

When Rubrics Fail: Error Enumeration as Reward for Reference-Free RL Post-Training

Wisdom Ikezogwo^{1,2*}, Mehmet Saygin Seyfioglu², Ranjay Krishna¹, and Karim Bouyarmane²

¹ University of Washington, Seattle, USA
 {wisdomik, ranjay}@cs.washington.edu

² Amazon, Seattle, USA
 {msefyiog, bouykari}@amazon.com

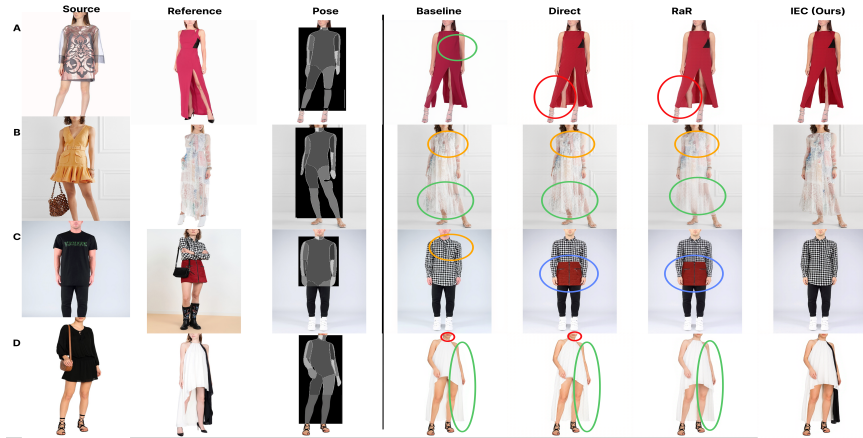


Fig. 1: Qualitative comparison of reward designs on non-flat references: Given the input, IEC avoids *attribute preservation* errors across 3 examples, *garment transfer* failures in example C, *source integrity* issues in A and D, and *realism* errors in examples B and C, over all other reward designs. Per-category radar plots are in the supplementary material (Fig. 2).

Abstract. Reinforcement learning with verifiable rewards (RLVR) and Rubrics as Rewards (RaR) have driven strong gains in domains with clear correctness signals and even in subjective domains by synthesizing evaluation criteria from ideal reference answers. But many real-world tasks admit multiple valid outputs and lack the single ideal answer that rubric generation depends on. We identify this *reference-free* setting as a gap in current post-training methods and propose **Implicit Error Counting (IEC)** to fill it. Instead of checking what a response gets right against a rubric, IEC enumerates what it gets wrong, applying severity-weighted scores across task-relevant axes and converting them into calibrated

* Work done during internship at Amazon.

per-aspect rewards. We show that naïve explicit enumeration is too noisy for stable optimization, and that two design choices: implicit score emission and group calibration are necessary to make error counting a reliable reward. As a case study, we validate IEC on virtual try-on (VTO), a domain that is simultaneously too constrained for holistic scoring and too permissive for rubric-based evaluation: subtle garment errors are unacceptable, yet many output variations are correct. We introduce Cascaded Error Counting (CEC) as an evaluation metric, which tracks human preferences well (60% top-1 vs. 30% others), and curate Mismatch-DressCode (MDressBench), a benchmark with maximal attribute mismatch to stress-test reward designs. On MDressBench, IEC outperforms RaR across all metrics (CEC: 5.31 vs. 5.60 on flat references; 5.20 vs. 5.53 on non-flat). On VITON-HD and DressCode, IEC matches or surpasses six baselines on 6 of 8 perceptual metrics. These results suggest that when ideal answers are unavailable, counting errors provide a stronger signal than constructing rubrics.

Keywords: RL · Error Enumeration as Reward · Reference-Free · Virtual Try-On

1 Introduction

Reinforcement learning has become a standard post-training step for aligning generative models with human expectations. In domains with verifiable outcomes—mathematics, code generation, game playing—reward is straightforward: a scoring function or test environment confirms correctness [26,31]. Reinforcement Learning with Verifiable Rewards (RLVR) [1, 9, 28, 31, 33, 45] produces dramatic gains precisely because the reward signal is cheap, unambiguous, and dense.

Real-world tasks are rarely so clean. Medical diagnosis, scientific reasoning, and creative generation require nuanced, multi-criteria evaluation with no single correct answer. Rubrics as Rewards (RaR) [12] addresses this gap: given a prompt and an *ideal reference answer*, it synthesizes instance-specific rubric criteria and scores responses against them. RaR provides structured, interpretable supervision that outperforms coarse Likert scoring. Its core assumption, however, is that an ideal answer exists and can ground rubric generation. When it does, as in medical QA with expert-written references, rubrics capture subtle quality dimensions that binary correctness cannot.

But many domains lack ideal answers altogether. In embodied control, multiple action sequences can achieve the same goal. In open-ended design tasks, “correctness” is defined not by what the output matches, but by what it avoids—wrong colors, broken geometry, identity drift. We refer to this as the **reference-free** setting: the task is subjective enough that rubrics are needed, yet no ideal output exists to generate them from. In this setting, the nature of quality assessment inverts. Rather than asking “does this response satisfy a checklist derived from an ideal answer?” one asks “what specific errors does this response exhibit?” When multiple outputs are plausible, enumerating failures is more stable than

enumerating successes, because the space of errors is typically smaller and more structured than the space of valid completions.

We test this hypothesis with **Implicit Error Counting (IEC)**, a reward framework for reinforcement post-training in reference-free domains. Given a candidate output and its conditioning inputs, a multimodal judge enumerates errors across task-relevant evaluation axes, weights them by severity, and emits calibrated per-aspect scores. The “implicit” qualifier is critical: asking the judge to expose raw error lists (Explicit Error Counting, or EEC) introduces high variance from inconsistent enumeration, where near-identical outputs receive different error counts due to surface-form variation in the judge’s language. IEC instead asks the judge to *internalize* the counting and output only scores, preserving the conceptual grounding of error enumeration while stabilizing the reward signal. A lightweight group calibration step further reduces prompt-to-prompt scale drift. Notably, IEC is also more efficient: it requires a single judge call per candidate, whereas RaR requires two (rubric generation followed by evaluation).

