarchy and readability; color harmony and creativity assess visual communication.

The six core dimensions are realism, lighting, anatomy, creativity, color harmony, and typography. Realism measures photographic plausibility and the absence of synthetic artifacts. Lighting measures exposure, direction, softness, and coherence with the scene. Anatomy measures the structural correctness of visible human or animal bodies. Creativity measures originality and the degree to which the result goes beyond a literal or repetitive rendering. Color harmony measures palette coherence, contrast, and saturation. Typography measures the visual quality of rendered text, including type hierarchy, spacing, legibility, and integration with the composition. Composition, mood, style coherence, text fidelity, layout, and information clarity are role-specific extensions used when required by the content.

Figure 7 summarizes the AWARE-RBench comparison across the reported criteria, including AWARE, HPSv3 Ma et al. (2025), and GPT-based evaluators.

A.2 Reward-Model and RL Training Details

Aesthetic reward model. To balance prediction quality and computational efficiency, we adopt Qwen2-VL-7B (Wang et al., 2024) as the vision-language backbone and initialize the model from



Figure 5: **Applicability examples in AWARE-RBench.** Panels **a–c** illustrate per-image applicability decisions across content types and criteria. An N/A label indicates that a criterion is irrelevant to the image, rather than that the image has low quality under that criterion.

the original HPSv3 checkpoint. For joint training of the six-dimensional reward model, the single reward head of HPSv3 is replicated to initialize six independent dimension-specific heads. We freeze the vision encoder and the pretrained language-model parameters, while optimizing the LoRA adapters, visual merger, and six score-and-uncertainty prediction heads. The LoRA rank, scaling factor, and dropout rate are set to 64, 128, and 0.05, respectively, resulting in approximately 228M trainable parameters, or 2.69% of the full model. We employ a cosine learning-rate schedule with linear warm-up over the first 3% of training steps. The peak learning rates are 2×10^{-5} for the LoRA adapters and visual merger and 1×10^{-4} for the reward heads, with a weight decay of 0.01. Training is conducted on 64 GPUs with a global batch size of 64. The model largely converges within 2 epochs and achieves its best validation performance at approximately epoch 3.

RL training. We perform online RL on SenseNova-U1.5 (Diao et al., 2026a) and Z-Image (Z-Image Team et al., 2025). For expert training, we use 10,000 prompts per content role (Photography, Artwork, and Infographic). The Photography subset contains 5,000 portrait and 5,000 non-portrait prompts. SenseNova-U1.5 uses FlowGRPO (Liu et al., 2025) with 16 candidates per prompt and 30-step CPS trajectories (Wang & Yu, 2025). We train for 400 iterations at learning rate 2×10^{-5} with reference-policy KL coefficient 0.01, using classifier-free guidance and conditional-only gradients. Z-Image uses Flow-Factory (Ping et al., 2026) with LoRA under the same fixed budget. Role-specific reward combinations follow Sec. 5. For Infographic prompts, the additional OCR reward computes multiset IoU between normalized target and OCR-recognized text, tokenized as English words and Chinese characters, penalizing missing, extra, and repeated tokens without relying on reading order. Although proprietary VLMs perform strongly, one GPT-5.6-sol low-thinking reward-model run costs over $1,000, making repeated training impractical.

A.3 Training Data Construction and Annotation

The corpus contains approximately 240K distinct images and 840K pairwise preference examples after source, quality, and distribution screening. It combines curated high-quality web images with outputs from 17 widely used image-generation systems grouped into high-, medium-, and low-capability tiers. Candidate groups cover Photography, Artwork, and Infographic content and are constructed as real–real, real–generated, and generated–generated comparisons. They include controlled seed variation, cross-model capability contrasts, model–real-image mixtures, and a real-image subset, together with variation in source model, checkpoint, sampling configuration, resolution, style, text density, object count, and visual difficulty. Human annotators rank each group under criterion-specific instructions. GPT-5.6-sol provides auxiliary labels for non-anatomy criteria, with three repeated judgments and unanimous-direction filtering. Splits are performed at the prompt, source, and concept levels to prevent correlated candidates from appearing in both training and evaluation.

