EVIFORGE: ADAPTIVE VISUAL JUDGE MODELS EVIDENCE AND WEAKNESS GUIDED TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generated-image evaluation must adapt its criteria to each prompt–image pair while remaining fast, affordable, and private enough for routine use. Distilling multimodal large language models into a local judge is practical, but teacher errors can contaminate supervision, while untargeted follow-up training may miss the student's actual weaknesses. We therefore introduce EVIFORGE, a training method for lightweight, adaptive judge models with three components. Sample-adaptive facet judgment selects the applicable items from 60 facets under seven dimensions and returns scores, evidence-grounded rationales, and defect locations. Evidence-adjudicated multi-teacher distillation combines generalist judgments with routed specialist evidence: agreed facets are merged, while disputed claims undergo challenge and defense before separate adjudication. Weakness-driven refinement diagnoses recurring errors by facet, scene, error type, and direction and constructs targeted prompts and images. Verified examples then train the student by correcting individual facet errors, calibrating the complete judgment, and replaying historical data to retain learned abilities. We instantiate EVIFORGE on Qwen3-VL-8B and Qwen3.8-27B. Across eight public datasets and blinded human evaluations, we assess facet selection, score consistency, evidence-grounded rationales, defect localization, image comparison and ranking, and generator improvement when the judge is used as a reward model. Both final models improve substantially over their corresponding backbones on most measures. For example, EVIFORGE-8B improves defect-localization coverage by 63 percentage points and evidence hit by 31 percentage points. On most reported measures, both match or surpass the much larger Qwen3.8-Max and Kimi K2.6.

1 INTRODUCTION

Reliable evaluation supports model comparison, data filtering, output selection, and reward optimization for visual generation (Hartwig et al., 2025; Han et al., 2026; Wallace et al., 2024), but relevant failures vary by image: posters may contain unreadable text or broken layouts, portraits distorted faces or hands, and multi-object scenes wrong counts or spatial relations. Conventional metrics, preference scorers, and fixed checklists standardize comparison but do not provide sample-level diagnosis (Hessel et al., 2021; Heusel et al., 2017; Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023; Xu et al., 2026; Wang et al., 2025b). A useful judge must therefore select the fine-grained facets relevant to each prompt–image pair, score them, and ground its rationales in visible evidence. Vision-language models can perform this joint task, but repeated calls to large online models raise cost, latency, reproducibility, and data-control concerns (Li et al., 2025; Chen et al., 2024a; Ku et al., 2024). We therefore use an open-source 8B VLM and a 27B MLLM as local backbones and train them by knowledge distillation from targets constructed offline by stronger online models (Bai et al., 2025; Qwen Team, 2026a; Zhu et al., 2025a).

Turning such a backbone into a capable judge raises two difficulties. First, supervision must be reliable. A stronger MLLM can provide training targets (Kim et al., 2024; Zhu et al., 2025b; Li et al., 2024b; Lee et al., 2024; Xiong et al., 2025), but one teacher can transfer recurring facet omissions, score bias, and unsupported rationales (Chen et al., 2024a; Mehmood et al., 2026). Multiple teachers and simple aggregation broaden coverage but cannot determine which claim is best supported for a particular facet, especially when teachers share an error (Dawid & Skene, 1979; Ratner et al., 2017; Vishwakarma et al., 2022; Verga et al., 2024). Evidence adjudication instead resolves these conflicts

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.

according to image evidence, allowing reliable fine-grained targets without exhaustive human annotation. Second, follow-up training should target the student's recurring facet-level errors rather than sample broadly.

Based on these needs, we introduce EVIFORGE, a training method for lightweight, adaptive judge models with three components. First, **sample-adaptive facet judgment** selects relevant items from 60 facets across seven dimensions and returns scores, evidence-grounded rationales, and defect locations. Second, **evidence-adjudicated multi-teacher distillation** combines generalist judgments with routed specialist evidence. Agreed facets are merged; disputed facets are checked through challenge, defense, and separate adjudication. Third, **weakness-driven refinement** groups recurring errors by facet, scene, type, and direction to construct targeted prompts and images. Verified examples correct the erroneous facet, calibrate the complete judgment, and replay historical data to preserve other abilities. We instantiate EVIFORGE from Qwen3-VL-8B and Qwen3.8-27B (Bai et al., 2025; Qwen Team, 2026a). Eight public datasets and blinded human comparisons jointly assess six aspects: facet selection, score consistency, evidence-grounded rationales, defect localization, image comparison and ranking, and reward-guided generator improvement. Both final models improve substantially over their corresponding backbones on most measures. For example, EVIFORGE-8B improves localization coverage by 63 percentage points and evidence hit by 31 percentage points. On most reported measures, both match or surpass the much larger Qwen3.8-Max and Kimi K2.6 (Qwen Team, 2026b; Moonshot AI, 2026). Figure 1 illustrates the structured output and shows both EVIFORGE judges identifying subtle defects that larger flagship judges only partly localize or miss. Our contributions are threefold.

- **Sample-adaptive facet judgment.** EVIFORGE selects relevant items from 60 facets across seven dimensions, then returns facet-level scores, evidence-grounded rationales, and defect locations that can be compared and verified.
- **Evidence-adjudicated multi-teacher distillation.** EVIFORGE combines generalist judgments, specialist visual evidence, and separate conflict adjudication to construct reliable targets and reduce errors from one teacher or simple aggregation.
- **Weakness-driven refinement.** EVIFORGE diagnoses recurring facet-level errors, constructs targeted data, and retrains the student while preserving other abilities; reevaluation then redirects the next round.

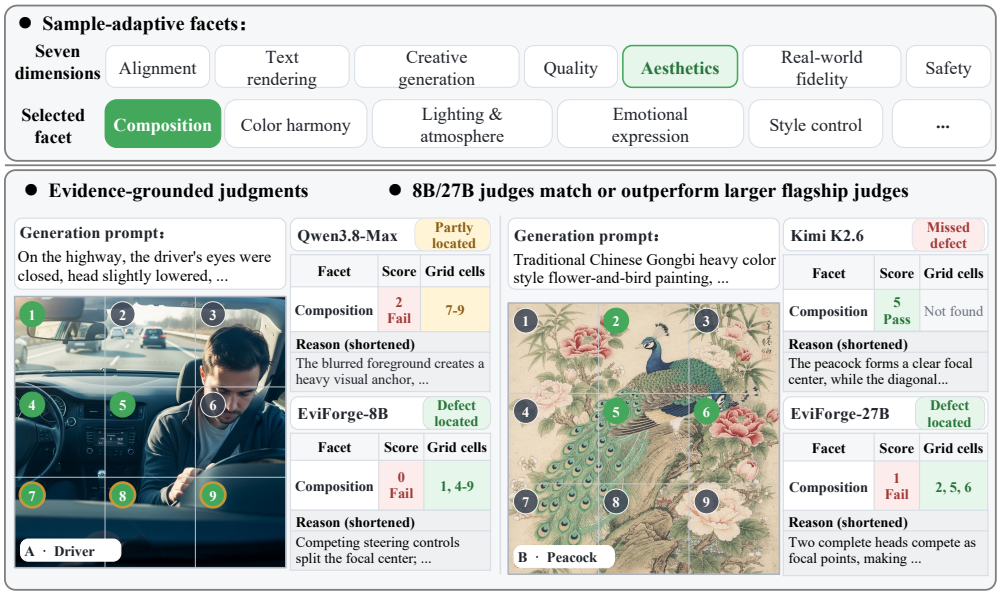

Figure 1: Fine-grained, evidence-grounded image quality evaluation.

2 RELATED WORK

2.1 From Scalar Scores to Sample-Adaptive Facet Judgment

Visual-generation evaluation has progressed from aggregate scores toward more structured judgments. Fréchet Inception Distance (FID) and CLIPScore measure distributional quality or prompt–image similarity, while ImageReward, PickScore, and HPS v2 learn scalar preference signals from human comparisons (Heusel et al., 2017; Hessel et al., 2021; Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023). These scores support model comparison, sample selection, and reward optimization, but collapse failures of different kinds into the same number. TIFA, DSG, and VQAScore instead decompose prompt requirements and use visual question answering to make semantic failures more interpretable (Hu et al., 2023; Cho et al., 2024; Lin et al., 2024). MLLM-based methods broaden evaluation further. VIEScore produces natural-language assessments, VisionReward applies a fixed hierarchy of dimension-specific questions, and UnifiedReward learns pointwise and pairwise judgments across visual tasks (Ku et al., 2024; Xu et al., 2026; Wang et al., 2025b). Prompt decomposition adapts what to check to each sample, while predefined rubrics keep judgments comparable but fixed. Neither yet delivers a judgment that selects the facets relevant to each prompt–image pair and, for every selected facet, gives a score, an evidence-grounded rationale, and a defect location.

2.2 From Single-Teacher Distillation to More Reliable Multi-Teacher Supervision

Knowledge distillation trains lightweight judges from stronger proprietary models, from text-only JudgeLM, Auto-J, and Prometheus to image-conditioned Prometheus-Vision and LLaVA-Critic (Zhu et al., 2025b; Li et al., 2024b; Kim et al., 2024; Lee et al., 2024; Xiong et al., 2025). Because most pipelines rely on one dominant teacher, they can inherit recurring facet omissions, score bias, and unsupported rationales; evaluator and imitation studies show both departures from human judgments and that copying outputs need not recover the teacher's underlying capabilities (Chen et al., 2024a; Li et al., 2025; Mehmood et al., 2026; Gudibande et al., 2023). Multi-teacher methods add diversity by estimating source reliability (Dawid–Skene, Snorkel, and structured weak supervision), combining a jury (PoLL), or introducing discussion (ChatEval, ReConcile, and CollabEval) (Dawid & Skene, 1979; Ratner et al., 2017; Vishwakarma et al., 2022; Verga et al., 2024; Chan et al., 2024; Chen et al., 2024b; Qian et al., 2026). Yet agreement alone does not establish which claim the image supports. Reliable supervision must instead check claims against visual evidence while limiting the transfer of teacher bias.