We validate IEC on virtual try-on (VTO) as a case study that exemplifies the reference-free challenge. VTO synthesizes images of users wearing target garments [4, 20, 25, 29, 51], and its failure modes are simultaneously subtle and consequential: a slightly wrong sleeve length, a discontinuous pattern, or a shifted neckline can render a result unusable. Crucially, many distinct outputs are correct; the garment can drape differently, lighting can vary, and fine details can differ. So no single ideal answer grounds rubric generation. At the same time, errors are *enumerable* and *localizable*: they fall along well-defined axes (garment transfer, attribute preservation, realism, lighting, source integrity) and can be detected by multimodal judges. Supervised objectives on *narrow* datasets [4, 25] cannot cover the combinatorial space of poses, body shapes, textures, and garment types, and tend toward shortcut learning [18, 32, 39], especially copy-paste behavior in masked-image editing. This motivates on-policy post-training that improves behavior without new paired supervision.

To stress-test reward designs under distribution shift, we curate **Mismatch-DressCode (MDressBench)**, a benchmark of 700 source-reference pairs selected for maximal attribute disagreement (e.g., short-sleeve source with long-sleeve reference). We also introduce **Cascaded Error Counting (CEC)**, an evaluation metric that builds a shared error vocabulary across candidates, enabling consistent measurement without ground truth. In a human study validating CEC as a selection metric, it achieves 60% top-1 accuracy in recovering annotator preferences, compared to 30% for both direct scoring and RaR-based selection.

Across MDressBench, IEC consistently outperforms both direct scoring and RaR on all metrics, achieving relative improvements of 5.96% (CEC), 2.32% (garment transfer), 3.48% (attribute preservation), 4.33% (realism), 3.3% (lighting), and 0.6% (source integrity) over RaR on non-flat references. On VITON-HD and DressCode, IEC outperforms all supervised and RL baselines on 6 of 8 perceptual metrics. Our ablations confirm that EEC underperforms IEC, and that group calibration provides consistent gains.

2 From Rubrics to Error Counting

We first formalize the gap between rubric-based and reference-free reward settings (§2.1), then present error counting as a general reward framework (§2.2), and finally instantiate it for virtual try-on (§2.3).

2.1 The Reference-Free Gap

Rubrics as Rewards [12] defines a structured reward function for each prompt x via a set of k rubric items $\{(w_j, c_j)\}_{j=1}^k$, where w_j is the weight of criterion j and $c_j : (x, \hat{y}) \mapsto \{0, 1\}$ is a binary check. The reward is:

$$r_{\text{rubric}}(x, \hat{y}) = \frac{\sum_{j=1}^k w_j \cdot c_j(x, \hat{y})}{\sum_{j=1}^k w_j} \quad (1)$$

Rubric generation requires a reference answer y^* to ground the criteria: each c_j encodes a property that a high-quality response should satisfy, derived from analyzing y^* . When y^* exists (e.g., expert-written medical answers), this yields precise, instance-specific supervision.

The reference-free setting arises when three conditions hold simultaneously:

1. **Multiple valid outputs exist.** Many responses $\hat{y}$ may be equally correct, so no single y^* captures what “good” means.
2. **Quality is defined by the absence of errors.** An output is good if it follows the conditioning and has no observable errors, not because it matches a particular reference.
3. **Errors are enumerable and localizable.** Failure modes cluster along well-defined axes and can be detected by inspection.

These conditions are satisfied whenever the space of acceptable outputs is large but the space of errors is structured, a property shared by visual generation, embodied control, and open-ended design tasks. Under these conditions, rubric generation from a reference answer either produces criteria that are too generic (because no single y^* captures the range of valid outputs) or too specific (because y^* encodes one valid solution and penalizes equally valid alternatives). Error counting exploits the structure of these domains by inverting the evaluation: instead of scoring how much an output resembles an ideal, it scores how few detectable errors it contains.

2.2 Error Counting as a Reward Framework

Let x denote the conditioning input and $\hat{y}$ a candidate output. We define A task-relevant evaluation axes $\{a_1, \dots, a_A\}$. For each axis a , a judge function $f_a(x, \hat{y})$ returns a score in $[0, 1]$ by enumerating errors along that axis and mapping error counts to a monotonically decreasing score.

Explicit Error Counting (EEC). The judge produces a list of errors $\mathcal{E}_a = \{(e_1, s_1), \dots\}$ for each axis, where e_i is an error label and $s_i \in \{\text{minor}, \text{major}\}$ is its severity. The axis score is computed as a deterministic function of the weighted count:

$$R_a^{\text{explicit}} = g\left(\sum_{(e,s) \in \mathcal{E}_a} w(s)\right) \quad (2)$$

where g is a monotonically decreasing mapping and $w(\cdot)$ assigns severity weights. This is conceptually clean but suffers from high variance: non-deterministic judges produce different error labels for near-identical outputs, destabilizing within-group rewards. Appendix B.6 illustrates this with a concrete training group where EEC assigns a perfect score to a generation with visible errors while penalizing better candidates. See Appendix D.3 for prompt details.

Implicit Error Counting (IEC). We instead instruct the judge to *internally* count errors for each axis and output only a calibrated score $R_a^{(i)} \in [0, 1]$, along with a short error summary for interpretability. Each error monotonically reduces the axis score; catastrophic errors exponentially reduce it toward zero. We define *catastrophic* errors as wrong garment type, identity change, or a missing major garment region (see Appendix B.5 / D.1). The judge implicitly determines the per-error penalty weight. The aggregate reward is $R_i^{\text{implicit}} = \frac{1}{|A|} \sum_{a=1}^{|A|} R_a^{(i)}$.

The key distinction from EEC is that the reward depends on the judge’s *internal* assessment rather than on parsed error labels, making it robust to surface-form variation of the judge, and brittleness of fixed weights [12]. The error summary provides interpretability without coupling directly to reward computation.

Group Calibration. Reward magnitudes vary across prompts due to difficulty and evaluator scale, introducing unnecessary variance in RL updates. We apply group-wise calibration using robust statistics. For a group reward vector $\mathbf{R}$:

$$z_i = \frac{R_i - \text{median}(\mathbf{R})}{1.4826 \cdot \text{MAD}(\mathbf{R}) + \varepsilon} \quad (3)$$

$$R_i^{\text{cal}} = \text{floor} + (1 - \text{floor}) \cdot \sigma(\alpha \cdot z_i) \quad (4)$$

where MAD is the median absolute deviation, 1.4826 rescales to standard-deviation scale under Gaussian assumptions, floor ≈ 0.05 prevents zero rewards, and σ is the logistic function with steepness $\alpha = 1$. This monotone transform preserves within-group ordering while reducing prompt-to-prompt scale variability.