For every criterion, annotators first decide whether the criterion is applicable and then score or rank quality only when it is applicable. N/A is a valid semantic label and is never converted into a zero score. Overall aesthetic is a separate holistic judgment, while subdimension scores are independent and have no fixed weights in the final overall score. The annotation instructions emphasize criterion-specific evidence, comparison within a group, and consistency across repeated items. Ambiguous or low-agreement items are adjudicated, while hard negatives and same-prompt rollouts are retained to improve the model's ability to distinguish close alternatives. We will provide the final annotation interface, condensed scoring rubric, GPT prompts, agreement statistics, adjudication rate, and source/split audit in the camera-ready version.

Because each group is annotated under multiple criteria, the same candidate pair can contribute different labels across dimensions. These cross-dimension labels are retained instead of being averaged to preserve criterion-specific preferences during training. The six core AWARE dimensions are selected because they provide direct, actionable signals for T2I alignment. For example, portrait or camera realism is marked N/A for non-photorealistic illustrations, typography and text fidelity are active when rendered text is present, and high-density infographics can omit portrait realism while low- or medium-density infographics retain broader visual criteria.

We repeat a subset of groups, include agreement checks and adjudication for ambiguous cases, and mine visually similar rollouts to improve within-prompt discrimination. Random pair-level splitting is avoided because pairs from one group are correlated. We report role and criterion distributions,

applicable/N/A ratios, the human and GPT-assisted fraction, repeated and adjudicated item counts, and split counts. These quantities are kept as editable commands in the main file so that the paper can be updated after the final data audit.

A.4 Taxonomy and Benchmark Statistics

We report role and criterion counts, applicable/N/A ratios, group sizes, pair counts, source-model distribution, prompt complexity, and train/validation/test sizes. Benchmark statistics include the number of evaluated pairs and images for AWARE-RBench and the number of prompts, systems, candidates, and human comparisons for AWARE-GenBench.

A.5 AWARE-GenBench Scoring Protocol

The benchmark translates broad aesthetic requirements into observable questions, such as coherent surface details, structured color palettes, and accurate text–layout relationships. Every generated image is independently evaluated twice by `gpt-5.6-sol` with medium reasoning effort. The judge receives only the image, its prompt, and the fixed questions, and assigns each answer a rubric-defined level from 0 to 3 or N/A. For role r, let $\mathcal{D}_r$ denote its dimensions, N_r its image count, and $\mathcal{Q}_{i,d}$ the questions for image i under dimension d. Given responses $\ell_{i,q}^{(t)} \in \{0, 1, 2, 3, \mathrm{N/A}\}$, the role score is

$$S_r = \frac{100}{TN_r|\mathcal{D}_r|} \sum_{t=1}^{T} \sum_{i=1}^{N_r} \sum_{d \in \mathcal{D}_r} \frac{1}{|\mathcal{Q}_{i,d}|} \sum_{q \in \mathcal{Q}_{i,d}} \mathbf{1}\left[\ell_{i,q}^{(t)} = 3\right], \qquad T = 2. \tag{8}$$

Photography aggregates realism and lighting, with anatomy included for portrait prompts. Artwork aggregates color harmony and creativity. Infographic aggregates text fidelity, typography, layout, and information organization. Unassessable answers remain in the denominator and contribute zero to S_r. Counting them as successes gives the upper bound.

A.6 Cross-Benchmark Comparison

The complete comparison with established aesthetic preference benchmarks is given in Table 4. These results are reported for context. The main analysis focuses on the criterion-wise AWARE-RBench results in Table 1. A visual audit of legacy benchmark pairs suggests that many use outputs from older generation systems and require a choice between two weak candidates, which limits their diagnostic value.

