Understanding Reward Hacking in Text-to-Image Reinforcement Learning

Supplementary Material

6. Automatic Prompt Optimization for ArtifactReward

We employ automatic prompt optimization (APO) [28] to optimize the prompt used for ArtifactReward. Qwen2.5-VL-7B-Instruct [2] serves as our backbone vision-language model, and APO is used to optimize the prompt so that the model can more reliably distinguish artifact-containing images from artifact-free ones.

Our initial prompt is given as:

Initial prompt.

```
Is there any artifacts in the image that
look not realistic?
```

We compute the ArtifactReward using the normalized probability of the model answering "no", where a higher "no" probability indicates a higher likelihood of being artifact-free (See Eq. 1). During optimization, we maximize the following log-likelihood objective:

$$LL = y\log(p_{\text{yes}}) + (1-y)\log(1-p_{\text{yes}}), \qquad (2)$$

where y is the ground-truth label, and $p_{\text{yes}} = 1 - p_{\text{no}}$ denotes the normalized probability assigned to the "yes" token. Algorithm 1 provides the complete APO procedure, with the scoring function $S(\cdot)$ defined as the log-likelihood above.

An example of the optimized prompt is:

Optimized prompt.

```
Analyze the images to identify any
unintentional digital artifacts,
concentrating on irregular lighting, object
placement, blending errors, or anomalies
that could affect realism. Disregard any
deliberate artistic styles or intentional
surreal elements. Respond with YES if
artifacts are detected; if not, respond with
NO.
```

This optimized prompt is more explicit and structurally aligned with common artifact patterns, enabling the model to produce more accurate and reliable artifact predictions. As shown in Table 3, our ArtifactReward consistently assigns higher scores to artifact-free images, clearly separating clean generations from those with distortions. This demonstrates the effectiveness of our optimized prompt in enabling a reward model that can reliably detect and penalize generated artifacts.

Algorithm 1 Automatic Prompt Optimization for ArtifactReward

Require: p_0: initial prompt, N: iterations, $\mathcal{D} = (x, y)$: training dataset where x is the input image and y is the artifact label, b: top prompts retained per iteration, l: number of error examples for reflections, $S(\cdot)$: scoring function

$P_0 \leftarrow \{p_0\}$ ▷ Initialize candidate prompt set
for $t = 1$ to N **do**
  $P_c \leftarrow P_{t-1}$
  for $p \in P_{t-1}$ **do**
    $J_{\text{error}} = \{(x_i, y_i) \mid p^i_{\text{yes}} > 0.5 \text{ if } y_i = 0, \text{ or } p^i_{\text{yes}} < 0.5 \text{ if } y_i = 1\}$
    $J^i_{\text{error}} \subset J_{\text{error}}$ is a sampled subset for each $i = 1, \ldots, l$
    $G = \bigcup_{i=1,\ldots l} \text{Reflect}(p, J^i_{\text{error}})$
    $H = \bigcup_{i=1,\ldots l} \text{Modify}(p, g_i, J^i_{\text{error}})$
    $P_c \leftarrow P_c \cup H$
  end for
  $\mathcal{S}_c = \{S(p) \mid p \in P_c\}$ ▷ Evaluate prompts
  $P_t \leftarrow \{p \in P_c \mid S(p) \geq \tau\}$, where τ is the b^{th} highest score in $\mathcal{S}_c$
end for
Return $p^* \leftarrow \arg\max_{p \in P_N} S(p)$