2.3 Instantiation: Virtual Try-On

We formulate VTO as an image-conditioned inpainting problem. Given a source image I_s , reference/target garment I_r , mask I_m , and pose I_p , the model generates I_t where I_r appears on I_s while preserving identity, pose, and background.

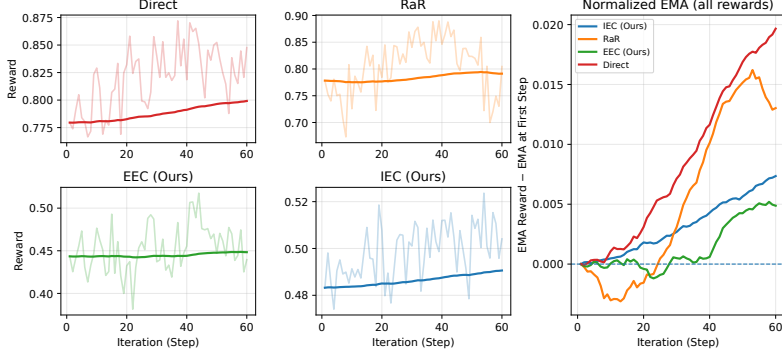


Fig. 2: Training dynamics for four reward designs during GRPO post-training. **Left 2×2 :** Per-method curves showing raw rewards (faded) and EMA (solid, $\alpha = 0.99$). **Right:** Normalized EMA showing relative improvement. IEC achieves smooth, monotonic improvement. Direct and RaR show high variance. EEC regresses early, confirming that explicit enumeration destabilizes optimization.

Architecture. We build on a rectified-flow backbone [21, 23], following the token concatenation scheme of DiT-VTON [20]. Each image is encoded into a latent space, patchified, and concatenated: $c = x_t \circ x_r \circ x_e$, where x_t is the noisy latent, x_r is the reference latent, and x_e is the masked source with pose stitched on. The flow-matching objective trains velocity $v_\theta(x_t, t, c)$ to approximate $v = x_1 - x_0$.

GRPO Post-Training. We apply Group Relative Policy Optimization [31] adapted to flow models [22, 45]. For each condition c , we sample K candidates and compute group-normalized advantages:

$$\hat{A}_i = \frac{R(x_i, c) - \text{mean}(\{R(x_j, c)\}_{j=1}^K)}{\text{std}(\{R(x_j, c)\}_{j=1}^K)} \quad (5)$$

Evaluation Axes. For VTO, we define $A = 5$ axes: (i) *garment transfer* (placement, sleeve/hem length, neckline, proportions); (ii) *attribute preservation* (color, pattern, texture); (iii) *realism* (drape, boundaries, halos, distortions); (iv) *lighting consistency*; and (v) *source integrity* (face, hair, background unchanged). Each axis is scored by the IEC judge (Equation 2.2), and the aggregate reward drives GRPO updates.

2.4 Baseline Rewards

We compare IEC against two standard reward designs, both evaluated by the same multimodal judge.

Direct Scoring. The judge outputs a scalar $R_i^{\text{direct}} \in [0, 1]$ for overall quality, guided by scoring bands: $[0.8, 1.0]$ for strong results, $[0.5, 0.8)$ for mixed quality, $[0.2, 0.5)$ for poor results, $[0.0, 0.2)$ for failures. It is efficient; however, it loses a fine-grained sensitivity to perceivable errors. See Appendix D.1.

Rubrics-as-Rewards (RaR). Following [12], we generate a prompt-specific rubric with criteria across categories {Essential, Important, Optional, Pitfall} with weights w_j . The judge evaluates each criterion with binary pass/fail. Let $\mathcal{P}_i = \{j : p_{ij} = 1, w_j > 0\}$ denote satisfied items and $\mathcal{F}_i = \{j : p_{ij} = 0, w_j < 0\}$ violated pitfalls:

$$R_i^{\text{rubric}} = \text{clip}\left(\frac{\sum_{j \in \mathcal{P}_i} w_j + \sum_{j \in \mathcal{F}_i} w_j}{\sum_{j: w_j > 0} w_j}; 0, 1\right) \quad (6)$$

This requires two judge calls (rubric generation + evaluation) and, crucially, requires generating rubrics *without an ideal reference answer* in the VTO setting.

Reward Cost. IEC and Direct scoring each require a single judge call per candidate, whereas RaR requires two (rubric generation followed by evaluation). For a group of $K=12$ candidates, this means IEC uses 12 judge calls per training step compared to 24 for RaR, a $2\times$ reduction in judge compute, while achieving stronger performance across all metrics.

2.5 Cascaded Error Counting as a Metric

Plain error counting exhibits high variance as a metric because the label space drifts across images. Cascaded Error Counting reduces this by sharing an evolving error vocabulary across candidates within a group. Without pooling, a judge may enumerate zero errors for one candidate and seventeen for a near-identical one (see Appendix B.6); re-evaluating against a shared pool forces the judge to reconsider missed errors. We process candidates in blocks:

1. **Pool Phase:** The judge returns preliminary per-image error sets $\mathcal{E}_{i \in B} = \{(e_i, \text{sev}), \dots\}$. We canonicalize and merge into a pool $\mathcal{D}^{(0)}$ with semantic duplication.
2. **Verification Phase:** We re-evaluate images conditioned on $\mathcal{D}^{(0)}$, instructing the judge to find applicable pool errors and verify prior detections. This yields refined error sets $\mathcal{E}_i^{(1)}$.

Error canonicalization applies deterministic normalization: lowercasing, removing non-alphanumeric characters, and synonym replacement. A cyclical refinement pass further stabilizes the vocabulary. Final scores are weighted counts $C_i = \sum_{(e, \text{sev}) \in \mathcal{E}_i} w(\text{sev})$ with $w(\text{major}) > w(\text{minor})$. See Appendix D.4 for details.