Table 4: **Pairwise preference accuracy (%) across multiple aesthetic benchmarks.** The table reports results on AWARE-RBench and established preference benchmarks, including HPDv3, HPDv3++, ImageReward, and PickScore. In the HPDv3++ column, the two values denote aesthetic and prompt-alignment accuracy, respectively. Schematic benchmark cases appear in Appendix A.12.

Models	AWARE-RBench ↑	HPDv3 ↑	HPDv3++ ↑	ImageReward ↑	PickScore ↑
Human	91.71	67.46	54.01/-	59.29	57.14
Gemini-Flash-3.7	77.64	50.18	37.79/50.38	43.05	50.63
Claude-Opus-5.0	80.97	48.82	43.72/52.32	43.21	50.61
GPT-5.6-sol	80.51	55.52	46.84/55.57	48.32	51.78
GPT-Astra	86.91	59.33	51.15/60.24	53.62	50.94
HPSv3 (Ma et al., 2025)	54.40	76.13	71.17/72.16	67.22	69.65
HPSv3++ (Liu et al., 2026b)	50.69	76.39	74.74/79.33	68.44	62.83
ImageReward (Xu et al., 2023)	50.34	58.74	58.57/63.56	65.42	59.25
PickScore (Kirstain et al., 2023)	50.51	65.67	57.83/63.09	62.59	74.21
ArtiMuse (Cao et al., 2025b)	64.97	57.09	54.79/54.58	53.19	56.6
UniPercept (Cao et al., 2025a)	53.20	56.07	54.27/55.34	52.02	55.86
AWARE (Ours)	87.54	69.56	62.15/66.27	58.57	64.62

Table 5: **Comparison with representative text-to-image generators.** AGB (AWARE-GenBench) reports Photography, Artwork, Infographic, and their mean. QIB (Qwen-Image-Bench (Li et al., 2026)) reports English (EN), Chinese (ZH), and overall scores. Green/light green indicate the best/second-best scores per column within closed-/open-source groups (Ours: open-source).

Model	AGB				QIB			GenEval	DPG
	Photo	Art	Info	Overall	EN	ZH	Overall		
Closed-source Models									
GPT-Image-2 (OpenAI, 2026)	69.66	85.00	95.71	83.46	65.23	64.69	64.96	89.00	85.98
Nano-Banana-2.0 (Raisinghani, 2026)	74.38	85.53	90.78	83.56	59.59	59.82	59.71	83.00	86.90
Qwen-Image-3.0-Pro (Qwen Team, 2026b)	76.00	80.38	89.97	82.12	62.30	61.39	61.85	87.23	85.93
Seedream-5.0-Pro (ByteDance Seed, 2026)	73.15	82.91	89.55	81.87	59.80	59.66	59.73	86.04	87.02
Open-source Models									
Qwen-Image-2.1 (Qwen Team, 2026a)	64.09	63.81	86.04	71.31	54.28	54.80	54.54	50.46	84.72
Z-Image (Z-Image Team et al., 2025)	60.69	61.13	39.79	53.87	48.79	49.43	49.11	73.37	87.41
Z-Image-Turbo (Z-Image Team et al., 2025)	64.16	56.50	32.75	51.14	46.65	46.91	46.78	82.00	84.86
SenseNova-U1 (Diao et al., 2026b)	37.16	53.00	77.96	56.04	48.28	45.99	47.14	90.96	87.78
SenseNova-U1.5-SFT (Diao et al., 2026a)	46.88	57.13	86.71	63.57	52.20	51.82	52.01	92.86	87.68
Z-Image + AWARE-RL (Ours)	67.25	71.44	40.33	59.67	48.11	49.15	48.63	68.38	89.19
SenseNova-U1.5-SFT + AWARE-RL (Ours)	59.63	72.31	89.67	73.87	53.49	53.34	53.42	90.10	87.48

A.7 Comparison with Other Generators

Table 5 compares our aligned models with representative generators on AWARE-GenBench (AGB), Qwen-Image-Bench (QIB) (Li et al., 2026), GenEval (Ghosh et al., 2023), and DPG-Bench (Hu et al., 2024).