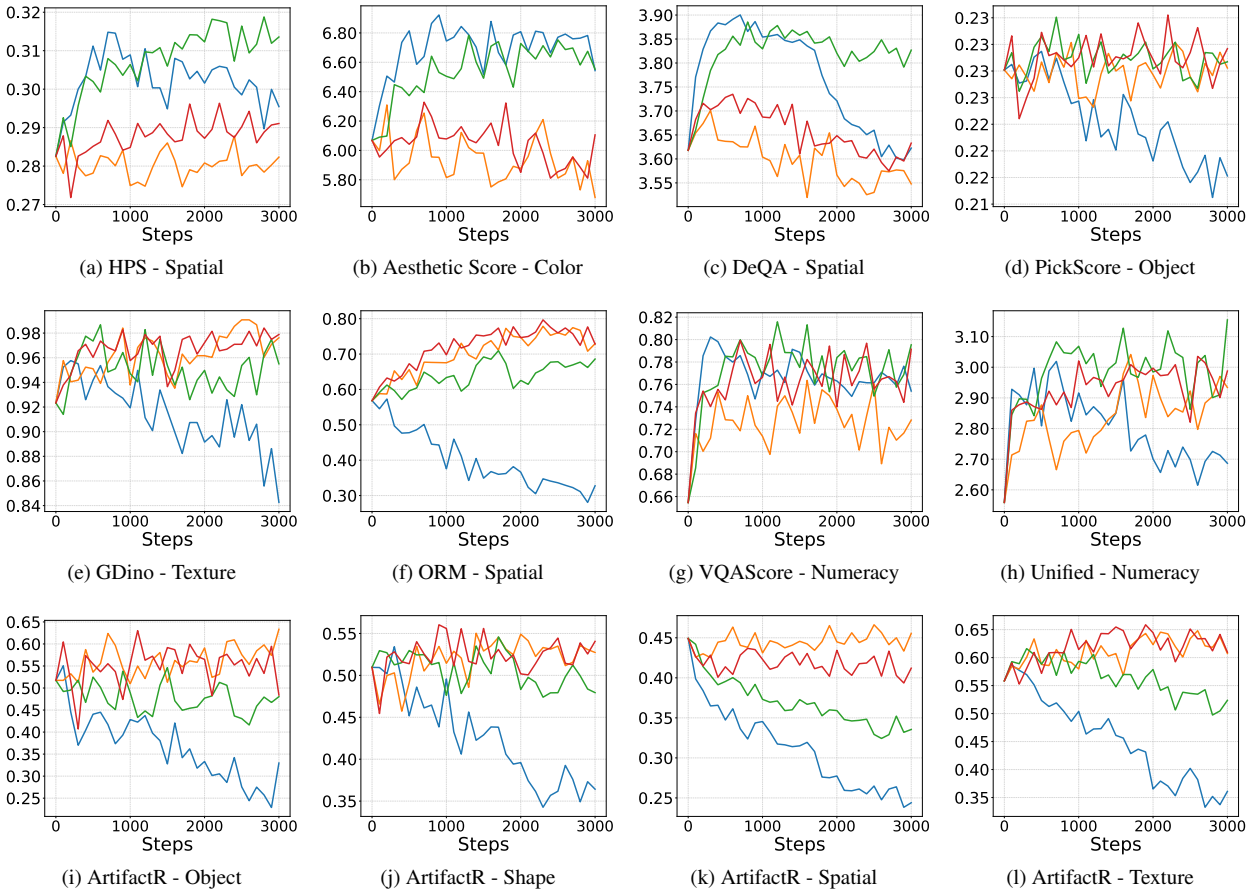
7. Additional Experiment Results

7.1. Additional Training Dynamics on Janus-Pro-7B

We conducted additional experiments using the larger 7B variant of Janus-Pro [4]. The training dynamics of different evaluation metrics are shown in Figure 4. Overall, we observe trends similar to those seen with the 1B model. Optimization under a given reward tends to improve metrics with closely related inductive biases, while degrading performance on metrics that capture different aspects of image quality. For example, the model trained with the GDino reward shows increased GDino and ORM scores, as both scores emphasize object recognition in generated images. In contrast, the model fails to improve HPS, Aesthetic Score, and DeQA, which focus more on color quality and visual appeal. Moreover, across all reward settings, we find that none effectively reduce artifacts in the generated images (Figures 4i–4l).

7.2. Detailed Results of WISE benchmark

Figure 5 presents the full WISE [26] benchmark results across different prompt subcategories and evaluation metrics for Janus-Pro-1B [4] trained with various reward configurations. Across nearly all categories,



(a) HPS - Spatial (b) Aesthetic Score - Color (c) DeQA - Spatial (d) PickScore - Object

(e) GDino - Texture (f) ORM - Spatial (g) VQAScore - Numeracy (h) Unified - Numeracy

(i) ArtifactR - Object (j) ArtifactR - Shape (k) ArtifactR - Spatial (l) ArtifactR - Texture

Figure 4. Evolution of metrics over training steps with different categories of prompts trained on Janus-Pro-7B [4]. Blue color denotes the model trained with HPS [42]; Orange color denotes the model trained with GDino [25]; Green color denotes the model trained with HPS and GDino; Red color denotes the model trained with finetuned ORM [14].