3 Experiments

We evaluate error counting as a reward signal using virtual try-on as our primary testbed. Below, we describe the benchmark, baselines, and evaluation protocol. **Mismatch-DressCode (MDressBench):** Existing VTO benchmarks randomly sample source-reference pairs within categories, yielding many pairs with similar

attributes that simplify the transfer task [25]. To expose the failure modes and out-of-distribution artifacts typically seen during model deployment, we curate MDressBench (Figure 3) from DressCode (train set) by selecting pairs with *maximal attribute disagreement*: short-sleeve sources with long-sleeve references, textured short dresses with ankle-length smooth references, and so on. We construct 10K pairs for RL training and reserve 350 pairs with the largest mismatch for evaluation, split across three categories (Top: 79, Lower: 122, Dress: 149) and two reference types (Flat and Non-flat), yielding 700 benchmark samples. For each pair, we follow DiT-VTON [20] to extract pose and inpainting masks for the source image, and align the reference garment crop. See Appendix A for details.

Table 1: Quantitative comparison on MDressBench (**flat** references). All metrics averaged across garment categories. $\uparrow$: higher is better; $\downarrow$: lower is better. Best values in **bold**. IEC outperforms all baselines across all eight metrics.

Method	CEC $\downarrow$	RaR $\uparrow$	Direct $\uparrow$	Garm. $\uparrow$	Attr. $\uparrow$	Real. $\uparrow$	Light. $\uparrow$	Src.Int. $\uparrow$
Baseline	6.018 $\pm$ 1.73	0.760 $\pm$ 0.14	0.796 $\pm$ 0.06	0.793 $\pm$ 0.16	0.815 $\pm$ 0.13	0.735 $\pm$ 0.12	0.760 $\pm$ 0.10	0.914 $\pm$ 0.08
Direct	5.373 $\pm$ 1.57	0.854 $\pm$ 0.11	0.840 $\pm$ 0.04	0.864 $\pm$ 0.09	0.873 $\pm$ 0.08	0.785 $\pm$ 0.08	0.771 $\pm$ 0.09	0.925 $\pm$ 0.07
RaR [12]	5.596 $\pm$ 1.69	0.826 $\pm$ 0.14	0.833 $\pm$ 0.05	0.858 $\pm$ 0.09	0.858 $\pm$ 0.09	0.771 $\pm$ 0.09	0.764 $\pm$ 0.10	0.929 $\pm$ 0.06
IEC (Ours)	5.312$\pm$1.72	0.875$\pm$0.09	0.854$\pm$0.04	0.873$\pm$0.08	0.886$\pm$0.07	0.803$\pm$0.07	0.794$\pm$0.07	0.933$\pm$0.06

Baseline Finetuning Dataset We train the baseline VTO model on a data mixture of VITON-HD [4, 13] and DeepFashion³, covering a range of poses, body shapes, viewpoints, and garment styles.

Implementation Details Our model builds on Kontext-dev [18], a 12B rectified-flow backbone, finetuned for 6.2k steps on the baseline mixture using the DiT-VTON token concatenation scheme [20]. Training uses resolution 512 $\times$ 512, batch size 8, learning rate 5×10^{-5} , AdamW, gradient checkpointing, and bf16 precision on 8 H100 (80GB) GPUs. Complete hyperparameter settings are listed in Appendix B.1.

RL Post-Training. We apply GRPO with $K = 12$ generations per group and best-of- n down-selection ($n = 8$) to increase reward variance. We optimize for 60 steps with batch size 3, gradient accumulation 4, and bf16 precision. Sampling uses 30 steps at 512 $\times$ 512 with guidance scale 2.5, so we capture fine-grained errors that show up in the latter stages of sampling, e.g., color/pattern mismatch; we use 50%



Fig. 3: MDressBench Examples. *Flat* references (*left*) show the garment as a lay-flat product image on a neutral background. *Non-flat* references (*right*) show the garment worn by a model, requiring the model to “extract” the garment from arbitrary poses and backgrounds.

³ <https://www.kaggle.com/code/rkuo2000/deepfashion-tryon>

of steps for optimization biased toward earlier sampling steps. All reward computations use GPT-5-mini. Each reward design uses one judge call per candidate, except RaR (two calls: rubric generation + evaluation). To avoid self-referential bias, we use GPT-5 for evaluations.

Table 2: Quantitative comparison on MDressBench (**non-flat** references). All metrics averaged across garment categories. $\uparrow$: higher is better; $\downarrow$: lower is better. Best values in **bold**. IEC outperforms all baselines across all eight metrics.

Method	CEC $\downarrow$	RaR $\uparrow$	Direct $\uparrow$	Garm. $\uparrow$	Attr. $\uparrow$	Real. $\uparrow$	Light. $\uparrow$	Src.Int. $\uparrow$
Baseline	5.489 $\pm$ 1.48	0.811 $\pm$ 0.13	0.815 $\pm$ 0.05	0.835 $\pm$ 0.12	0.848 $\pm$ 0.11	0.754 $\pm$ 0.11	0.767 $\pm$ 0.10	0.902 $\pm$ 0.10
Direct	5.284 $\pm$ 1.54	0.867 $\pm$ 0.11	0.837 $\pm$ 0.04	0.870 $\pm$ 0.07	0.879 $\pm$ 0.07	0.782 $\pm$ 0.08	0.770 $\pm$ 0.09	0.906 $\pm$ 0.11
RaR [12]	5.533 $\pm$ 1.65	0.850 $\pm$ 0.14	0.829 $\pm$ 0.05	0.862 $\pm$ 0.08	0.863 $\pm$ 0.09	0.769 $\pm$ 0.09	0.770 $\pm$ 0.10	0.905 $\pm$ 0.13
IEC w/o GC	5.259 $\pm$ 1.51	0.890 $\pm$ 0.11	0.852 $\pm$ 0.04	0.878 $\pm$ 0.07	0.887 $\pm$ 0.06	0.799 $\pm$ 0.07	0.799 $\pm$ 0.08	0.918 $\pm$ 0.10
IEC (Ours)	5.203 $\pm$ 1.48	0.883 $\pm$ 0.10	0.852 $\pm$ 0.04	0.882 $\pm$ 0.07	0.893 $\pm$ 0.06	0.803 $\pm$ 0.07	0.796 $\pm$ 0.08	0.910 $\pm$ 0.11

Evaluation Protocol For each prompt, we generate 8 outputs per model and average metrics over two runs for stability, making the effective benchmark size 5600 samples. We evaluate robustly with multiple metrics, including: (i) **CEC** ($\downarrow$) as our primary metric (§2.5); (ii) RaR score ($\uparrow$); (iii) Direct score ($\uparrow$); and (iv) its aspect scores ($\uparrow$) for Garment Transfer, Attribute Preservation, Realism, Lighting, and Source Integrity. Additionally, we evaluate on the standard paired test splits of VITON-HD and DressCode using conventional perceptual metrics (LPIPS, SSIM, FID, KID) to assess generalization beyond MDressBench (Table 6).