AWARE-RL improves all three AGB content roles on both base generators, yielding overall gains of 5.80 points for Z-Image and 10.30 points for SenseNova-U1.5-SFT. Among the listed open-source models, aligned SenseNova-U1.5-SFT ranks first in AGB overall, Artwork, and Infographic, while aligned Z-Image leads on Photography.

Beyond AGB, aligned SenseNova-U1.5-SFT improves QIB scores in both English and Chinese, while aligned Z-Image achieves the highest DPG-Bench score among all listed models. Transfer is selective: GenEval decreases for both models, and QIB decreases slightly for Z-Image.

A.8 Advantage Normalization for Applicability-aware RL

After the criterion-wise normalization and applicability-weighted aggregation in Sec. 5, we standardize the combined advantages from Eq. 7. Let $\mathcal{B}_+$ contain the rollout images whose prompt groups have at least one enabled criterion with nonzero quality variance. Denoting the mean and sample standard deviation of A^{raw} over $\mathcal{B}_+$ by μ_+ and σ_+, respectively, we compute

$$\hat{A}_{q,j} = \begin{cases} \dfrac{A^{\text{raw}}_{q,j} - \mu_+}{\sigma_+ + \epsilon}, & (q,j) \in \mathcal{B}_+, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Constant criterion scores contribute zero advantage during within-group normalization. Prompt groups with no enabled criterion of nonzero quality variance are excluded from the final normalization and receive zero advantage. If $\mathcal{B}_+$ is empty, all advantages are set to zero without computing μ_+ or σ_+. The resulting $\hat{A}_{q,j}$ is used in the clipped GRPO objective with reference-policy KL regularization described in the main text.

A.9 Optional AWARE-guided SFT Filtering

For offline supervised fine-tuning data selection, an image–prompt pair can first be assigned a content role, after which only the criteria active for that role are considered. A pair is retained when its applicable AWARE quality scores satisfy a selected quality threshold. N/A criteria are excluded rather than treated as failures. This interface preserves the distinction between an image that is poor on a relevant criterion and one for which the criterion is irrelevant. We do not report an independent AWARE-SFT training condition in this paper. The main experiments use AWARE for applicability-aware RL.

A.10 Evaluation Protocol and Reproducibility

AWARE-RBench reports criterion-wise and overall pairwise preference accuracy, with the Hard protocol for proprietary models specified in Sec. 7.2. Applicability is a binary decision evaluated per image after thresholding at 0.5. Criterion-wise accuracy is the primary metric, while the additional statistics in Table 6 diagnose class imbalance. The alignment experiments report unified applicability variants and the single-versus-multiple reward ablation described in Section 7. Full prompts, annotation templates, model hyperparameters, aggregation formulas, and additional failure cases will be included with the final release. The full content-type and criterion breakdown is given in Table 6.

Table 6: Applicability classification accuracy (%) on AWARE-RBench, grouped by content type and criterion.

Image Type	**Realism**	**Lighting**	**Anatomy**	**Creativity**	**Color Harmony**	**Typography**
Photography	98.79	98.79	88.38	98.88	98.97	99.82
Artwork	93.81	93.90	96.99	92.44	92.35	98.18
Infographic	100.00	100.00	100.00	100.00	100.00	100.00

A.11 Reward Ablation Details

Figure 8 illustrates the visual effects of combining rewards. Adding Realism to Lighting encourages natural illumination and realistic texture, while adding Typography to OCR encourages coherent lettering and integration with the scene.

Table 7 compares single and multiple rewards on Z-Image under the same RL budget. Adding Realism to Lighting increases the Photography score from 58.12 to 75.22, a gain of 17.10 points. Adding Typography to OCR decreases the Infographic score from 47.83 to 46.42. Thus, complementary rewards improve Photography in this configuration, while balancing text correctness and typographic aesthetics remains an opportunity for further optimization.