Table 6. Performance on WISE [26] benchmark trained on Janus-Pro-7B [4]. WiScore = 0.7 × Consistency + 0.2 × Realism + 0.1 × Aesthetic by the default setting.

Method	Consistency	Realism	Aesthetic	**WiScore**
Janus-Pro-7B [4]	0.4340	0.5470	0.5528	0.4685
HPS [42]	0.4160	0.4745	0.6530	0.4514
HPS + ArtifactR	0.4255	**0.6605**	0.5815	0.4881
GDino [25]	0.4445	0.5170	0.5435	0.4689
GDino + ArtifactR	0.4470	0.6295	0.5465	0.4935
ORM [14]	0.4420	0.5800	0.5975	0.4852
ORM + ArtifactR	0.4665	0.6355	0.5840	0.5121
HPS + GDino	0.4780	0.5697	**0.6755**	0.5161
HPS + GDino + ArtifactR	0.4775	0.6470	0.5880	0.5225
T2I-R1 [18]	0.4655	0.5925	0.6730	0.5157
T2I-R1 + ArtifactR	**0.4945**	0.6450	0.6000	**0.5352**

incorporating our ArtifactReward leads to substantial improvements in image realism, consistently outperforming models trained with baseline rewards. In addition, ArtifactReward enhances image consistency in most subcategories, indicating that reducing structural artifacts not only improves visual plausibility but also strengthens text-image alignment across diverse prompt types. More image illustrations can be found in Figure 7 and Figure 8.

We additionally evaluate our method on the larger Janus-Pro-7B [4] model. As shown in Table 6 and Figure 6, the results mirror those observed with the 1B model, demonstrating that ArtifactReward generalizes effectively across model scales.

7.3. Detailed Results of LLM4LLM benchmark

Figure 9 presents the LLM4LLM [37] benchmark results across different prompt subcategories and evaluation metrics for Janus-Pro-1B [4] trained with various reward configurations. More image illustrations can be found in Figure 11 and Figure 12.

Additional results trained on Janus-Pro-7B [4] are shown in Table 7 and Figure 10.

Table 7. Performance on LLM4LLM [37] benchmark trained on Janus-Pro-7B [4].

Method	Perception	Correspondence	All
Janus-Pro-7B [4]	0.4390	0.5416	0.9806
HPS [42]	0.4360	0.5205	0.9564
HPS + ArtifactR	0.4454	0.5278	0.9732
GDino [25]	0.4417	0.5570	0.9987
GDino + ArtifactR	0.4506	0.5519	1.0024
ORM [14]	0.4499	0.5561	1.0059
ORM + ArtifactR	0.4576	0.5613	1.0188
HPS + GDino	**0.4615**	0.5540	1.0156
HPS + GDino + ArtifactR	0.4558	0.5594	1.0152
T2I-R1 [18]	0.4580	0.5559	1.0139
T2I-R1 + ArtifactR	0.4608	**0.5622**	**1.0230**