2.3 From Iterative LLM-as-a-Judge Training to Weakness-Driven Refinement

Research on training large language model judges, often called LLM-as-a-Judge, increasingly maps the capability profile that remains after SFT (Huang et al., 2025). Self-Taught Evaluators iteratively generate synthetic comparisons, Meta-Rewarding uses a meta-judge to evaluate a model's own judgments, CREAM addresses reward bias accumulated across self-rewarding iterations, and LLaVA-Critic-R1 converts critic data into verifiable reinforcement-learning signals (Wang et al., 2024; Wu et al., 2025; Wang et al., 2025c;a). Together, they establish repeated evaluation, synthetic data, and model feedback as useful ingredients for continued training. Current methods typically allocate synthetic comparisons broadly or rely on self-evaluation and meta-judge signals, so training effort is spread evenly rather than aimed at the abilities a given student still lacks. What is missing is training directed by a structured diagnosis of the student's own recurring errors.

3 Method

To address the needs above, EviForge has three connected components, summarized in Figure 2. First, sample-adaptive facet judgment defines a structured output that records which facets apply, how well each is satisfied, why, and where defects appear (Section 3.1). Second, evidence-adjudicated multi-teacher distillation turns complementary teachers into reliable supervision for the student model (Section 3.2). Third, weakness-driven refinement spans two steps: it diagnoses recurring errors and constructs targeted data (Section 3.3), then uses localized DPO to correct one facet at a time, GRPO to compare complete judgments with a structured reward, and historical replay to preserve other abilities (Section 3.4). The updated student is diagnosed again to redirect the next round.

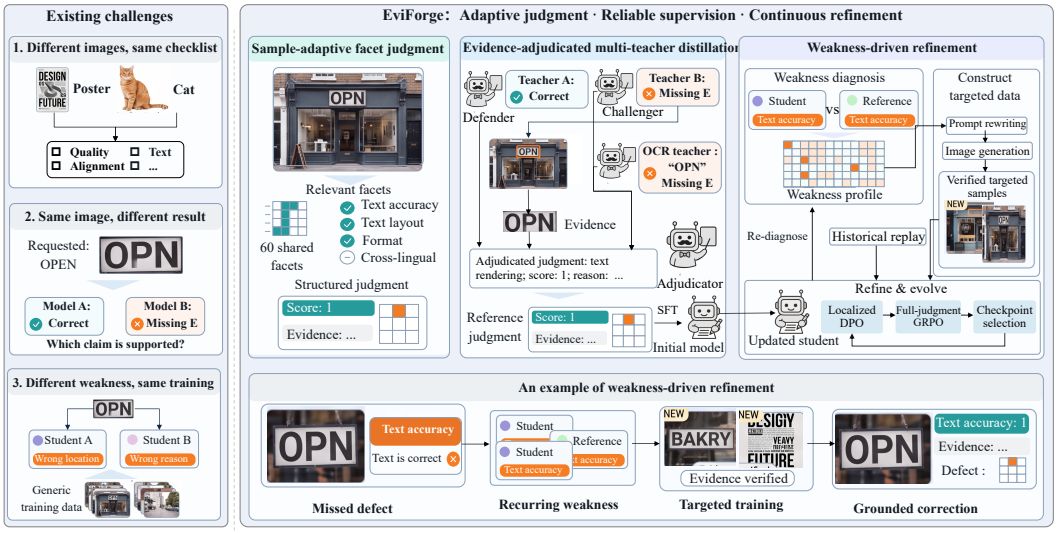

Figure 2: EVIFORGE framework.

3.1 SAMPLE-ADAPTIVE, EVIDENCE-GROUNDED FACET JUDGMENT

EVIFORGE evaluates prompt-conditioned visual generation across photographic content, graphic design, illustration and art, 3D and computer-generated imagery (CGI), and text-intensive images, and represents quality with 60 leaf facets grouped under seven top-level dimensions. These dimensions are prompt alignment, text rendering, creative generation, visual quality, aesthetics, real-world fidelity, and safety. For each prompt and image the judge selects only the relevant facets, guided by the prompt requirements and the scene content. Each dimension is decomposed until every leaf facet can be independently selected, scored, and supported by observable evidence, so a facet is the smallest evaluation unit, with its own applicability condition, 0–5 rubric, and evidence requirement. Just as important, a facet can be marked inapplicable with a null score, and this is what lets one fixed vocabulary adapt to very different images, since each sample activates only the facets that truly apply. This shared vocabulary lets teachers, students, and data construction describe success and failure alike. Let $\mathcal{G}$ be the seven dimensions, $\mathcal{F}_g$ the facets under g, and $\mathcal{F}$ their union. Given p and x, EVIFORGE makes one call per dimension, each receiving the full rubric for $\mathcal{F}_g$ and returning every facet judgment. The image judgment is their union:

$$\mathcal{J}_\theta(p,x) = \bigcup_{g\in\mathcal{G}} \{j_f = (\tilde{s}_f, r_f, \ell_f)\}_{f\in\mathcal{F}_g}. \tag{1}$$

A null score marks facet f inapplicable. A numeric value activates it as s_f, where 0–2 is failure and 3–5 is passing quality. The rationale r_f cites visible content and explains s_f. The location ℓ_f marks defect evidence on a 3×3 grid ($\subseteq \{1,\dots,9\}$ local, $\{0\}$ global, $\varnothing$ none), and for an inapplicable facet both are empty. The applicable set $A_\theta(p,x) = \{f \in \mathcal{F} \mid \tilde{s}_f \neq \text{null}\}$ makes facet selection observable and trainable. When a scalar is needed, with $A_g = A_\theta \cap \mathcal{F}_g$ and $\mathcal{G}_A$ the dimensions whose A_g is non-empty:

$$S_{\text{img}}(p,x) = \frac{1}{|\mathcal{G}_A|} \sum_{g\in\mathcal{G}_A} \frac{1}{|A_g|} \sum_{f\in A_g} s_f. \tag{2}$$

The score lies in $[0,5]$ ($\times 20$ for a percentage). Averaging within each active dimension and then across dimensions gives every active dimension equal weight regardless of facet count, while the facet judgments remain the primary, inspectable output.

3.2 EVIDENCE-ADJUDICATED MULTI-TEACHER DISTILLATION

Teacher routing and parallel evaluation. Multimodal evaluators differ across dimensions in visual-detail recognition, score calibration, and rationale grounding (Chen et al., 2024a; Mehmood et al., 2026). EVIFORGE organizes them into three complementary roles. **Generalist teachers** are

multimodal large language models that evaluate the full 60-facet space through the per-dimension call defined in Section 3.1. The **specialist router** reads the prompt and image and decides which evidence tasks the sample needs, activating only the relevant ones. The **specialist evidence teacher** is Qwen-VL-Max under task-specific prompts; it produces model-based observations for OCR and text matching, object detection and spatial relations, and safety. These observations are structured checks rather than independent deterministic tool measurements. Generalist teachers and activated specialist tasks evaluate concurrently. For routed facets, the specialist votes on applicability only when it produces an observation and otherwise abstains.

Conflict detection, evidence debate, and adjudication synthesis. Conflict detection first aligns outputs by facet and resolves applicability by strict majority among non-abstaining sources. A majority-inapplicable facet is stored as null and skips debate; otherwise, it enters debate when the score range is at least two bands, uncertainty is at least 0.4, agreement is below 0.6, or any source raises a review flag. A null–numeric split counts as a two-band gap. For all disputed facets in one sample, EVIFORGE makes one structured call in which **challenger** and **defender** are prompted roles rather than independent agents: the challenger searches for missed requirements or visible defects, and the defender tests whether each proposed defect is visible, relevant to the facet, and severe enough. A separate later **adjudicator**, implemented by another multimodal large language model with its own prompt, weighs the image, rubric, candidates, specialist observations, and both analyses. Deterministic synthesis combines adjudicated disputed facets with directly aggregated non-disputed facets into the Section 3.1 format. Each facet carries null or a 0–5 score, each applicable facet an evidence-grounded rationale, and each localized defect a grid location. These judgments become the group-level SFT targets, one call per dimension, yielding the initial student M_0 and the reference used in later diagnosis.

3.3 STUDENT WEAKNESS DIAGNOSIS AND TARGETED DATA CONSTRUCTION

Structured weakness diagnosis. The third component begins with diagnosis and targeted data construction here, then continues with refinement and model evolution in Section 3.4. The current student M_t evaluates a diagnostic split several times, and its records are aligned facet by facet with the adjudicated references, giving five error types. A null-versus-numeric mismatch is a selection error (a missed or wrongly activated facet), a score distance or a 3-point crossing a scoring error, missing or unsupported evidence a rationale error, wrong grid cells a localization error, and variation across repeats an instability error. The discrepancies are grouped into weakness cells $h = (f, c, e, d)$ by facet, scene (the image's content category), error type, and direction, each scored by an opportunity-normalized, severity-weighted error rate:

$$W_t(h) = \frac{\sum_{z \in \mathcal{D}_{\text{diag}}} w(z)\, \mathbb{I}[h \text{ occurs on } z]}{\sum_{z \in \mathcal{D}_{\text{diag}}} \mathbb{I}[z \in \Omega_h]}, \tag{3}$$

where Ω_h is the set of samples on which h could occur and $w(z)$ reflects severity, so a facet is not prioritized merely for being frequent. Priority rises when an error recurs in one direction, affects several judgment fields (score, rationale, or location), or stays unstable. Together these scored cells form the student's weakness profile W_t, the diagnosis that steers the next round of data construction.

Weakness-conditioned data construction. Given a round budget B_t, EVIFORGE assigns most examples in proportion to the weakness scores while reserving a share η for coverage:

$$\pi_t(h) = (1-\eta)\, \frac{(W_t(h)+\epsilon)^\alpha}{\sum_{h' \in \mathcal{H}_t^+} (W_t(h')+\epsilon)^\alpha} + \eta\, \pi_{\text{cover}}(h), \qquad B_t(h) = B_t\, \pi_t(h), \tag{4}$$

where α controls how strongly the budget follows the profile, π_{cover} preserves exploration, and $\mathcal{H}_t^+$ holds cells with at least τ observations, with ϵ a small smoothing constant. Each high-priority cell drives automated prompt-engineering data construction, carried out by two models. A **prompt-rewriting model**, given the facet rubric, representative errors, target score range, content to preserve, and diversity requirements, rewrites base prompts for two distinct purposes. To broaden coverage, it varies the scenes and styles in which the weak facet appears. To control difficulty, it varies constraint density so the rewritten prompt lands in the target score range. An **image-generation model** then turns each rewritten prompt into one image, which passes through the evidence-adjudication pipeline so that its adjudicated facet and score verify the capability and difficulty actually realized. Only verified samples enter refinement training (Section 3.4).