Figure 5 provides representative image-level applicability decisions. For each image and criterion, annotators first determine whether the criterion is applicable. An N/A label means that the criterion is irrelevant to the image and is not treated as a low-quality judgment.



Figure 8: Complementary rewards. Lighting alone causes overexposure, while adding Realism encourages natural illumination. OCR improves text correctness, while adding Typography preserves textual aesthetic.

Table 7: Single versus multiple rewards on Z-Image. AGB scores on each subset; all variants use the same RL budget.

AGB subset	Reward configuration	AGB ↑
Photography	+ Lighting	58.12
	+ Lighting + Realism	75.22
Infographic	+ OCR	47.83
	+ OCR + Typography	46.42

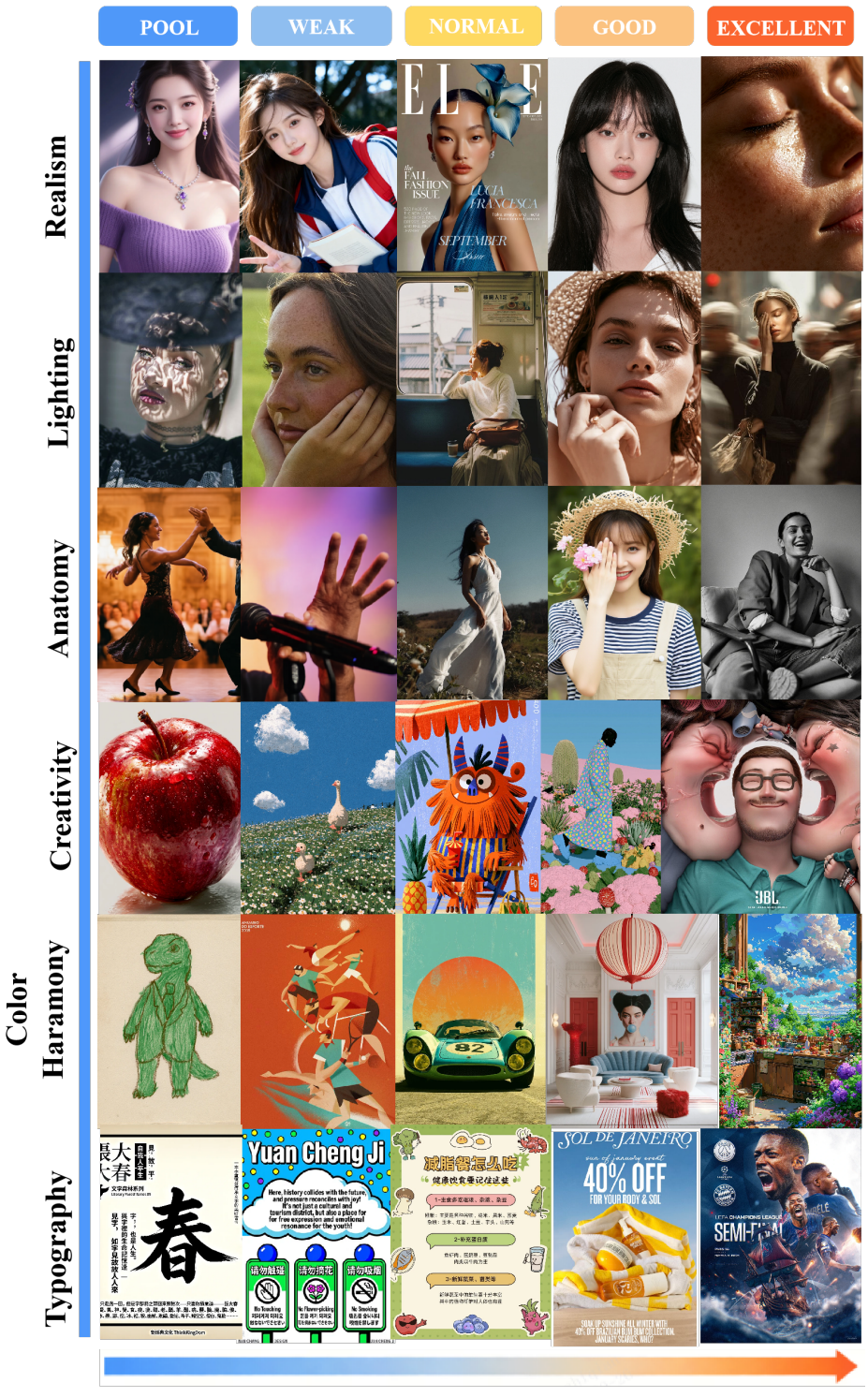

Figure 6: **Criterion-wise quality-level exemplars.** Representative examples for the six aesthetic criteria, *i.e.*, anatomical correctness, color, creativity, lighting, realism, and infographic-text quality, are organized into five quality levels. Each row corresponds to one criterion, while the columns progress from *poor* to *excellent* from left to right. These examples provide qualitative references for criterion-specific annotation and evaluation.