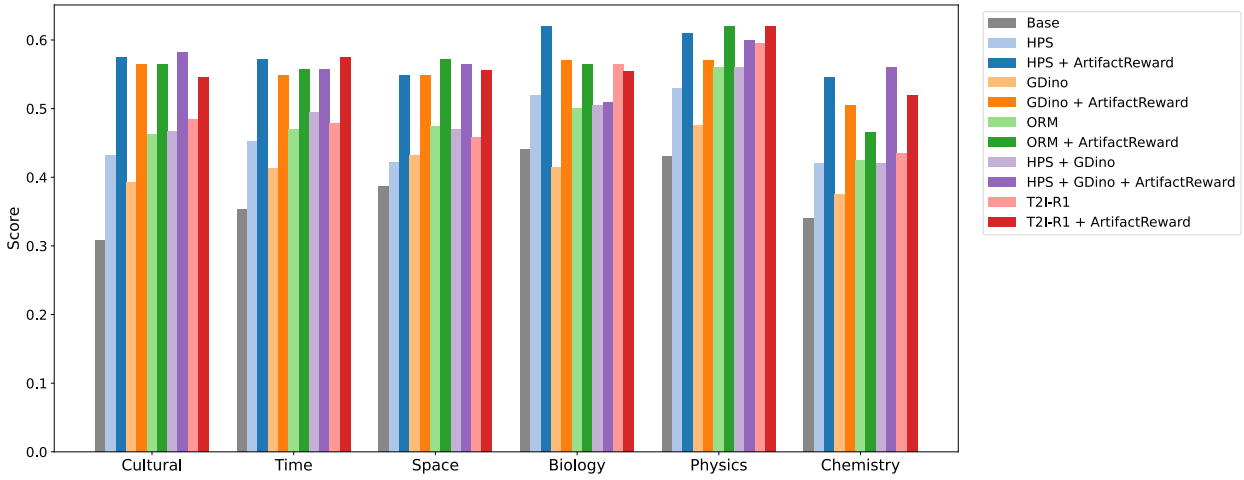
Table 8. Performance on EvalAlign [35] benchmark trained on Janus-Pro [4].

Method	**Faithfulness**	
	1B	7B
Janus-Pro [4]	0.7642	0.8694
HPS [42]	0.7569	0.8538
HPS + ArtifactR	**0.9302**	**1.0676**
GDino [25]	0.7832	0.8956
GDino + ArtifactR	**0.8952**	**0.9837**
ORM [14]	0.7560	0.8607
ORM + ArtifactR	**0.9140**	**0.9245**
HPS + GDino	0.7840	0.8956
HPS + GDino + ArtifactR	**0.8756**	**0.9837**
T2I-R1 [18]	0.7692	0.9059
T2I-R1 + ArtifactR	**0.9015**	**0.9638**

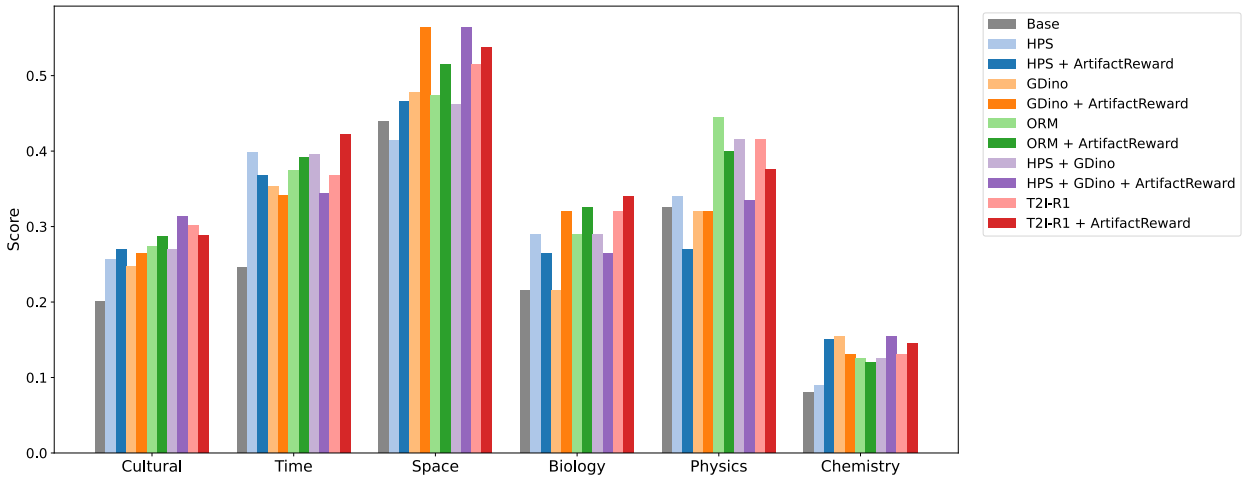
7.4. Detailed Results of EVALALIGN benchmark

We additionally evaluate our method on **EVALALIGN** [35], a fine-grained, MLLM-based evaluation benchmark for text-to-image models. It uses image faithfulness, how accurately the visual content matches reality or semantics, as one of their image quality measurements. To highlight the impact of our method on reducing such artifacts, we focus on this faithfulness metric.