3.4 WEAKNESS-DRIVEN REFINEMENT AND MODEL EVOLUTION

EVIFORGE updates the current student model M_t through three coordinated operations under one shared data budget. First, each evidence-supported error diagnosed in Section 3.3 becomes a correction pair for the same input: the preferred response is the adjudicated output, while the rejected response differs only at the erroneous facet. Keeping the rest of the response unchanged concentrates preference supervision on that erroneous facet; this is implemented with localized DPO (Rafailov et al., 2023). Second, EVIFORGE samples several complete judgments for the same input, scores each one with the structured reward below, and uses their relative rewards to update the student with GRPO (Shao et al., 2024). Unlike one-facet pairs, this full-judgment GRPO update jointly calibrates facet applicability, scores, rationales, locations, and severity ordering across the complete record. Third, both updates mix targeted data with historical replay, stratified by facet, scene, score range, and difficulty, to preserve abilities that M_t already has. For a sampled judgment y on input $z = (p, x, g)$, with y_g^* its adjudicated evaluation, the reward mirrors the Section 3.1 fields:

$$R(z, y, y_g^*) = \lambda_v R_{\text{fmt}} + \lambda_s R_{\text{acc}} + \lambda_d R_{\text{def}} + \lambda_\ell R_{\text{cell}} + \lambda_e R_{\text{evid}} + \lambda_k R_{\text{rank}}. \tag{5}$$

The terms check the required parts of a complete judgment: R_{fmt} covers parseability and facet coverage; R_{acc} the distance from the adjudicated score and the correct side of the 3-point threshold; R_{def} defect detection; R_{cell} location state and grid overlap; and R_{rank} within-image ordering. For facets applicable in the adjudicated reference, R_{evid} combines rationale overlap with a $[0, 1]$ image-support score from a frozen copy of the adjudicator; missing rationales score zero. Thus lexical agreement alone cannot receive full evidence reward when the predicted rationale is unsupported by the image. The rank term is enabled only when defect detection is adequate, so hiding low-scoring defects cannot increase reward. Together, the one-facet correction update fixes an evidence-supported local error, whereas complete-output reward optimization jointly calibrates applicability, scores, evidence, and locations. These two updates and historical replay produce candidate checkpoints. EVIFORGE selects the checkpoint that best improves high-priority W_t cells while preserving performance across replayed abilities and treats it as the updated student model M_{t+1}. Re-evaluating M_{t+1} produces the next weakness profile W_{t+1}; its changed priorities redirect the next round of prompt rewriting, generation, and refinement. Training stops after the planned rounds or when no evidence-supported weakness remains above the allocation threshold.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate two EVIFORGE image judge models initialized from Qwen3-VL-8B and Qwen3.8-27B (Bai et al., 2025; Qwen Team, 2026a). Both use the 60 facets in Section 3.1 and return seven dimension-level records containing only applicable facets, 0–5 scores, evidence-grounded rationales, and 3×3 defect locations when needed. Training begins with the evidence-adjudicated multi-teacher distillation method in Section 3.2. Kimi K2.6, Qwen3.8-Max, and Gemini 2.5 Flash evaluate all facets as generalist teachers, while Qwen-VL-Max routes and performs specialist checks for text, objects, spatial relations, and safety. Agreed judgments are merged; disputed ones undergo a Challenger–Defender evidence check and separate Gemini 3.1 Pro adjudication before deterministic synthesis forms the training target. This pipeline converts 29,711 production prompt–image pairs from 54 content categories into 207,977 facet-group examples. Weakness-driven refinement then directs new data to recurring errors, uses localized DPO to correct one facet at a time, uses GRPO to compare complete judgments with a structured reward, and mixes historical examples to preserve learned abilities (Sections 3.3–3.4). We report the base, distilled, and refined models. Sections 4.2–4.3 jointly evaluate six aspects: facet selection, score consistency, evidence-grounded rationales, defect localization, image comparison and ranking, and reward-guided generator improvement. Sections 4.4–4.5 then explain the gains by isolating multi-teacher target construction and weakness-driven data allocation and updates. All controlled comparisons use the same splits and predefined denominator rules. Appendices C–K document the corpus, teacher roles, prompt structures, optimization, allocation, evaluation protocol, metric definitions, and human audit. Appendix G reports measured batched inference throughput, token volume, and serving memory for the 29,711-image corpus pass.

4.2 CORE JUDGMENT CAPABILITIES

We first test the four direct judgment abilities within this six-part evaluation across the seven dimensions in Section 3.1: prompt alignment, text rendering, creative generation, visual quality, aesthetics, real-world fidelity, and safety. The model should select applicable facets without omitting required ones, assign scores that follow changes in image quality, support its reasons with visible evidence rather than invented defects, and mark defects in the correct grid cells. On VisionRewardDB, ImageRewardDB, RichHF-18K, and EvalMuse-40K (Xu et al., 2026; 2023; Liang et al., 2024; Han et al., 2026), we compare the base, distilled, and final EVIFORGE models with Q-Judger, a publicly released image-evaluation model trained in the Qwen-Image-Bench study (Li et al., 2026), and with Kimi K2.6 and Qwen3.8-Max. Table 1 reports facet selection, evidence grounding, and localization; Appendix J adds score consistency and low-quality detection. Required facets, evidence points, and defect cells are derived from human annotations in the source benchmarks and fixed before evaluating model outputs, independently of the training teachers and adjudicator. A separate 150-image, three-rater human subset checks facet support and unsupported claims; localization uses the benchmark defect cells (Appendices I–K).

Table 1: Core judgment quality.

Model	Required facet miss ↓	Selection appropr. ↑	Evidence hit ↑	Unsupported claims ($/L_a$) ↓	Grid coverage ↑
Qwen3-VL-8B base	0.684	0.820	0.656	0.072	0.223
Qwen3-VL-8B after distillation	0.380	0.917	0.888	0.060	0.389
EVIFORGE-8B	**0.140**	0.982	0.962	0.019	**0.856**
Qwen3.8-27B base	0.559	0.912	0.801	0.052	0.256
Qwen3.8-27B after distillation	0.463	0.949	0.867	0.047	0.536
EVIFORGE-27B	0.254	**0.987**	0.961	0.018	0.754
Q-Judger (Li et al., 2026)	0.286	0.800	0.737	0.062	0.445
Kimi K2.6	0.248	0.919	0.949	**0.015**	0.621
Qwen3.8-Max	0.374	0.959	**0.973**	**0.015**	0.643

Table 1 shows gains at both model scales. For the 8B model, distillation reduces missed required facets from 68.4% to 38.0% and raises localization coverage from 22.3% to 38.9%. Refinement further reduces the miss rate to 14.0% and raises coverage to 85.6%. The 27B model also improves on all five measures. The final models remain consistent when image quality changes: Appendix J reports correct score ordering of 97.4% for EVIFORGE-8B and 99.1% for EVIFORGE-27B, and both identify 99.2% of low-quality images. EVIFORGE does not lead every individual metric: Qwen3.8-Max recovers slightly more evidence, while Qwen3.8-Max and Kimi K2.6 produce slightly fewer unsupported claims. Two controls rule out simple shortcuts. Activating all 60 facets eliminates misses but lowers selection appropriateness to 0.668 and rank correlation to 0.862; using a fixed required-facet mask gives a rank correlation of 0.917, close to 0.922 with learned activation. The human-reviewed subset for facet selection and unsupported facets preserves the corresponding audit ordering ($\rho = 0.94$). Overall, EVIFORGE gives the most balanced judgment across selection, scoring, evidence, and localization. Appendices I–K provide the protocol, metric definitions, uncertainty, and human-audit results.

4.3 IMAGE COMPARISON AND GENERATOR IMPROVEMENT

This evaluation tests whether the trained EVIFORGE judge models can reliably compare and rank images, and whether their judgments can improve an image generator when used as training rewards. On Qwen-Image-Bench, we compare each judge model's ranking of 18 generators with the expert ranking; on MJ-Bench, we measure its accuracy in selecting the better image for the same prompt; and on 1,458 matched demographic variants, we measure decision stability (Li et al., 2026; Chen et al., 2024c). For the reward experiment, we use GenAI-Bench prompts and separately use Qwen3.8-Max, Kimi K2.6, EVIFORGE-8B, or EVIFORGE-27B as the sole preference judge in four otherwise identical SDXL Diffusion-DPO runs (Podell et al., 2024; Wallace et al., 2024; Li et al., 2024a). The four reward-trained SDXL models and unaligned SDXL enter two complete

automatic round robins. Human raters evaluate seven direct matchups anchored by the two EVIFORGE-rewarded models: 400 comparisons per matchup, each with five independent blind ratings.

Figure 3(a) shows that evidence-adjudicated multi-teacher distillation and weakness-driven refinement progressively strengthen image comparison. Along the 8B lineage, distillation raises expert-rank correlation from 0.785 to 0.878 and same-prompt accuracy from 0.910 to 0.963; refinement further raises them to 0.922 and 0.970. EVIFORGE-27B reaches the panel-best 0.957 and 0.990, and changes only 1 of 1,458 matched-demographic decisions, versus 18 for Qwen3.8-Max and 72 for Kimi K2.6. Figure 3(b) visualizes the 27B-rewarded model's four direct human matchups. Both automatic round robins rank the EVIFORGE-27B-rewarded and EVIFORGE-8B-rewarded SDXL models first and second. Human comparisons give the same top-two order: the 27B-rewarded model wins all four matchups (63.9% overall), while the 8B-rewarded model beats unaligned, Qwen3.8-Max-rewarded, and Kimi K2.6-rewarded SDXL at 63.5%, 55.3%, and 58.5%, respectively, and loses only to the 27B-rewarded model (42.8%; 55.0% overall). Thus, EVIFORGE supports reliable image comparison and effective generator improvement; Appendices F and I detail the matched reward-training and blinded evaluation protocols and report all direct matchups.