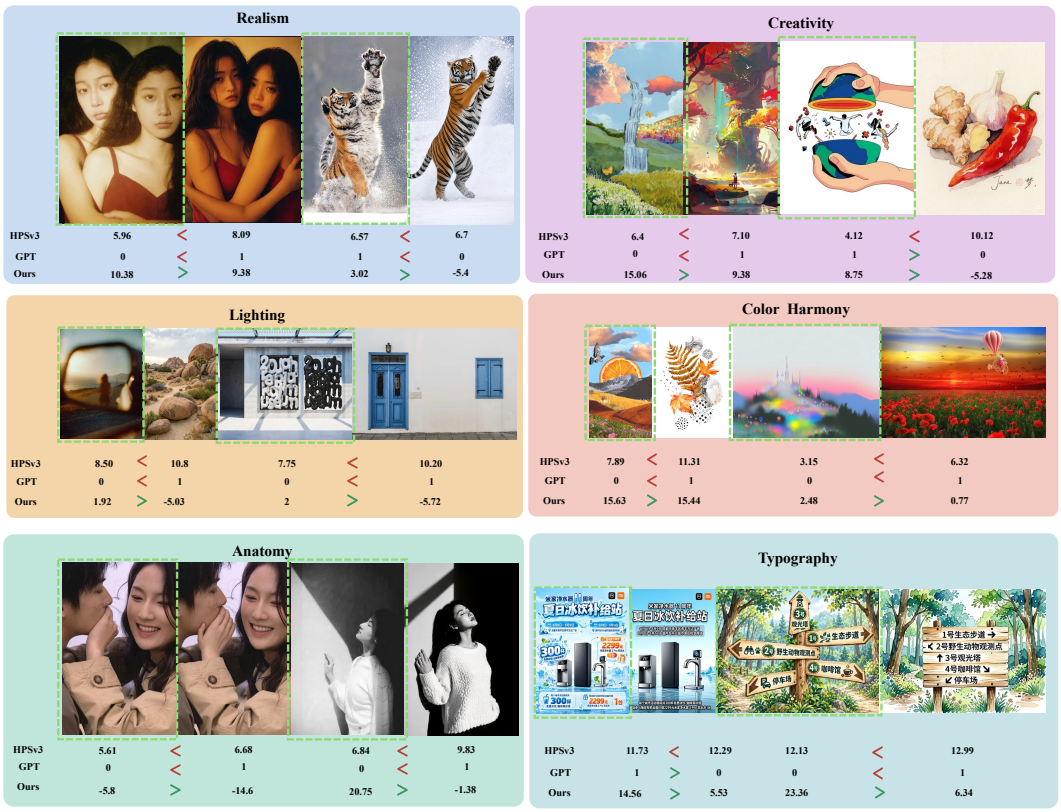

Figure 7: **AWARE-RBench comparison with existing evaluators.** AWARE is compared with HPSv3 Ma et al. (2025) and GPT-Astra on pairwise preference accuracy across the six criteria. For each criterion, two image pairs are shown. HPSv3 and AWARE assign a score to each image under the active criterion, with the higher-scoring image selected as preferred, whereas GPT-Astra directly selects the preferred image for that criterion. The outputs of HPSv3, GPT-Astra, and AWARE are displayed immediately below each pair. The red or green greater-than/less-than sign below each result indicates whether the prediction is incorrect or correct, respectively, relative to the annotated preference.

A.12 Benchmark Examples and Evaluation Criteria

Table 8 illustrates per-image applicability decisions and criterion-conditioned pairwise preferences. Table 9 presents actual AWARE-GenBench prompt excerpts, evaluation questions, and level-3 attainment criteria for the three content roles, translated into English and condensed for presentation.

Table 8: **AWARE-RBench: schematic evaluation examples.** Examples illustrate applicability decisions and criterion-conditioned preferences. Different dimensions can favor different candidates; the judgments shown here illustrate the task definitions.

Case / role	Criterion-conditioned pairwise case	Evaluation focus
Photography	*Prompt:* "natural portrait at golden hour." *Active:* realism, lighting, anatomy. *N/A:* typography.	Pairwise preferences under realism, lighting, and anatomy. Per-image applicability labels distinguish these criteria from N/A typography.
Artwork	*Prompt:* "expressive concept painting." *Active:* creativity, color harmony. *N/A:* photographic realism for a non-photorealistic painting.	Pairwise preferences under creativity and color harmony. Per-image applicability labels identify criteria relevant to the intended style.