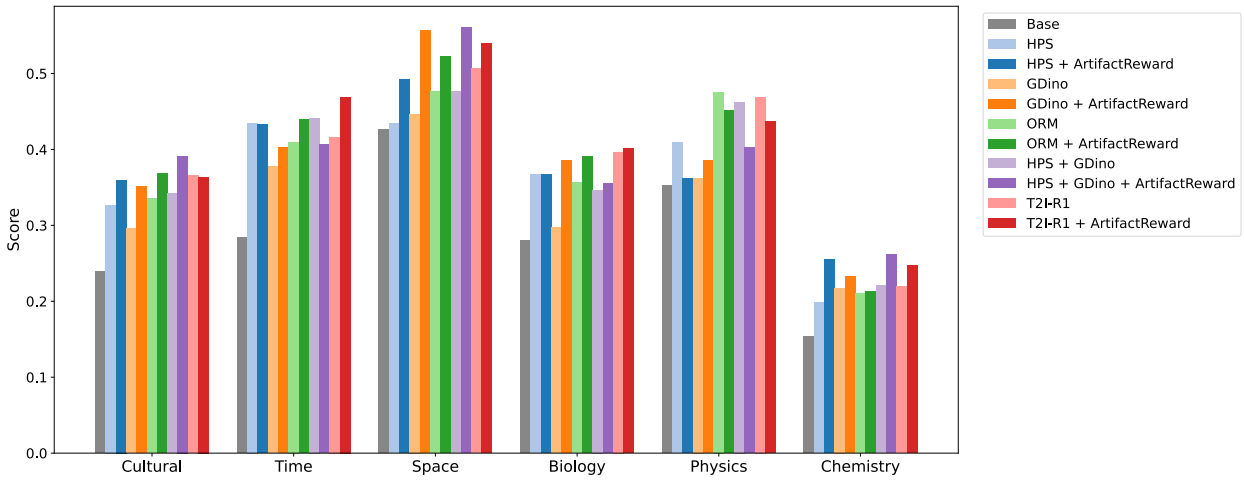
As shown in Table 8, Figure 13 and 14, incorporating our ArtifactReward consistently leads to substantial improvements, achieving higher scores across nearly all subcategories. These results further confirm that ArtifactReward can enhance semantic plausibility and structural correctness beyond what existing reward models capture.