Human Evaluation of CEC. To validate that CEC reflects human preferences as a *selection metric*, we conduct a 25-sample study where annotators rank try-on outputs from the baseline model across 8 images generated per sample, then evaluate each group using Direct, RaR, and CEC scoring and report top- k ranking accuracy. This study validates CEC as a metric, not IEC as a training reward; a dedicated human preference study of IEC outputs at scale is left to future work. Prior work on decomposed evaluation supports the principle that counting structured violations correlates well with human quality judgments: GenEval [10] counts compositional errors in text-to-image task, and TIFA [15] counts question-answering failures for text-image alignment.

Cross-Family Judge Evaluation. Because both our training (GPT-5-mini) and evaluation (GPT-5) judges come from the same model family, we additionally evaluate, as an ablation, with a frozen open-source judge, **Qwen3.6-35B-A3B** (with temperature set to 0), using the identical CEC, RaR, and Direct pipelines, see §5.

4 Results

We evaluate on MDressBench (*flat* and *non-flat*). Prior to evaluation, we validate CEC as a selection metric: Table 3 shows CEC matches the human top choice

Table 3: Human alignment of automated selection rules ($n = 25 \times 8$). CEC achieves $2 \times$ higher top-1 accuracy than alternatives, validating it as a reliable selection metric.

Selector	Top-1 $\uparrow$	Top-3 $\uparrow$	Top-5 $\uparrow$
Direct	0.30	0.80	1.0
RaR [12]	0.30	0.90	1.0
CEC (Ours)	0.60	1.0	1.0

twice as often as Direct or RaR (Top-1: 0.60 vs. 0.30), supporting its use as the primary evaluation signal. We organize our results around three questions: Does error counting outperform rubric-based and direct rewards? (§4.1) Why does explicit counting fail while implicit counting succeeds? (§5) What role does group calibration play? (§5)

4.1 Error Counting Outperforms Rubrics and Direct Scoring

Tables 1 and 2 present results on MDressBench. All RL methods improve over the SFT baseline. **IEC vs. RaR.** IEC outperforms RaR on all eight metrics under both flat and non-flat references. On flat references, IEC achieves CEC 5.31 vs. 5.60 for RaR. On the more challenging non-flat subset, IEC exceeds RaR by 5.96% (CEC), 3.88% (RaR score), 2.32% (garment transfer), 3.48% (attribute preservation), 4.33% (realism), 3.3% (lighting), and 0.6% (source integrity). This gap is notable because RaR must generate rubrics *without* an ideal reference to anchor criteria; rubrics become generic and miss prompt-specific failure modes, producing misaligned reward signals. **IEC vs. Direct Scoring.** Direct scoring reduces CEC from 6.02 (SFT) to 5.37 on flat references, however, IEC still outperforms on all metrics. Direct scoring compresses quality into a single scalar, losing the fine-grained error sensitivity that IEC’s aspect decomposition provides. The gap widens on non-flat references, where transfer errors are more varied and harder to capture with a direct score. **RaR underperforms Direct.** Strikingly, RaR scores worse than Direct on CEC (5.60 vs. 5.37 on flat; 5.53 vs. 5.28 on non-flat). This confirms that in the reference-free setting, attempting to construct rubrics without ideal answers can be worse than not using rubrics at all. Binary pass/fail judgments across many criteria compound stochastic judge errors, and generic criteria miss the specific failure modes of each input.

4.2 Garment Category Analysis

Figure 1 and the supplement (Fig. 1, Tabs. 2 and 3) show that IEC improves performance *across all garment types* rather than shifting errors between categories. The largest gain is on lowers (CEC 4.66 vs. 5.65 for SFT), while tops are the only case where Direct is slightly better on CEC. IEC attains the best score in 20 of 24 (metric, category) combinations on flat references, indicating gains are not driven by a single category. On non-flat references, IEC wins 14/24 combinations, with gains concentrating on dresses (all 8 metrics) and tops (5/8). Figure 2 of

Table 4: Explicit vs. implicit error counting. EEC fails to improve over baseline; IEC achieves consistent gains.

Method	Flat			Non-Flat		
	CEC↓	Rub.↑	Dir.↑	CEC↓	Rub.↑	Dir.↑
Baseline (SFT)	6.02	0.76	0.80	5.49	0.81	0.82
EEC	6.03	0.76	0.81	5.72	0.77	0.81
IEC (Ours)	5.31	0.88	0.85	5.20	0.88	0.85

the supplement (Appendix B.3) summarizes this per-category structure: IEC achieves the largest area across all categories, with the most pronounced gains on Dresses. See Appendix B.5 for a qualitative discussion of reward hacking modes across reward designs.

4.3 Generalization to Standard Benchmarks

To assess whether IEC’s gains transfer beyond MDressBench, we evaluate on the paired test splits of VITON-HD and DressCode using perceptual metrics. Table 6 compares our RL post-trained models against six external SFT baselines [5–7, 44, 46, 47]. IEC achieves the best LPIPS (0.047) and SSIM (0.892) on VITON-HD, and the best FID (3.610) and ties for best SSIM (0.927) on DressCode. Across both benchmarks, IEC attains the top score on 6 of 8 metrics despite using only 60 RL steps with no additional paired data. The consistent gains over external SFT models, trained on substantially larger paired datasets, validate that error-counting rewards provide an efficient path to quality improvement.

5 Ablations

Why Not Explicit Error Counting? A natural question is whether exposing error lists (EEC) would provide even finer-grained signal than implicit scoring. Table 4 and Figure 2 answer this decisively: EEC *fails to improve* CEC on flat references (6.03 vs. 6.02 SFT) and *degrades* on non-flat references (5.72 vs. 5.49 SFT). IEC yields large improvements in both settings. This behavior is consistent with the reward dynamics in Fig. 2. EEC produces high-variance rewards and degrades early. Explicit label enumeration is brittle, as semantically identical errors can be counted under multiple surface forms, injecting noise into within-group comparisons and destabilizing rankings. IEC, by contrast, asks the judge to implicitly aggregate errors into a calibrated scalar, yielding more stable preference signals. More broadly, in reference-free settings, error-based rewards from multimodal VLM judges are inherently susceptible to variance from linguistic and perceptual ambiguity (see Appendix B.6). Implicit scoring is therefore not merely a convenience; it is a stability-critical design choice for optimization.