Infographic	*Prompt:* "editorial portrait poster with a headline." *Pair:* A has accurate anatomy but weak typography; B has attractive typography but a distorted face. *Preference:* A $>$ B for anatomy; B $>$ A for typography.	Each criterion has its own pairwise preference label. Anatomy and typography are applicable to both candidates.

Table 9: **AWARE-GenBench: examples of prompt-specific evaluation.** Each row presents a prompt excerpt, one evaluation question, and its level-3 attainment criterion from the benchmark. Text is translated into English and condensed for presentation.

Subset	Prompt excerpt	Evaluation question	Level-3 criterion
Photography	Rolling green slopes at sunset, with dark pines along the ridge and golden cumulus clouds.	*Realism:* Do grass-surface details form a credible natural surface consistent with terrain, viewing distance, and vegetation growth?	Grass texture varies naturally in scale, density, and sharpness with distance and follows the slopes, producing a coherent natural surface; minor simplifications are acceptable.
Artwork	A stylized landscape illustration of a vast green mountain overlooking a village, a winding river, and flocks of sheep.	*Color harmony:* Do controlled variations in green distinguish the main terrain surfaces without monotony or arbitrary segmentation?	Coordinated variations in green distinguish ridges, slopes, and valley surfaces, creating depth while preserving the overall green palette; minor simplifications are acceptable.
Infographic	A seven-stage industrial R&D infographic. Stage 06, "SCALE-UP," includes descriptive text, large storage tanks, and an upward growth arrow.	*Readability and layout:* Do type size, weight, and spacing preserve the hierarchy of the SCALE-UP stage without crowding or overlap?	The title is emphasized without crowding the body text; both stay within their stage and do not conflict with the number, neighboring stages, or specified illustrations.