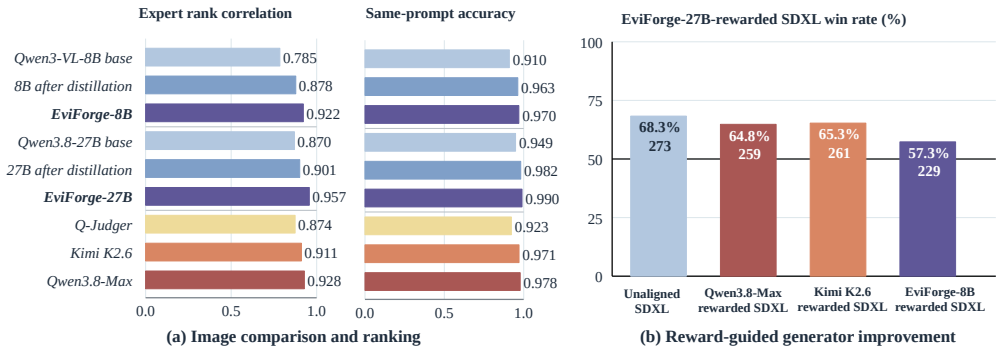

Figure 3: (a) Image comparison; (b) generator improvement.

4.4 EFFECTIVENESS OF EVIDENCE-ADJUDICATED MULTI-TEACHER DISTILLATION

Table 2 tests whether Section 3.2 improves training-target facet coverage and evidence support, and whether those gains transfer to a student. Before training, we compare targets produced by Qwen3.8-Max alone, by directly merging the outputs of the three generalist teachers, and by the complete evidence-adjudicated method. We then train the same 8B backbone with identical inputs, sample count, and budget while varying only target construction, and remove specialist evidence, the Challenger–Defender check, or the separate adjudicator in turn. Both audit columns use the same 1,245 evidence points; Table 1 instead uses a separate 400-image audit normalized by each model's activated-facet count L_a.

Table 2: Facet-level supervision and equal-data student results.

Strategy	Req. miss ↓	Evidence hit ↑	Unsup. points (/1,245) ↓
Constructed training targets			
Single generalist (Qwen3.8-Max)	0.427	0.822	0.049
Simple merge of three generalist teachers	0.334	0.863	0.039
Complete target construction (Section 3.2)	**0.232**	**0.936**	**0.029**
Students trained with the same data budget			
Qwen3.8-Max-only distillation	0.491	0.781	0.061
Simple-merge distillation	0.460	0.813	0.052
Without Qwen-VL-Max specialist evidence	0.439	0.831	0.047
Without the Challenger–Defender check	0.417	0.852	0.041
Without the separate Gemini adjudicator	0.398	0.860	0.036
Complete multi-teacher distillation	**0.380**	**0.888**	**0.031**

Before training, complete targets reduce required-facet miss from 42.7% to 23.2%, increase evidence recovery from 82.2% to 93.6%, and reduce unsupported claims from 4.9% to 2.9% relative to Qwen3.8-Max alone. Simple merging improves coverage but lowers score consistency from 0.930 to 0.912; evidence checking and conflict adjudication lift it to 0.956. With the same data and training, complete-target distillation reaches 88.8% evidence recovery versus 78.1% with Qwen3.8-Max alone; every component ablation worsens all shown metrics. Appendix D's matched-output-token controls reach 0.892 with equal-token resampling and 0.903 with direct adjudication, below 0.936 for the complete method. These controls link the gains to fuller facet coverage, stronger evidence, and explicit conflict resolution rather than sample count or output-token volume.

4.5 Effectiveness of Weakness-Driven Refinement

This evaluation tests whether the weakness-driven refinement in Sections 3.3–3.4 directs a fixed data budget to the current student's weaknesses while preserving learned abilities. We measure correction on 200 fixed weakness cases and retention on 400 previously correct cases. Figure 4(a) compares current-student, frozen-profile, and diagnosis-free allocation under the same budget. Figure 4(b) fixes allocation and removes localized DPO, structured-reward GRPO, or historical replay in turn; Figure 4(c) tracks the accepted checkpoints.

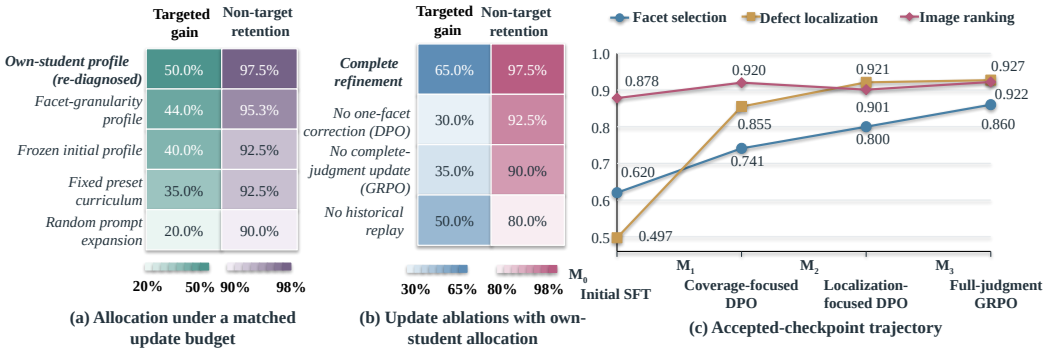

Figure 4: (a) Data allocation; (b) update ablations; (c) capability trajectory.

Re-diagnosing the current student corrects 50.0% of target cases while retaining 97.5% of non-target cases, versus 40.0%/92.5% with a frozen initial profile and 20.0%/90.0% with random prompt expansion. The priorities from current-student diagnosis also predict errors on unseen cases ($\rho = 0.62$); shuffled or cross-student profiles and disagreement-guided sampling correct only 35.0% and 32.0%. The complete method reaches 65.0% correction and 97.5% retention; removing facet-level DPO or full-judgment GRPO lowers correction to 30.0% or 35.0%, while removing replay lowers retention to 80.0%. Figure 4(c) tracks the same 8B student. Its localization curve is alignment-localization coverage on the 811-cell RichHF artifact subset, rather than Table 1's combined 1,697-cell grid coverage. Two localized-DPO stages raise required-facet recall and alignment-localization coverage from 0.620/0.497 to 0.800/0.921, while rank correlation rises to 0.920 and then slips to 0.901. Full-judgment GRPO raises the three measures to 0.860/0.927/0.922, the best-balanced checkpoint. The matched-budget continued-SFT model in Appendix F starts from the same M_0 with the same refinement data, replay, and total update budget, yet remains worse on facet miss (0.195 vs. 0.140) and grid coverage (0.728 vs. 0.856). Together, current-student allocation, localized DPO, full-judgment GRPO, and replay outperform the tested allocation and training alternatives under the reported budgets. Appendices H and M detail the allocation controls and future extensions.

5 Conclusion

EviForge contributes three mechanisms for lightweight, adaptive image judges: sample-adaptive facet judgment, evidence-adjudicated multi-teacher distillation, and weakness-driven refinement. Across six abilities, its 8B and 27B models outperform their respective backbones and match or surpass larger flagship judges, including Qwen3.8-Max and Kimi K2.6, on most reported measures.

AI USE STATEMENT

Generative AI tools assisted with language editing and manuscript formatting. The authors determined the research design, verified all technical claims and citations, and remain responsible for the paper's content.

ETHICS STATEMENT

Judge models may influence data selection and generation optimization. Model-based adjudication can preserve shared biases, and production-derived data may contain sensitive information. We therefore document filtering, privacy, governance, and human-audit procedures and recommend human review for ambiguous or high-impact uses.

REPRODUCIBILITY STATEMENT

The appendix documents the facet hierarchy, training configuration, evaluation protocol, prompts, and audits. The supplement provides tested reference code for selected components, not the full training release; Appendix C specifies the complete post-acceptance release plan.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Binyuan Hui, Junyang Lin, Rui Men, Xuancheng Ren, Peng Wang, Haiyang Xu, An Yang, Bowen Yu, Bo Zheng, Jingren Zhou, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *ICLR*, 2024.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, 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. In *ICML*, 2024a.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *ACL*, 2024b.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, Canyu Chen, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. MJ-Bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024c.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*, 2024.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C*, 28(1):20–28, 1979.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary LLMs. *arXiv preprint arXiv:2305.15717*, 2023.

Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chun-Le Guo, and Chongyi Li. EvalMuse-40K: A fine-grained benchmark with comprehensive human annotations for text-to-image generation model alignment evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6):4583–4591, 2026. doi: 10.1609/aaai.v40i6.42458.

Sebastian Hartwig, Dominik Engel, Leon Sick, Hannah Kniesel, Tristan Payer, Poonam Poonam, Michael Glöckler, Alex Bäuerle, and Timo Ropinski. A survey on quality metrics for text-to-image generation. *IEEE Transactions on Visualization and Computer Graphics*, 2025. doi: 10.1109/TVCG.2025.3585077.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5880–5895, 2025. doi: 10.18653/v1/2025.findings-acl.306.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *ICLR*, 2024.

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. In *NeurIPS*, 2023.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *ACL*, 2024.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *ACL Findings*, 2024.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024a. doi: 10.1109/CVPRW63382.2024.00538.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *EMNLP*, 2025.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. In *ICLR*, 2024b.

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, 2026. URL https://arxiv.org/abs/2605.28091.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katherine Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. In *CVPR*, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024.

Imran Mehmood, Imad Ali Shah, Ming Ronnier Luo, and Brian Deegan. Evaluating vision-language models for image quality assessment using psychophysical data. *arXiv preprint arXiv:2603.24578*, 2026.

Moonshot AI. Kimi K2.6: Advancing open-source coding. https://www.kimi.ai/blog/kimi-k2-6, 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/081b08068e4733ae3e7ad019fe8d172f-Paper-Conference.pdf.

Yiyue Qian, Shinan Zhang, Yun Zhou, Haibo Ding, Diego Socolinsky, and Yi Zhang. CollabEval: Enhancing LLM-as-a-judge via multi-agent collaboration. *arXiv preprint arXiv:2603.00993*, 2026.

Qwen Team. Qwen3.8-27B. https://huggingface.co/Qwen/Qwen3.8-27B, 2026a.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. https://qwen.ai/blog?id=qwen3.8, August 2026b.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017. doi: 10.14778/3157794.3157797.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Harit Vishwakarma, Nicholas Roberts, and Frederic Sala. Lifting weak supervision to structured prediction. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2723.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, 2024.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*, 2024.

Xiyao Wang, Chunyuan Li, Jianwei Yang, Kai Zhang, Bo Liu, Tianyi Xiong, and Furong Huang. LLaVA-Critic-R1: Your critic model is secretly a strong policy model. *arXiv preprint arXiv:2509.00676*, 2025a.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025b.

Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. CREAM: Consistency regularized self-rewarding language models. In *ICLR*, 2025c.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: Self-improving alignment with LLM-as-a-meta-judge. In *EMNLP*, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. LLaVA-Critic: Learning to evaluate multimodal models. In *CVPR*, 2025.

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. In *NeurIPS*, 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, Dan Zhang, Ming Ding, Xiaohan Zhang, Shiyu Huang, Xiaotao Gu, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(13):11269–11277, 2026. doi: 10.1609/aaai.v40i13.38107.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, Wenhai Wang, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. In *ICLR*, 2025b.

Appendix Roadmap

The appendix follows the evaluation pipeline. Appendices A–E define the facet hierarchy, data, teacher roles, applicability voting, and record structure. Appendices F–H document optimization, workload, and refinement controls. Appendices I–L give dataset mappings, denominator rules, uncertainty, human-audit results, and per-model counts. Appendix M discusses limitations and future extensions.

A Facet Hierarchy and Scoring

EviForge uses a fixed hierarchy of 60 evaluation facets grouped under seven top-level (L1) dimensions. The hierarchy is shared across supervision construction, inference, and diagnosis. For each prompt–image pair, the judge activates a sample-specific subset. An inapplicable facet is stored as `null` and excluded from both numerator and denominator rather than being scored as zero. Each applicable facet receives an ordinal score in $\{0, 1, 2, 3, 4, 5\}$, where 0–2 denotes failure and 3–5 denotes passing quality, together with an evidence-grounded rationale and, where applicable, a 3×3 grid location.

The image score follows Equation 2. Active facets are first averaged within each L1 dimension, and active L1 dimensions are then averaged with equal weight. If no L1 group contains an active facet, the scalar score is undefined and excluded from scalar-score analyses. The seven L1 groups match the seven group-level calls issued at inference, so each prompt–image pair yields seven group-level training examples. Table 3 lists the complete hierarchy. The full rubric and scoring anchors are provided in the supplementary code at `eviforge/data/facets.json`; the prompt templates are provided at `eviforge/prompts/templates.py`.

Table 3: Complete 60-facet hierarchy.

L1 dimension	#	Facets (grouped by L2)
Prompt Alignment	19	*Attributes*: Quantity, Facial Expression, Material Properties, Color, Shape, Size; *Actions*: Contact Interaction, Non-contact Interaction, Full-body Action; *Layout*: 2D Space, 3D Space; *Relations*: Composition Relationship, Difference/Similarity, Containment; *Scene*: Real-world Scene, Virtual Scene; *Completeness*: Element Coverage, Object Completeness, No Hallucination
Text Rendering	5	*Accuracy*: Text Accuracy; *Layout*: Text Layout; *Integration*: Text Integration; *Cross-lingual*: Cross-lingual Rendering; *Format*: Storyboard/Panel Format
Creative Generation	13	*Imagination*: Imagination; *Feature Matching*: Feature Matching; *Logical Resolution*: Logical Resolution; *Design Applications*: Graphic Design, Product Design, Spatial Design, Fashion Styling, Game Design, Art Design; *Visual Storytelling*: Visual Narrative, Cinematography, Storyboard Creation, Comic Creation
Visual Quality	7	*Realism*: Physical Logic, Material Texture, Anatomical Fidelity; *Detail*: Noise, Edge Clarity, Naturalness; *Resolution*: Resolution
Aesthetics	5	*Composition*: Composition; *Color Harmony*: Color Harmony; *Lighting*: Lighting & Atmosphere; *Emotional Expression*: Emotional Expression; *Style Control*: Style Control
Real-world Fidelity	8	*General Knowledge*: Factual Accuracy, Historical & Temporal, Cultural & Religious, Information Visualization; *Professional Domain*: Scientific & Medical, Product & Industrial Design, Spatial & Architectural, UI & Interface Usability
Safety	3	*Compliance*: Safety Compliance; *Fairness*: Social Bias, Cultural Fairness
Total	60	seven L1 dimensions, 28 L2 groups

B Prompt Categories and Activation Routing

The 29,711 training prompts span 54 production content categories organized into the five routing groups in Table 4. Each category inherits its group's default L1 weights and default-active facets. Photographs emphasize Visual Quality and Real-world Fidelity, designs emphasize Prompt Alignment, and illustrations emphasize Aesthetics. Six facets are default-active regardless of category: Safety Compliance, Social Bias, Cultural Fairness, Element Coverage, Object Completeness, and No Hallucination. This is an applicability prior rather than a hard mask: the judge may return `null` for a proposed facet or activate another facet when the sample requires it. Categories without a confident match fall back to learned activation. L1 routing weights affect which facets are proposed, not the equal-weight aggregation of the final image score.

Table 4: Routing groups and default L1 emphasis.

Group	Categories	Default L1 emphasis
Photography	12	Visual Quality 1.5; Real-world Fidelity 1.5
Graphic and Product Design	12	Prompt Alignment 1.5; Creative Generation 1.5
Illustration and Art	15	Aesthetics 1.5; Creative Generation 1.5
3D and CGI	8	Visual Quality 1.5; Aesthetics 1.3
Text and Information	7	Text Rendering 2.0; Prompt Alignment 1.5

C Training Corpus and Governance

Initial supervision covers 29,711 prompt–image pairs across 54 production content categories, about 550 pairs per category. Each pair produces one example per L1 dimension, giving 207,977 group-level SFT examples: 205,924 for training and 2,053 for validation. Table 5 summarizes the data inventory by role.

Provenance and label policy. The pairs come from large-scale production requests to a text-to-image service between June 2025 and June 2026. Each example contains only a prompt and its

generated image. User feedback, preference votes, likes, downloads, and other behavior signals are never used as supervision; all facet-level labels are produced by the evidence-adjudication pipeline. Candidate pairs are screened for privacy, sensitive content, and generation validity, after which the final pairs are sampled randomly within each category. The corpus is random within the eligible production pool, not uniformly random over natural web data, and remains specific to the service's users, generator versions, and collection window.

Table 5: Data inventory by role.

Slice	Source	Count	Role
Distillation	Production prompt–image requests, 54 categories	29,711 pairs	Initial SFT; seven group examples per pair
Targeted refinement	Prompt rewriting and rendered images	≈26,500 prompts / ≈23,800 images per round	Diagnosed weaknesses
Historical replay	Production-pool history	≈4,480 per round	Non-target retention
Evaluation	Seven public judge sets, GenAI-Bench prompts, and audits	Fixed denominators	Held-out testing

Split hygiene. Of the 29,711 pairs, 29,710 keep all seven records in one split; one stratification-boundary pair contributes five training and two validation records. Apart from this recorded boundary case, template families and near-duplicate prompt clusters stay within one split, and weakness-driven prompt rewrites inherit the parent pair's split. The terminal SFT checkpoints, ckpt-4640 and ckpt-2342, are fixed terminal checkpoints rather than checkpoints selected on this record-level validation partition. Refinement checkpoint selection uses a separate development split, and every reported result uses held-out evaluation data.

Publication plan. The supplementary material provides a tested reference implementation of selected processing, objective-construction, and evaluation components; it does not yet include the complete training stack, model weights, or training corpus. Upon acceptance, we will release the trained model weights, the complete processing, training, and evaluation codebase, fixed data splits, and a public subset of the training data, subject to applicable permissions and privacy review. The benchmark test sets used in this work are all public and can be downloaded from their original sources.

D Teacher Pool, Debate, and Adjudication

The teacher pool combines generalist multimodal models with one routed specialist evidence teacher. Kimi K2.6, Qwen3.8-Max, and Gemini 2.5 Flash evaluate the full facet space through the group-level interface. Qwen-VL-Max serves as both specialist router and specialist evidence teacher, using task-specific prompts to produce model-based observations for OCR, text matching, object detection, spatial relations, and safety. These observations are structured specialist checks, not outputs from independent deterministic OCR or detector tools. Challenger and Defender are two roles within one prompted debate call per sample. Adjudication is a separate later call to `gemini-3.1-pro-preview`, and synthesis composes the selected facet records deterministically.

Candidate outputs are aligned by facet. Divergent facets enter evidence-focused analysis: the Challenger searches for missed requested content or visible defects, while the Defender tests whether each alleged defect is visible, relevant, and severe enough. A separate adjudicator resolves the dispute from candidates and evidence, and synthesis builds one complete target. Table 6 reports the matched-adjudicator control. During weakness-driven refinement, Qwen3.8-Max rewrites prompts and Qwen-Image-2 generates images; samples enter training only after evidence verification.

Consensus-but-wrong analysis. Table 7 shows that, among 1,245 audited facets, teachers agree on 971 (78%). Adjudication resolves 226 of 274 disputes correctly (82%). Of the 80 residual errors,

Table 6: Matched-adjudicator control.

Adjudication variant	Evidence hit ↑
Simple teacher aggregation	0.863
Equal-token resampling	0.892
Direct adjudication without Challenger/Defender	0.903
Full evidence adjudication	**0.936**

32 are consensus-but-wrong and 48 are failed disputes. Agreement is therefore a useful default, but consensus errors account for 40% of the residual errors. The complete pipeline substantially improves target reliability without guaranteeing that every agreed judgment is correct.

Table 7: Teacher agreement and final correctness.

Teacher relation	Final correct	Final wrong	Total
Teachers agree	939	32	971
Teachers disagree	226	48	274
Total	1,165	80	1,245

E Prompt and Record Structures

All teacher and evaluation-model instructions are rendered in English per L1 group. We record the operative inputs, decision rules, and output fields below; complete machine-readable templates and frozen facet anchors are provided in the supplementary code.

Group-level judgment. Each prompt–image pair produces one call per L1 group. The call contains: (1) the text prompt, image, L1 group, and facet score anchors; (2) the applicability rule that `null` is allowed only when the prompt does not request the facet and the image does not contain it; (3) evidence and location rules using cells 1–9 for local defects, `[0]` for a global defect, and `[]` when no local defect is found; and (4) one strict-JSON record per facet, ordered as `dimension`, `cells`, `reason`, and `score`. Requested but missing content receives a numeric failure score rather than `null`.

Per-facet anchors. Every facet uses the same five-field schema: `b0` for complete failure, `b12` for a confirmed defect, `b3` for the passing boundary, `b4` for minor residual uncertainty, and `b5` for fully verified quality. The frozen English anchors are the actual model input.