Group Calibration Table 2 includes IEC w/o GC (group calibration removed). Without calibration, IEC still outperforms all baselines, confirming that the core

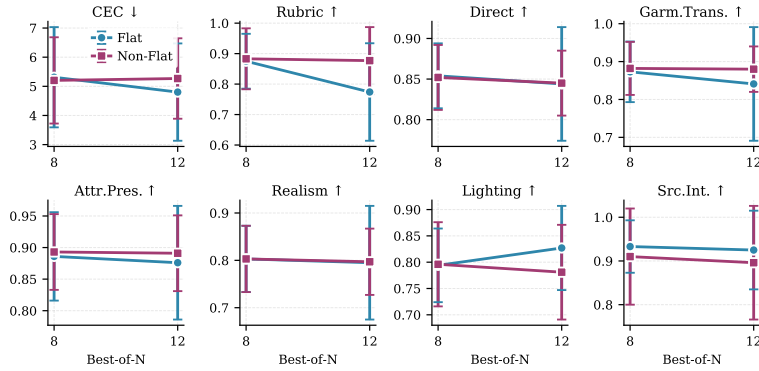


Fig. 4: Effect of Best-of- N with 12 generations. Higher N typically reduces reward diversity and performance.

Table 5: Judge-family ablation on MDressBench (non-flat). Evaluation using a different judge shows that IEC dominates Direct/RaR across both judges, while absolute rubric values shift under the stricter Qwen judge, the relative ordering holds.

Method	GPT-5			Qwen3.6-35B-A3B		
	CEC↓	Rub↑	Dir↑	CEC↓	Rub↑	Dir↑
Baseline	5.489	0.811	0.815	7.114	0.087	0.879
Direct	5.284	0.867	0.837	5.550	0.107	0.919
RaR [12]	5.533	0.850	0.829	5.905	0.106	0.907
IEC (Ours)	5.203	0.883	0.852	5.377	0.110	0.919

error-counting signal is strong. However, group calibration improves CEC (5.203 vs. 5.259 on non-flat) by reducing prompt-to-prompt scale variability, which stabilizes GRPO’s advantage estimates. The effect is more pronounced on harder prompts where reward magnitudes vary most. Note that calibration is effective for IEC precisely because the within-group ranking is reliable; for EEC, where stochastic enumeration corrupts the ranking itself (Appendix B.6), calibration does not provide equivalent benefit.

Robustness to Underlying Judge Model Because our training and evaluation judges share a model family, we ablate judge identity by evaluating the non-flat subset of the benchmark with a different and open-weight judge, as described in §3. Table 5 shows the relative performance across methods and judges is preserved: IEC outperforms Baseline, Direct, and RaR under *both* judges on all three metrics, and the $\text{IEC} > \text{Direct} > \text{RaR}$ ordering holds cross-family. This indicates IEC’s gains are not an artifact of self-preference within a single judge family. Full per-category results are in Appendix B.8.

Effect of Best-of- N Sampling Best-of- N selects the best $N/2$ and worst $N/2$ within a group to increase diversity. We vary $N \in \{8, 12\}$ with 12 generations fixed (Figure 4). Increasing N generally reduces aspect scores, consistent with a quality-diversity trade-off. CEC responds asymmetrically: it slightly worsens on

non-flat references (5.20→5.27) but improves on flat (5.31→4.80). We use $N=8$ as it increases diversity within groups and is robust across both reference types. **Qualitative Analysis** In the supplement (Fig. 9), RaR-trained models blur garment details, reflected in poor realism scores. IEC preserves fine-grained textures and patterns. Figure 10 of the supplement shows IEC can occasionally fail on source integrity when source and reference backgrounds differ, accounting for the smaller 0.6% improvement. This small gain is structural rather than a signal failure, because source integrity has the smallest within-group reward variance of any axis at every training window ($\sigma_{\text{src}}=0.060\text{--}0.074$ vs. $\sigma_{\text{garm}}=0.102\text{--}0.170$), so GRPO’s group-normalized advantage on this axis is correspondingly weaker. The residual failures are predominantly pose/limb edits (38% of labels) rather than identity/face changes (30%). See Appendix B.5 for details.

Training Dynamics Figure 2 shows IEC yields smooth, monotonic reward improvement, while Direct and RaR are noisier and EEC regresses early. Figure 5 compares IEC and RaR across iterations (20, 40, 60): IEC consistently achieves lower CEC, reaching minimum of 4.80 at iteration 40 (20% below baseline) vs. 5.06 for RaR. Both methods regress slightly after iteration 40, consistent with reward overfitting observed in prior RL post-training work [22,36]. At iteration-40, with IEC a typical generation carries a tight error distribution (median 5, p90 6, max 10) with *no* zero-error candidates, whereas EEC reports zero errors on 24.5% of candidates while reaching 26 on others, direct evidence of the enumeration variance behind EEC’s instability. We report at iteration-60 as RaR starts to regress; over the 0-60 window IEC’s calibrated reward stays above its starting value ($\Delta_{60}=+0.018$), whereas RaR peaks around iteration-40 and is already declining by iteration-60, with its per-rubric-item fail rate climbing from 3% to 15% respectively. See Appendix B.7 for the full per-axis analysis and training dynamics beyond iteration 60.

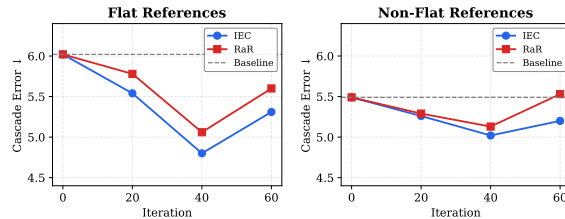


Fig. 5: Training dynamics. CEC ↓ for IEC and RaR across post-training iterations. IEC consistently achieves lower error.