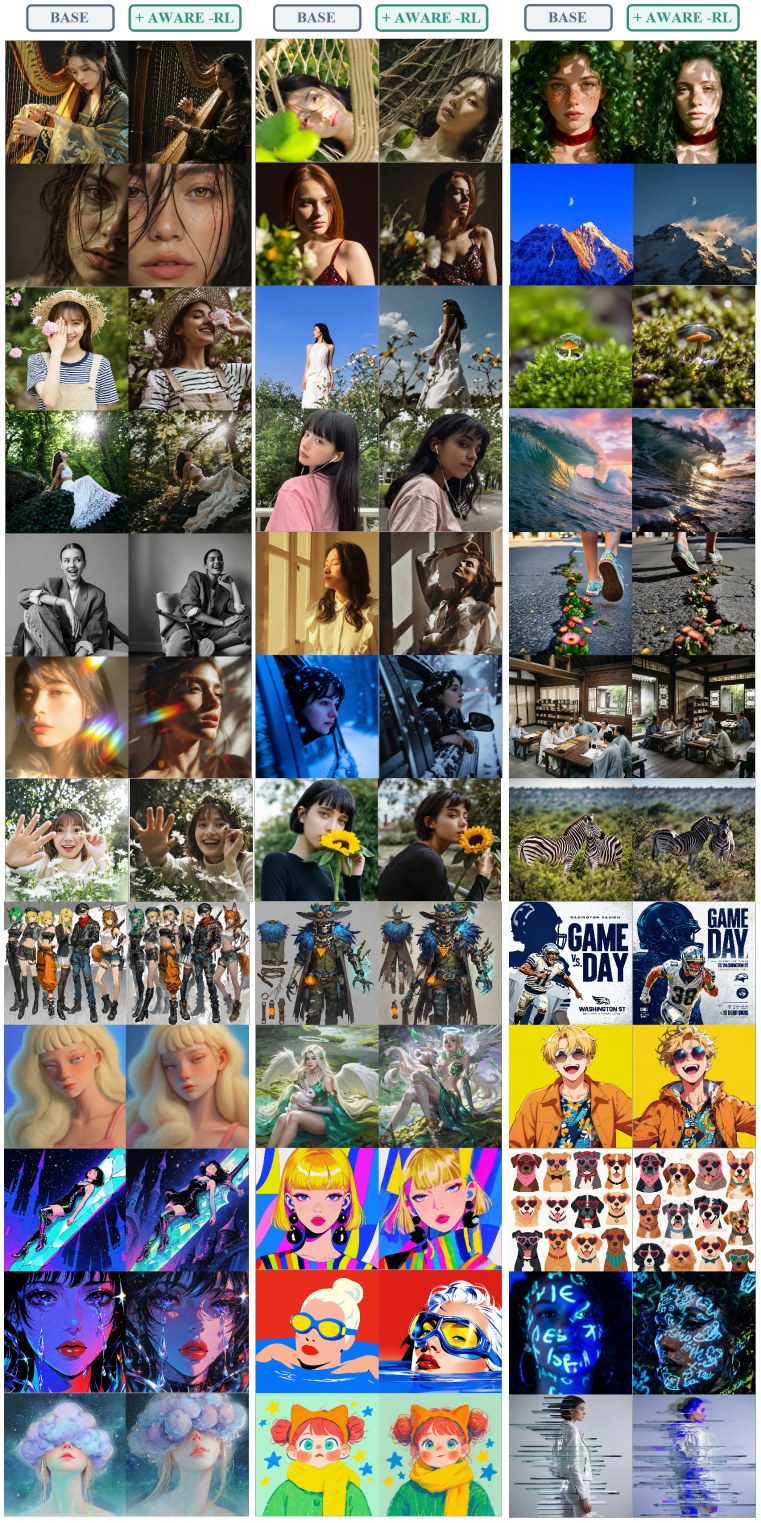

Figure 9: **Qualitative results on Z-Image.** Each adjacent pair compares the Z-Image base model (left) with its AWARE-RL enhanced counterpart (right) under the same prompt. Across diverse subjects and styles, including portraits, natural scenes, animals, graphic designs, and stylized illustration, AWARE-RL generally improves structural coherence, detail rendering, lighting, and overall aesthetic quality while preserving visual diversity.