(a) Realism

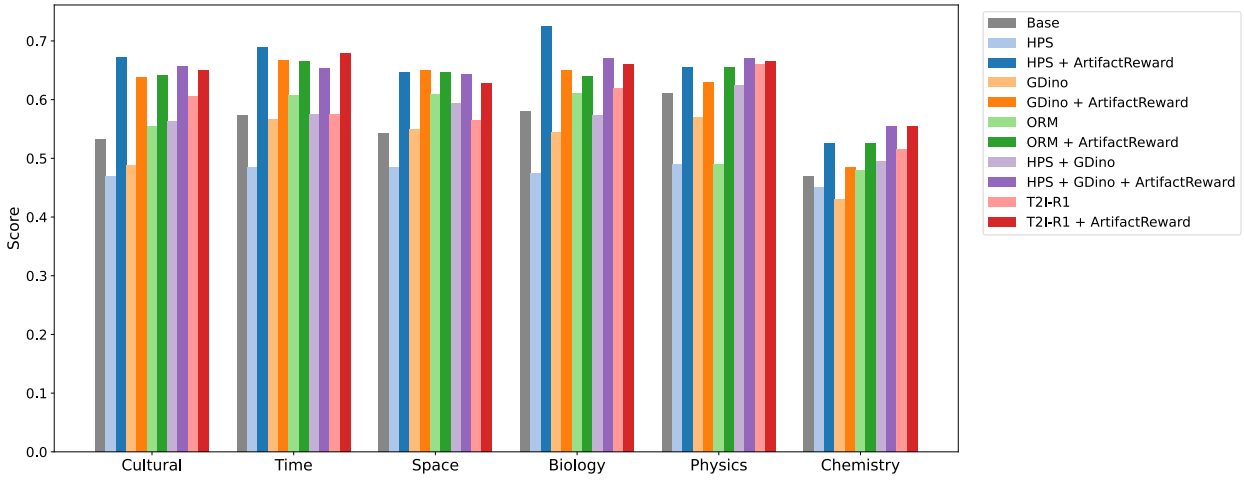


(b) Consistency

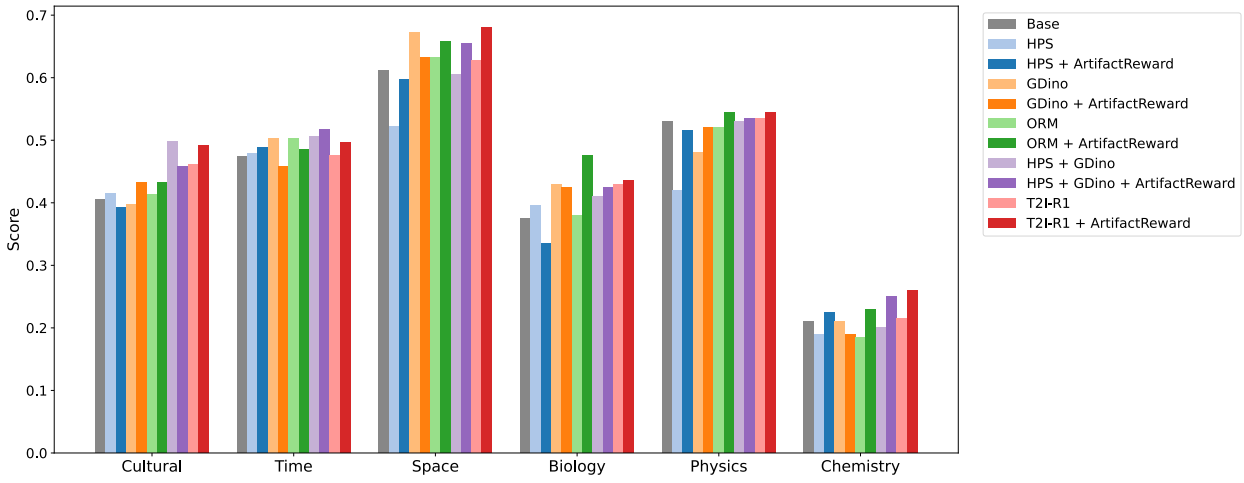


(c) WiScore

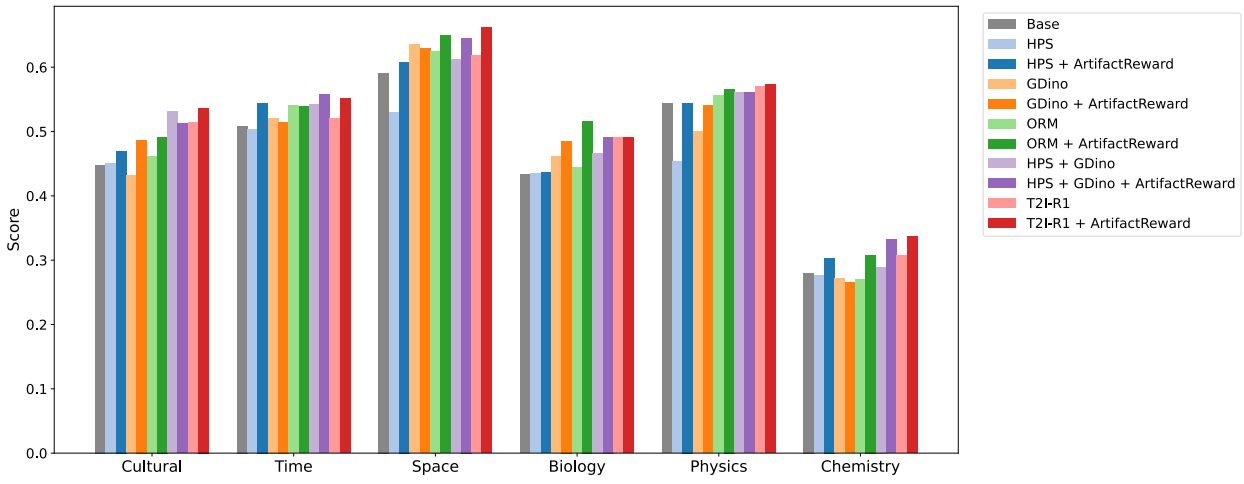
Figure 5. Performance on WISE [26] benchmark across different categories trained on Janus-Pro-1B [4].