6 Related Work

RL for Visual Generation. RLHF is standard for aligning language models [26]; DPO [28] removes the explicit reward model, while GRPO [31] normalizes

Table 6: We evaluate SFT-only models trained on task-specific paired datasets vs. our RL post-training methods. Best in **bold**, second best underlined. IEC matches or surpasses all external baselines on 6/8 metrics despite using only 60 RL steps with no additional paired data.

Method	VITON-HD				DressCode			
	LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	SSIM↑	FID↓	KID↓
<i>SFT Methods</i>								
CatVTON [6]	0.057	0.870	5.430	0.410	0.046	0.892	3.990	0.820
CatVT2ON [6]	0.057	<u>0.890</u>	8.100	2.250	0.037	<u>0.922</u>	5.720	2.340
CAT-DM [47]	0.080	<u>0.877</u>	5.600	0.830	—	—	—	—
IDM-VTON [5]	0.102	0.870	6.290	—	0.062	<u>0.920</u>	8.640	0.900
OOTDiffusion [44]	0.071	0.878	8.810	0.820	0.045	0.927	4.200	0.370
OmniVTON [46]	0.145	0.832	7.760	—	0.119	0.865	5.340	—
<i>Ours (SFT + RL Post-Training)</i>								
Baseline (SFT)	0.059	0.872	5.490	0.820	0.038	0.893	4.130	0.990
Direct	<u>0.049</u>	0.888	4.980	<u>0.650</u>	0.029	0.921	<u>3.680</u>	0.840
RaR [12]	0.055	0.878	5.210	<u>0.690</u>	<u>0.031</u>	0.915	<u>3.920</u>	0.940
IEC w/o GC	0.052	0.875	5.190	0.660	0.034	0.913	3.980	0.930
IEC (Ours)	0.047	0.892	<u>5.120</u>	0.660	0.035	0.927	3.610	<u>0.790</u>

advantages within candidate groups. Several works adapt policy optimization to diffusion models [1, 9, 33]. Flow-GRPO [22] introduces ODE-to-SDE conversion for flow-matching models; MixGRPO [19] uses mixed ODE-SDE sampling, SuperFlow [3] enables on-the-fly training, and TreeGRPO [8] constructs search trees for step-level credit assignment. DiffusionNFT [50] proposes a negative-aware finetuning objective. GRPO-Guard [36] addresses over-optimization via regulated clipping, directly relevant to the reward regression we observe (Fig. 5). Our work focuses on reward design rather than optimization mechanics.

Reward Design for Subjective Tasks. ImageReward and HPSv2 [41, 43] train reward models on human preference datasets but are aesthetics-centric. Using VLMs as evaluators [2, 11, 37, 40, 49] has been applied to both evaluation and training-time rewards. RaR [12] generates task-specific rubrics, improving interpretability; Rubicon [16] extends this with defense rubrics, exploring explicit vs. implicit aggregation; and DR Tulu [30] evolves rubrics over training to counter generic, trivially-satisfied criteria, the same dilution we observe for reference-free rubrics. QA-LIGN [17] decomposes a constitution into hierarchical principle-specific QA checks scored by an LLM judge and optimized with GRPO for safety alignment using a structured per-axis reward complementary to our error-counting method. UniWorld-V2/Edit-R1 [32] uses VLM implicit feedback for image editing post-training via logit-based scoring, which is structurally parallel to our implicit approach, though applied to general editing. CoCoEdit [42] shows exclusive reliance on VLM rewards degrades content consistency, motivating region-based regularization—paralleling our observation that source integrity sees the smallest IEC improvement. GenEval [10] and TIFA [15] demonstrate that decomposing quality into countable criteria (compositional errors and QA failures, respectively) yields more stable signals than holistic scores. IEC extends this principle from evaluation to training rewards in a reference-free setting.

Virtual Try-On and Evaluation. VTO synthesizes images of a person wearing a target garment while preserving identity and background. Early methods used geometric warping [13, 35]; VITON-HD [4] introduced high-resolution try-on, DressCode [25] extended to multiple categories, LADI-VTON [24] introduced latent diffusion with textual inversion, and TryOnDiffusion [51] proposed parallel U-Nets. Recent methods explore DiT [27] and flow-matching architectures, but rely on supervised fine-tuning on narrow datasets. VTON-VLLM [34] is the closest prior work: it aligns VTO via a VLLM fine-tuned on human-validated feedback, whereas we perform on-policy GRPO post-training using error-counting rewards without human preference labels or a trained reward model. Traditional VTO metrics (FID [14], LPIPS [48], SSIM [38]) correlate poorly with human perception of VTO-specific quality [25] and cannot detect semantic errors like incorrect sleeve length. Our CEC metric addresses this via a shared error vocabulary for consistent, fine-grained measurement.

7 Limitations and Broader Considerations

Domain breadth. We validate IEC only on VTO; testing on other reference-free domains (image editing, robotic manipulation, creative writing) would establish generality. **Missing baselines.** We do not compare against VTON-VLLM [34], which requires a fine-tuned VLLM reward model, whereas our method is training-free. **Metric validity.** CEC improves human alignment in our 25-sample study but remains an automated proxy; larger studies evaluating IEC outputs directly are future work. **Reproducibility.** IEC relies on non-deterministic LLMs as judges (GPT-5-mini/GPT-5); API changes could silently alter reward signals, motivating future work with open-source judges. **Fairness.** Our benchmark is curated from DressCode, which does not fully cover the diversity of body shapes, skin tones, and garment styles, and reward designs optimizing for judge-perceived quality may inherit judge biases. See Appendix B.4 for expanded discussion.

8 Conclusion

We identified the reference-free setting as a gap in RL post-training: domains where ideal answers are unavailable, multiple outputs are valid, and quality is defined by the absence of errors. We proposed Implicit Error Counting (IEC), showing that implicit score emission and group calibration are necessary for stable optimization. On VTO, IEC consistently outperforms rubric-based and direct-scoring rewards and matches or exceeds supervised baselines trained on far more paired data, at half the judge compute of RaR. When you cannot define what an ideal output looks like, define what a bad one looks like, and count.

Acknowledgements

This work was conducted during an internship at Amazon. The authors thank the Amazon team for their support and resources.