Conflict triggers. Applicability voting is applied first. A strict majority of inapplicable votes stores the facet as `null` and skips debate. Otherwise, numeric candidates remain active; any minority `null` vote counts as an applicability disagreement. An active facet enters debate when that disagreement or any other trigger holds: the numeric score range is at least two bands, aggregated uncertainty is at least 0.4, teacher agreement is below 0.6, or any teacher emits a non-empty review flag. Group agreement is $\max(0, 1 - \text{score range}/2)$, and aggregated uncertainty is one minus the maximum soft-label probability. Facets triggering none of these rules keep their aggregated records and skip debate.

Challenger–Defender debate. One debate call covers all disputed facets in a sample. The Challenger identifies missed requested content or visible defects and returns supporting evidence and a percentage-coordinate region. The Defender checks whether each alleged defect is visible, relevant, and severe enough and records any supported alternative explanation. The strict-JSON output contains `challenger`, `defense`, and `region` for every disputed facet.

Adjudication and normalization. The adjudicator receives the full image, cropped disputed region, candidate judgments, both arguments, and specialist evidence. For each disputed facet it returns `score`, `reason`, `adopted`, `confidence`, and `rejected_sources`. Deterministic synthesis combines adjudicated and non-disputed records into `dimension`, `score`, `reason`, and

`cells`. The score is an integer from 0 to 5 or `null`. Non-disputed facets keep aggregated teacher cells; a sustained Challenger-supported region is mapped to grid cells, and a rejected defect receives an empty cell list.

All three generalists vote on every facet; a specialist votes only where its check produces an observation and otherwise abstains. A facet is stored as `null` only when strictly more than half of voting sources mark it inapplicable. Majority-inapplicable facets skip debate. Otherwise the remaining numeric candidates are aggregated, and any minority `null` vote remains a conflict signal; a numeric score never coexists with `null` in the synthesized record.

F Optimization Settings

All trainable language-model parameters are updated while the vision encoder remains frozen. Training uses BF16, gradient checkpointing, AdamW, cosine schedules, seed 42, and a maximum sequence length of 10,240. Table 8 summarizes the phase-specific configuration. The 27B DPO phase uses PyTorch AdamW; all other phases use fused AdamW. Localized DPO builds single-facet contrastive pairs whose chosen and rejected records differ only on the disputed facet. Let $q = (p, x, g)$ denote the dimension-level model input. The deployed objective adds an SFT-style chosen-response term to standard DPO:

$$\mathcal{L}_{\mathrm{DPO}} = -\mathbb{E} \log \sigma \left(\beta \left[\log \frac{\pi_\theta(y^+|q)}{\pi_{\mathrm{ref}}(y^+|q)} - \log \frac{\pi_\theta(y^-|q)}{\pi_{\mathrm{ref}}(y^-|q)} \right] \right), \tag{6}$$

$$\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{DPO}} + \lambda_{\mathrm{chosen}} \mathcal{L}_{\mathrm{chosen}}, \qquad \mathcal{L}_{\mathrm{chosen}} = \frac{1}{|y^+|} \sum_{t=1}^{|y^+|} - \log \pi_\theta(y_t^+ \mid q, y_{<t}^+). \tag{7}$$

The chosen-response weight is 1.0 in the first localized-DPO stage and 1.5 in the second localized-DPO stage and the 27B student run. GRPO scores complete judgments with six weights: format 0.05, accuracy 0.12, defect 0.40, cells 0.18, rationale evidence 0.05, and rank 0.20. The defect term uses the squared F-measure $(F_1)^2$, and the cell term uses $(F_1)^2$ over grid cells. Over facets applicable in the adjudicated reference, the rationale-evidence term multiplies content-word $(F_1)^2$ by the frozen verifier's image-support score in $[0, 1]$ and averages the products; a missing predicted rationale has content-word $F_1 = 0$. Rank is continuously gated by defect quality, and rewards are normalized within each group of eight completions. Sampling uses temperature 0.7, top-p 0.95, and a 4,096-token completion budget. The 27B run uses 12 GPUs for training and four for rollout. Its GRPO set contains 9,535 error cases and 4,465 resampled correct cases.

Table 8: Training configuration by optimization phase.

Model	Opt.	Records	Epochs / updates	GPUs / global batch	Key settings
8B	SFT	205,924 train + 2,053 val	2 / 4,640	16 / 32	LR 1.414e-5; ZeRO-2; packing; ckpt-4640
8B	DPO	124,319 (stage 1) + 42,517 (stage 2)	2 stages / 5,213	32→16 / 32	LR 5e-7; $\beta = 0.1$; chosen weight 1.0/1.5; ckpt-1329
8B	GRPO	13,999	1 / 3,499	32 / 32	LR 1e-6; $\beta = 0.04$; 8 completions; ckpt-2000
27B	SFT	205,924 train + 2,053 val	2 / 2,342	32 / 64	LR 1.414e-5; ZeRO-2; packing; ckpt-2342
27B	DPO	88,152	1 / 2,755	16 / 32	LR 5e-7; $\beta = 0.1$; chosen weight 1.5; ZeRO-3 + offload
27B	GRPO	14,000	1 / 2,334	12 train + 4 rollout / 48	LR 1e-6; $\beta = 0.04$; 8 completions; ckpt-1500

Both continued-SFT controls start from the distilled model M_0. The chosen-only variant applies plain SFT to the 166,836 preferred responses from the DPO pairs. The matched-budget variant uses the complete refinement data and historical replay under the same total update budget as EVIFORGE, but replaces DPO and GRPO with plain SFT. Table 9 therefore compares the full update strategy with additional SFT alone under the reported update budget.

Table 9: Continued-SFT controls from the distilled model M_0.

Model	Req. miss ↓	Selection ↑	Ordinal ↑	Low-Q ↑	Evidence ↑	Unsup. ↓	Grid ↑
M_0 after distillation	0.380	0.917	0.939	0.945	0.888	0.060	0.389
Continued SFT, chosen only	0.260	0.940	0.947	0.961	0.915	0.040	0.519
Continued SFT, matched budget	0.195	0.958	0.956	0.969	0.929	0.031	0.728
EVIFORGE-8B	**0.140**	**0.982**	**0.974**	**0.992**	**0.962**	**0.019**	**0.856**

For the downstream reward study, four copies of SDXL Base 1.0 are optimized with full-parameter UNet Diffusion-DPO in fp32, changing only the image judge that supplies the preference reward: Qwen3.8-Max, Kimi K2.6, EVIFORGE-8B, or EVIFORGE-27B. GenAI-Bench contributes 1,200 training, 200 calibration, and 200 test prompts. Each judge produces approximately 20,984 retained preference pairs after positive-gap filtering. Training runs for 1,500 steps on eight GPUs with 8-bit AdamW, effective batch 32, $\beta = 1{,}000$, and learning rate 1e-6. The four reward-trained generators and unaligned SDXL each generate two seeds for the same 200 test prompts. Human raters evaluate seven unique direct matchups: the EVIFORGE-27B-rewarded SDXL against all four alternatives and the EVIFORGE-8B-rewarded SDXL against the three non-EVIFORGE alternatives. Each matchup contains 400 comparisons; each comparison receives five independent blinded ratings combined by majority vote. The shared 8B-versus-27B matchup is counted once, giving 2,800 comparisons and 14,000 individual ratings; no automatic judge contributes to these human-win results.

G COMPUTE AND TOKEN WORKLOAD

Table 10 reports training-time token workload. Construction totals are order-of-magnitude output-token estimates derived from the recorded construction counts, while rollout totals are upper bounds from the completion cap. Training-time input tokens were not recorded consistently and are excluded.

Measured batched inference on the 29,711-image corpus pass (207,977 dimension-level calls) gives the following deployment profile. EVIFORGE-8B completes the pass in 4.85 hours on 16 H100 GPUs across two nodes (0.59 amortized s/image; 383 images/GPU-hour). EVIFORGE-27B processes 204,302 valid calls (98.2%) in 7.5 hours on eight L20X GPUs (0.92 amortized s/image; 485 images/GPU-hour), which extrapolates to 7.6 hours for the full pass. A 1,500-call sample averages 14.3k text-input and 4.3k output tokens per image across seven calls; vision tokens are at most 0.6k under the pixel cap. With tensor parallelism 1, the 27B service uses about 112 GiB for weights and KV cache on one 141 GB GPU. The 8B BF16 weights use about 16 GiB, excluding configuration-dependent KV cache. Both judges are therefore locally deployable; these batched rates are not single-request latency.

Table 10: Approximate output-token workload.

Component	Token volume	Basis
Multi-teacher target construction	≈ 0.8B output tokens	≈ 0.6B generalist outputs + ≈ 0.2B debate and adjudication
Weakness-driven data construction	≈ 0.1B output tokens per round	Targeted prompt construction and verification
8B structured rollout	≤ 0.46B sampled tokens	111,992 completions, each capped at 4,096 tokens
27B structured rollout	≤ 0.46B sampled tokens	112,000 completions, each capped at 4,096 tokens

Table 11 reports measured iteration speed and the corresponding wall-clock and GPU-hour estimates under each phase's recorded GPU allocation. The GPUs provide 141 GB of memory each.

Table 11: Approximate training time under the recorded GPU allocation (141 GB per GPU).

Model	Optimization	Updates	Measured s/it	Est. wall-clock	Est. GPU-hours
8B	SFT	4,640	5.65	≈7.3 h	≈117
8B	DPO	5,213	4.26	≈6.2 h	≈172
8B	GRPO	3,499	12.76	≈12.4 h	≈397
27B	SFT	2,342	44.3–82.4	≈28.8–53.6 h	≈922–1,716
27B	DPO	2,755	62.4	≈47.8 h	≈764
27B	GRPO	2,334	119	≈77.2 h	≈1,234

H Allocation, Acceptance, and Additional Controls

For weakness cell $h = (f, c, e, d)$, the severity-weighted diagnostic error rate and round allocation are

$$W_t(h) = \frac{\sum_{z \in \mathcal{D}_{\text{diag}}} w(z)\, \mathbb{I}[h \text{ occurs on } z]}{\sum_{z \in \mathcal{D}_{\text{diag}}} \mathbb{I}[z \in \Omega_h]}, \tag{8}$$

$$\pi_t(h) = (1-\eta)\frac{(W_t(h)+\epsilon)^\alpha}{\sum_{h' \in \mathcal{H}_t^+}(W_t(h')+\epsilon)^\alpha} + \eta \pi_{\text{cover}}(h), \qquad B_t(h) = B_t \pi_t(h). \tag{9}$$

$\mathcal{H}_t^+$ contains cells with at least τ diagnostic observations. The coverage term assigns equal mass to each eligible facet–scene pair and divides that mass equally among its eligible error-type and direction cells. If no cell has enough support, we set $\pi_t = \pi_{\text{cover}}$, so the full budget follows this coverage distribution. Table 12 lists the fixed settings, which are shared by both backbones unless noted.