References

1. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
2. Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., Sun, L.: Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In: Forty-first International Conference on Machine Learning (2024)
3. Chen, K., Xu, Z., Shen, Y., Lin, Z., Yao, Y., Huang, L.: Superflow: Training flow matching models with rl on the fly. arXiv preprint arXiv:2512.17951 (2025)
4. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14131–14140 (2021)
5. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for authentic virtual try-on in the wild. In: European Conference on Computer Vision. pp. 206–235. Springer (2024)
6. Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., Jiang, D., Liang, X.: Catvton: Concatenation is all you need for virtual try-on with diffusion models. arXiv preprint arXiv:2407.15886 (2024)
7. Chong, Z., Zhang, W., Zhang, S., Zheng, J., Dong, X., Li, H., Wu, Y., Jiang, D., Liang, X.: Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. arXiv preprint arXiv:2501.11325 (2025)
8. Ding, Z., Ye, W.: Treegrp: Tree-advantage grp for online rl post-training of diffusion models. arXiv preprint arXiv:2512.08153 (2025)
9. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Reinforcement learning for fine-tuning text-to-image diffusion models. In: Advances in Neural Information Processing Systems (2024)
10. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
11. Gong, Y., Wang, X., Wu, J., Wang, S., Wang, Y., Wu, X.: Onereward: Unified mask-guided image generation via multi-task human preference learning. arXiv preprint arXiv:2508.21066 (2025)
12. Gunjal, A., Wang, A., Lau, E., Nath, V., He, Y., Liu, B., Hendryx, S.: Rubrics as rewards: Reinforcement learning beyond verifiable domains. arXiv preprint arXiv:2507.17746 (2025)
13. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7543–7552 (2018)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
15. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20406–20417 (2023)
16. Huang, Z., Zhuang, Y., Lu, G., Qin, Z., Xu, H., Zhao, T., Peng, R., Hu, J., Shen, Z., Hu, X., et al.: Reinforcement learning with rubric anchors. arXiv preprint arXiv:2508.12790 (2025)

17. Jacob Dineen, A.R., Liu, Q., Xu, Z., Ye, X., Shen, M., Li, Z., Lu, S., Baral, C., Chen, M., Zhou, B.: Qa-align: Aligning llms through constitutionally decomposed qa. arXiv preprint arXiv:2506.08123 (2025)
18. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontekst: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
19. Li, J., Cui, Y., Huang, T., Ma, Y., Fan, C., Yang, M., Zhong, Z.: Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. arXiv preprint arXiv:2507.21802 (2025)
20. Li, Q., Qiu, S., Han, J., Xu, X., Seyfioglu, M.S., Koo, K.K., Bouyarmane, K.: Dit-vton: Diffusion transformer framework for unified multi-category virtual try-on and virtual try-all with integrated image editing. arXiv preprint arXiv:2510.04797 (2025)
21. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
22. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470 (2025)
23. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. The International Conference on Learning Representations (2022)
24. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Ladt-vton: Latent diffusion textual-inversion enhanced virtual try-on. In: Proceedings of the ACM International Conference on Multimedia. pp. 8580–8589 (2023)
25. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress code: High-resolution multi-category virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2231–2235 (2022)
26. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35**, 27730–27744 (2022)
27. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
28. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* **36**, 53728–53741 (2023)
29. Seyfioglu, M.S., Bouyarmane, K., Kumar, S., Tavanaei, A., Tutar, I.B.: Diffuse to choose: Enriching image conditioned inpainting in latent diffusion models for virtual try-all. arXiv preprint arXiv:2401.13795 (2024)
30. Shao, R., Asai, A., Shen, S.Z., Ivison, H., Kishore, V., Zhuo, J., Zhao, X., Park, M., Finlayson, S.G., Sontag, D., Murray, T., Min, S., Dasigi, P., Soldaini, L., Brahman, F., tau Yih, W., Wu, T., Zettlemoyer, L., Kim, Y., Hajishirzi, H., Koh, P.W.: Dr tulur: Reinforcement learning with evolving rubrics for deep research (2026), <https://arxiv.org/abs/2511.19399>, accessed: 2026-06-30
31. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)

32. UniWorld Team: UN I W O R L D-V2: Reinforce Image Editing with Diffusion Negative-Aware Finetuning and MLLM Implicit Feedback. Tech. rep., Peking University, Rabbitpre AI (2025), technical Report, arXiv:2510.16888
33. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purber, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. arXiv preprint arXiv:2311.12908 (2024)
34. Wan, S., Chen, J., Cai, Q., Pan, Y., Yao, T., Mei, T.: VTON-VLLM: Aligning virtual try-on models with human preferences. In: NeurIPS (2025)
35. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European Conference on Computer Vision. pp. 589–604 (2018)
36. Wang, J., Liang, J., Liu, J., Liu, H., Liu, G., Zheng, J., Pang, W., Ma, A., Xie, Z., Wang, X., Wang, M., Wan, P., Liang, X.: Grpo-guard: Mitigating implicit over-optimization in flow matching via regulated clipping. arXiv preprint arXiv:2510.22319 (2025)
37. Wang, Y., Zang, Y., Li, H., Jin, C., Wang, J.: Unified reward model for multimodal understanding and generation. arXiv preprint arXiv:2503.05236 (2025)
38. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004)
39. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
40. Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., et al.: Q-bench: A benchmark for general-purpose foundation models on low-level vision. In: ICLR (2024)
41. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
42. Wu, Y., Xie, C., Li, R., Chen, L., Yi, Q., Zhang, L.: Cooedit: Content-consistent image editing via region regularized reinforcement learning. arXiv preprint arXiv:2602.14068 (2026)
43. Xu, J., Liu, X., Wu, Y., Wang, Y., Ye, W., Geng, S., Zhao, Y., Li, J., Li, C., Sun, H., et al.: Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems* **36** (2023)
44. Xu, Y., Gu, T., Chen, W., Chen, C.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on (2024), <https://arxiv.org/abs/2403.01779>, accessed: 2026-06-30
45. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
46. Yang, Z., Li, Y., He, S., Li, X., Xu, Y., Dong, J., Du, Y.: Omnivton: Training-free universal virtual try-on. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16702–16711 (2025)
47. Zeng, J., Song, D., Nie, W., Tian, H., Wang, T., Liu, A.A.: Cat-dm: Controllable accelerated virtual try-on with diffusion model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8372–8382 (2024)
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)

- 49. Zhang, Z., Wu, H., Zhang, E., Zhai, G., Lin, W.: Q-bench: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)
- 50. Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., Liu, M.Y.: Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117* (2025)
- 51. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: A tale of two unets. *arXiv preprint arXiv:2306.08276* (2023)