Table 12: Weakness-driven allocation settings.

Parameter	Value
Sharpness α	1.0
Coverage mix η	0.20
Smoothing ϵ	0.01
Minimum support τ	30 diagnostic samples
π_{cover}	Uniform over facet–scene pairs, then over eligible error-type/direction cells
Round budget	3 rounds for 8B; 2 rounds for 27B
Per-round budget B_t	≈26,500 targeted prompts
Replay per round	4,480 stratified examples
Retention threshold	0.90

Acceptance criterion. A checkpoint advances only if non-target retention on development observations of the 400 non-target cases is at least 0.90. Among advancing candidates, each round maximizes target-case gain. The final model is the accepted checkpoint with the highest development-set rank correlation under the same retention floor. These checkpoint-selection observations are separate from the registered test-split cases reported in Figure 4. This selects M_3 for EviForge-8B; M_3 also matches or exceeds all earlier rounds on required-facet coverage and localization.

Per-seed allocation results. Table 13 reports targeted gain over three data-shuffle seeds for the random, frozen-profile, and re-diagnosed controls; Table 14 gives the complete allocation and update controls. Retention is reported for seed 42. The current-model profile exceeds the frozen profile by 0.10 on every seed.

The rank reward scores within-image ordinal consistency. The candidate judgment orders evaluated elements by its own facet scores, while the reference ordering comes from adjudicated scores. The normalized rank correlation is continuously gated by the defect term so that the model cannot improve ordering by suppressing low scores. Rewards are group-normalized over eight completions per prompt.

Table 13: Targeted gain across data-shuffle seeds.

Seed	Random	Frozen M_0 profile	Re-diagnosed profile
1	40 (0.20)	70 (0.35)	90 (0.45)
2	30 (0.15)	80 (0.40)	100 (0.50)
3	50 (0.25)	90 (0.45)	110 (0.55)
Mean	40 (0.20)	80 (0.40)	100 (0.50)

Table 14: Complete weakness-driven refinement controls for the 8B student. Allocation rows use the same localized-DPO schedule and cumulative budget. Random, frozen-profile, own-profile, and auxiliary-baseline gains are three-seed means; other rows use seed 42.

Test	Variant	Targeted gain ↑	Retention ↑
Allocation	Random prompt expansion	0.20	0.900
	Fixed preset curriculum	0.35	0.925
	Shuffled weakness profile	0.35	0.875
	Other-student profile	0.35	0.900
	Frozen initial profile	0.40	0.925
	Own-student profile, re-diagnosed	0.50	0.975
Baselines	Facet-granularity profile	0.44	0.953
	Disagreement-guided sampling	0.32	0.928
	Candidate-pool mining	0.27	0.923
Update	Without localized DPO	0.30	0.925
	Without full-judgment GRPO update	0.35	0.900
	Without historical replay	0.50	0.800
	Complete refinement	**0.65**	**0.975**

I Dataset-to-Facet Mapping and Evaluation Protocol

Eight public evaluation sets are used. Seven judge-evaluation sets come from six benchmarks, and GenAI-Bench supplies prompts for the reward-guided generation and human study. Table 15 lists their roles. Each annotation is mapped to the facet vocabulary before scoring.

Table 15: Evaluation datasets and primary roles.

Dataset	Evaluation unit	Human evidence	Primary role
VisionRewardDB	Image / attribute	Applicability and ordered quality levels	Facet activation; score ordering
ImageRewardDB	Image / prompt	Pointwise quality and same-prompt preference	Quality detection; pairwise order
RichHF-18K	Image / defect	Defect descriptions and spatial evidence	Rationale evidence; localization
EvalMuse-40K	Image / element	Element-level failures and locations	Required facets; evidence grounding
Qwen-Image-Bench	Generator / prompt	Expert generator ranking	Generator-level ranking
MJ-Bench quality	Image pair	Human quality preference	Pairwise decisions
MJ-Bench bias	Controlled group	Matched demographic variants	Score stability
GenAI-Bench	Prompt / generated pair	Blinded human preferences	Reward-guided generation

All image-level tasks use dimension-balanced aggregation. Every model is scored on the same split under the same predefined denominator rules; selection and unsupported-claim rates use each model's activated-facet count L_a. Parsing failures count as failed trials. Automatic pairwise comparisons randomize left and right positions; ties remain in the denominator and do not count as wins. Human generator evaluation uses seven unique direct matchups anchored by the two EviForge-rewarded generators. Confidence intervals resample prompt clusters rather than individual images.

AI-assisted audits are reported as blinded evaluations with a human-reviewed stratified subset, inter-annotator agreement, and AI–human agreement; AI reviewers are never labeled as human ground truth.

Before scoring model outputs, source annotations are mapped to the fixed 60-facet vocabulary. The required-facet set contains 4,439 human-confirmed image–facet requirements from the VisionRewardDB test split and is shared by every model. The separate evidence audit contains 1,245 claim-level evidence points derived from human-written benchmark annotations. Table 1 reports combined grid coverage over 1,697 annotated defect cells from the RichHF quality (886) and artifact (811) subsets. Figure 4(c) instead tracks alignment-localization coverage on the 811-cell artifact subset across refinement checkpoints. These evaluation inputs are fixed independently and are not interchangeable denominators.

Table 16: Evaluation inputs and denominator rules.

Setting	Metric family	Evaluation input	Denominator
Core capabilities	Required-facet miss	VisionRewardDB test split	4,439
Core capabilities	Evidence hit	Blinded audit set	1,245
Core capabilities	Selection / unsupported	400-image blinded audit	L_a
Core capabilities	Score monotonicity	VisionRewardDB ordered pairs	114
Core capabilities	Low-quality detection	ImageRewardDB fixed threshold	128
Core capabilities	Grid coverage	RichHF quality + artifact	1,697
Refinement trajectory	Alignment localization	RichHF artifact	811
Generator ranking	Rank ρ	QIB generators	18
Pairwise / stability	Accuracy; NDS / GES	MJ-Bench quality / bias	1,120 / 1,458
Reward study	Human win	Seven direct generator matchups	2,800
Distillation validation	Required-facet miss	Equal-data runs	4,439
Distillation validation	Evidence / unsupported audit	Fixed audit points	1,245
Refinement validation	Targeted gain / retention	Registered cases	200 / 400

J Metric Definitions and Supplementary Results

Core-judgment metrics form four capability groups. Facet selection uses required-facet miss and selection appropriateness. Score consistency uses ordinal monotonicity and low-quality detection, reported in Table 17. Evidence grounding uses evidence hit and unsupported claims. Defect localization uses grid coverage. The paired metrics within the first three groups prevent shortcuts: activating every facet can reduce misses but lowers selection appropriateness, while adding more explanation can recover evidence but also introduce unsupported claims.

Table 17: Score-consistency results.

Model	Ordinal ↑	Low-quality detection ↑
Qwen3-VL-8B base	0.877	0.844
Qwen3-VL-8B after distillation	0.939	0.945
EVIFORGE-8B	**0.974**	**0.992**
Qwen3.8-27B base	0.982	0.945
Qwen3.8-27B after distillation	0.947	0.945
EVIFORGE-27B	**0.991**	**0.992**
Q-Judger	0.947	**1.000**
Kimi K2.6	0.956	0.992
Qwen3.8-Max	0.956	0.984

Ratio metrics follow the denominator rules in Table 18. For the tie-free expert ranking of 18 generators, Spearman's ρ is $1 - \Sigma d^2/969$. NDS is $1 - \sigma/\mu$ over min–max-normalized scores, and GES is $1 - G_{\mathrm{Gini}}$, where G_{Gini} is the standard Gini coefficient of the same normalized scores. Human win counts blinded head-to-head decisions by human raters.

Table 18: Metric definitions and denominators.

Metric	Denominator	Definition
Required-facet miss	4,439	Human-confirmed facets not activated
Selection appropriateness	L_a	Activated facets supported by prompt or image
Ordinal monotonicity	114	Ordered pairs with correct score direction
Low-quality detection	128	Low-quality images correctly flagged
Evidence hit	1,245	Human-required evidence points recovered
Unsupported claim	L_a	Unsupported claims per activated facet
Grid coverage	1,697	Annotated defect cells covered
Rank ρ	18	Spearman correlation with expert ranking
Pairwise accuracy	1,120	Correct same-prompt choices
Bias gap	1,458	Unstable demographic-variant units
Human win	400 per matchup	Wins in each blinded generator matchup

Precision–recall robustness. For EVIFORGE-8B, low-quality detection has precision 0.941, recall 0.992, and F_1 0.966. Grid localization has precision 0.837, recall 0.856, and F_1 0.846. Its unsupported-claim and unsupported-facet rates are 0.019 and 0.017, respectively; the recall-focused metrics are therefore not produced by flagging everything.

Activation controls. Table 19 reports the activation controls. Under a fixed required-facet mask, the 18-generator rank correlation is 0.9174, within 0.005 of the learned-mask value 0.9216. The mean activated-facet count changes by only -2% from M_2 to M_3. Forcing all 60 facets active gives zero required-facet miss but lowers selection appropriateness to 0.668 and rank correlation to 0.862. A category-rule mask misses sample-specific facets, with a miss rate of 0.451.

Table 19: Facet-activation controls.

Activation strategy	Miss ↓	Selection ↑	QIB ρ ↑
All facets active	0.000	0.668	0.862
Category-rule mask	0.451	0.920	0.901
Learned activation, EVIFORGE-8B	0.140	0.982	0.922

Confidence intervals. Table 20 reports selected 95% confidence intervals. Small-denominator proportions use Clopper–Pearson intervals, shared-image and preference metrics use cluster resampling, and correlations use Fisher-z intervals. The three repeated allocation controls report per-seed targeted gain and standard deviation rather than treating three seeds times 200 cases as 600 independent trials. Intervals are reported per model; interval overlap is not used as a significance test.

Table 20: Selected 95% confidence intervals.

Metric	Point	95% CI
Required-facet miss, 4,439	0.140	[0.130, 0.150]
Evidence hit, targets, 1,245	0.936	[0.921, 0.949]
Ordinal monotonicity, 114	0.974	[0.925, 0.995]
Low-quality recall, 128	0.992	[0.957, 1.000]
Low-quality precision, 135	0.941	[0.887, 0.974]
Weakness ρ, 60 facets	0.62	[0.434, 0.755]
QIB rank ρ, 18	0.922	[0.799, 0.971]

K HUMAN-AUDIT PROTOCOL

Selection appropriateness and unsupported-claim rate use a blinded audit of 400 images per model. Qwen3.8-Max receives a dedicated audit prompt, anonymized model outputs, the source prompt and image, and the applicable facet rubric; it labels each activated facet as supported, partially relevant, or unsupported and checks each rationale claim against visible image content. Qwen3.8-Max also serves as one generalist teacher and as a comparison baseline. The teacher–student comparison tests

whether the complete distillation pipeline transfers supervision beyond using this single teacher alone. Because the same model performs the two AI-assisted audits, these audit outputs are not treated as independent human ground truth; the stratified human subset and benchmark-native evidence and localization annotations provide complementary checks. Unsupported claims are counted individually and normalized by activated-facet count L_a; facets containing at least one define the unsupported-facet rate. A stratified 150-image subset per model receives three independent human ratings under the same three-way rubric (Fleiss $\kappa = 0.71$). AI–human agreement is $\kappa = 0.68$ with 86.4% raw agreement, and the human subset preserves the full-audit ordering ($\rho = 0.94$; Table 27). This human subset audits facet support and unsupported-facet labels. Evidence hit is evaluated separately on the fixed 1,245 claim-level evidence points, while localization is evaluated against the benchmark grid cells in Table 16. Table 26 reports the fully human downstream preference study.

L Per-Model Numerators and GenAI-Bench Outcomes

Tables 21–23 report the per-model counts underlying the main results. Table 24 reports all ten head-to-head automatic generator matchups, and Table 25 aggregates each generator's four matchups. Under both judges, the EviForge-27B-rewarded model directly defeats the EviForge-8B-rewarded model, which directly defeats each of the other three generators; the two tournaments therefore agree on the first- and second-place models. Table 26 reports the seven unique human matchups anchored by the two EviForge-rewarded generators.

Table 21: Core-capability per-model numerators. B, D, and E denote the backbone, the checkpoint after evidence-adjudicated distillation, and final EviForge. Ratios use the denominator rules in Table 18; selection and unsupported rates use each model's activated-facet count L_a.

Metric	8B-B	8B-D	**8B-E**	27B-B	27B-D	**27B-E**	Q-J	Kimi	QMax
Req. miss (/4439)	3036	1689	621	2483	2054	1127	1268	1100	1662
Selection app. ($/L_a$)	13215	13972	16125	14532	14972	16081	12194	14023	15629
Ord. mono. (/114)	100	107	111	112	108	113	108	109	109
Low-Q (/128)	108	121	127	121	121	127	128	127	126
Evidence hit (/1245)	817	1106	1198	997	1079	1197	917	1181	1211
Unsup. claim ($/L_a$)	1157	912	305	821	734	291	945	225	237
Unsup. facet ($/L_a$)	1081	838	280	753	674	267	919	207	218
Grid cover. (/1697)	378	660	1452	434	909	1279	756	1054	1092
L_a	16123	15243	16424	15930	15782	16299	15236	15267	16296

Table 22: Aggregate-score values. B, D, and E denote the backbone, the checkpoint after evidence-adjudicated distillation, and final EviForge. Rank ρ is Spearman over 18 generators; pairwise and bias entries are numerators over predefined denominators.

Metric	8B-B	8B-D	**8B-E**	27B-B	27B-D	**27B-E**	Q-J	Kimi	QMax
Rank ρ (n=18)	0.7853	0.8782	0.9216	0.8700	0.9009	0.9567	0.8741	0.9112	0.9278
Pair acc. (/1120)	1019	1078	1086	1063	1100	1109	1034	1087	1095
NDS	0.931	0.924	0.932	0.940	0.939	0.951	0.905	0.883	0.885
GES	0.965	0.959	0.963	0.967	0.966	0.967	0.947	0.935	0.935
Bias gap (/1458)	165	1	2	0	0	1	28	72	18

Table 23: Distillation-control numerators. Ratios in Table 2 use denominators of 4,439 required facets, 114 ordered score pairs, and 1,245 audit points.

Level	Variant	Req. miss (/4439)	Ord. mono. (/114)	Evidence hit (/1245)	Unsup. (/1245)
Evaluation	Single Generalist Teacher (Qwen3.8-Max)	1894	106	1023	61
	Simple teacher aggregation	1482	104	1075	48
	Full evidence adjudication	1032	109	1165	36
Student	Qwen3.8-Max-only distillation	2178	106	972	76
	Simple-aggregation distillation	2041	104	1012	65
	w/o Specialist Evidence Teachers	1949	106	1035	58
	w/o Evidence Debate	1853	107	1061	51
	w/o separate adjudication	1765	106	1071	45
	Complete distillation	1689	107	1106	38

Table 24: Automatic head-to-head generator comparisons (400 trials per matchup).

		EviForge-8B judge			Qwen3.8-Max judge		
Generator A	Generator B	A win	B win	Tie	A win	B win	Tie
Unaligned	Qwen3.8-Max reward	167	232	1	149	251	0
Unaligned	Kimi K2.6 reward	210	189	1	171	229	0
Unaligned	EviForge-8B reward	166	234	0	132	268	0
Unaligned	EviForge-27B reward	134	266	0	137	263	0
Qwen3.8-Max reward	Kimi K2.6 reward	234	166	0	239	161	0
Qwen3.8-Max reward	EviForge-8B reward	196	203	1	187	213	0
Qwen3.8-Max reward	EviForge-27B reward	147	253	0	144	256	0
Kimi K2.6 reward	EviForge-8B reward	149	251	0	149	251	0
Kimi K2.6 reward	EviForge-27B reward	137	263	0	134	266	0
EviForge-8B reward	EviForge-27B reward	179	221	0	169	231	0

Table 25: Aggregate automatic generator rankings. Each row aggregates 1,600 comparisons against the other four generators; models are ranked by win rate, with ties retained in the denominator but not counted as wins.

	EviForge-8B judge				Qwen3.8-Max judge			
Generator	Win	Loss	Tie	Win rate	Win	Loss	Tie	Win rate
Unaligned SDXL	677	921	2	42.3%	589	1,011	0	36.8%
Qwen3.8-Max-rewarded	809	789	2	50.6%	821	779	0	51.3%
Kimi K2.6-rewarded	641	958	1	40.1%	673	927	0	42.1%
EviForge-8B-rewarded	867	732	1	54.2%	901	699	0	56.3%
EviForge-27B-rewarded	1,003	597	0	**62.7%**	1,016	584	0	**63.5%**

The automatic ranking is computed separately for each judge by sorting the five generators by aggregate win rate, $W/(W+L+T)$, over the complete round robin. The human study contains seven unique matchups: EVIFORGE-27B against all four alternatives and EVIFORGE-8B against the three non-EVIFORGE alternatives, with their shared matchup counted once. These direct outcomes determine the top-two order; no order is assigned among the three non-EVIFORGE generators because they are not compared with one another by human raters.

Table 26: Direct human comparisons between reward-trained generators (400 blinded comparisons per matchup; five binary votes per comparison, majority aggregation).

Generator A	Generator B	A win	B win	A rate
EVIFORGE-27B reward	Unaligned	273	127	68.3%
EVIFORGE-27B reward	Qwen3.8-Max reward	259	141	64.8%
EVIFORGE-27B reward	Kimi K2.6 reward	261	139	65.3%
EVIFORGE-27B reward	EVIFORGE-8B reward	229	171	57.3%
EVIFORGE-8B reward	Unaligned	254	146	63.5%
EVIFORGE-8B reward	Qwen3.8-Max reward	221	179	55.3%
EVIFORGE-8B reward	Kimi K2.6 reward	234	166	58.5%

Across its four matchups, the EVIFORGE-27B-rewarded generator wins 1,022 of 1,600 comparisons (63.9%). From the EVIFORGE-8B side, the corresponding total is 880 of 1,600 (55.0%), including its 171 wins in the shared matchup with EVIFORGE-27B. The observed direct human outcomes therefore support EVIFORGE-27B reward $>$ EVIFORGE-8B reward $>$ each non-EVIFORGE alternative.

Table 27: Human-only audit subset (150 images per model; three blind raters). Unsupported rates are facet-level; the final column reports full-audit selection appropriateness / unsupported-facet rate.

Model	Supported ↑	Unsup. ↓	Full audit: selection / unsup. facet
8B base	0.790	0.085	0.820 / 0.067
8B distilled	0.900	0.070	0.917 / 0.055
EVIFORGE-8B	0.970	0.022	0.982 / 0.017
27B base	0.890	0.060	0.912 / 0.047
27B distilled	0.935	0.052	0.949 / 0.043
EVIFORGE-27B	0.980	0.020	0.987 / 0.016
Q-Judger	0.765	0.080	0.800 / 0.060
Kimi K2.6	0.900	0.017	0.919 / 0.014
Qwen3.8-Max	0.945	0.017	0.959 / 0.013

M LIMITATIONS AND FUTURE EXTENSIONS

The present study has four main boundaries, each pointing to future work. First, the current 60-facet hierarchy cannot exhaust open-ended or culturally specific quality criteria; it can be expanded as new evaluation needs emerge. Second, the precision of evidence grounding and the reliability of model-based supervision can be improved with finer region grounding, additional specialist visual tools, and stronger independent verification. Third, Qwen3.8-Max serves as a teacher and as the auditor for two AI-assisted metrics, so model-specific audit bias may remain despite blinded outputs and the human subset checks. Fourth, the current experiments focus on text-to-image generation; task-specific facets and evidence tools can extend and test EVIFORGE on text-and-image-to-image editing and text-to-video generation.