(a) Realism



(b) Consistency



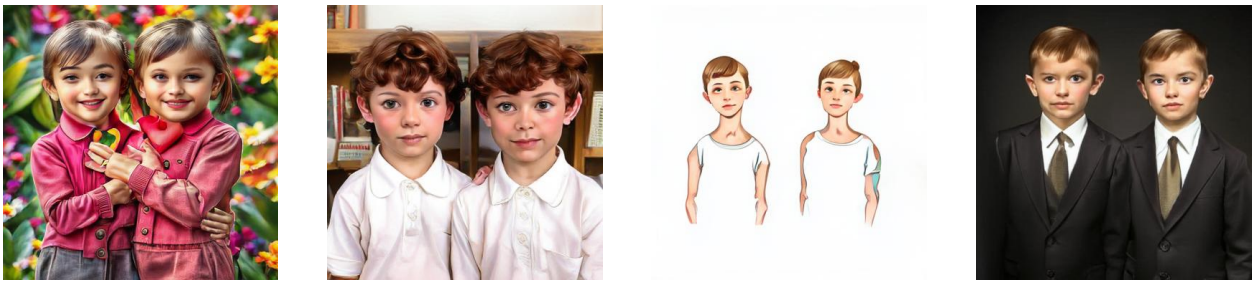
(c) WiScore

Figure 6. Performance on WISE [26] benchmark across different categories trained on Janus-Pro-7B [4]

(a) HPS (b) HPS + ArtifactR (c) GDino (d) GDino + ArtifactR

(e) HPS + GDino (f) HPS + GDino + ArtifactR (g) ORM (h) ORM + ArtifactR

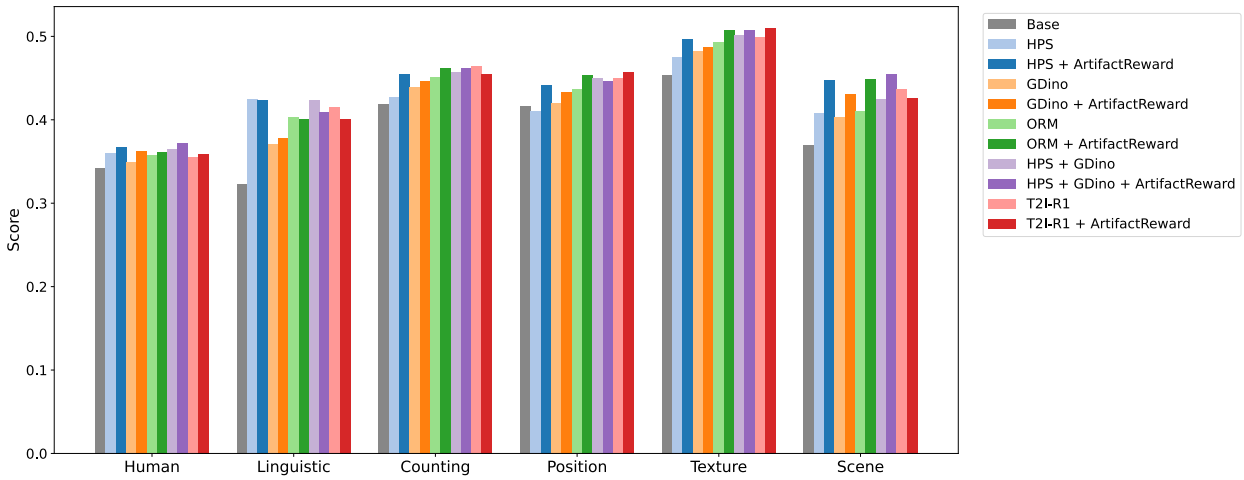
Figure 7. Images generated with prompt "*An iconic bridge, known for its red hue and location over a famous bay in San Francisco*" in WISE [26] benchmark under different training reward configurations trained on Janus-Pro-1B [4].



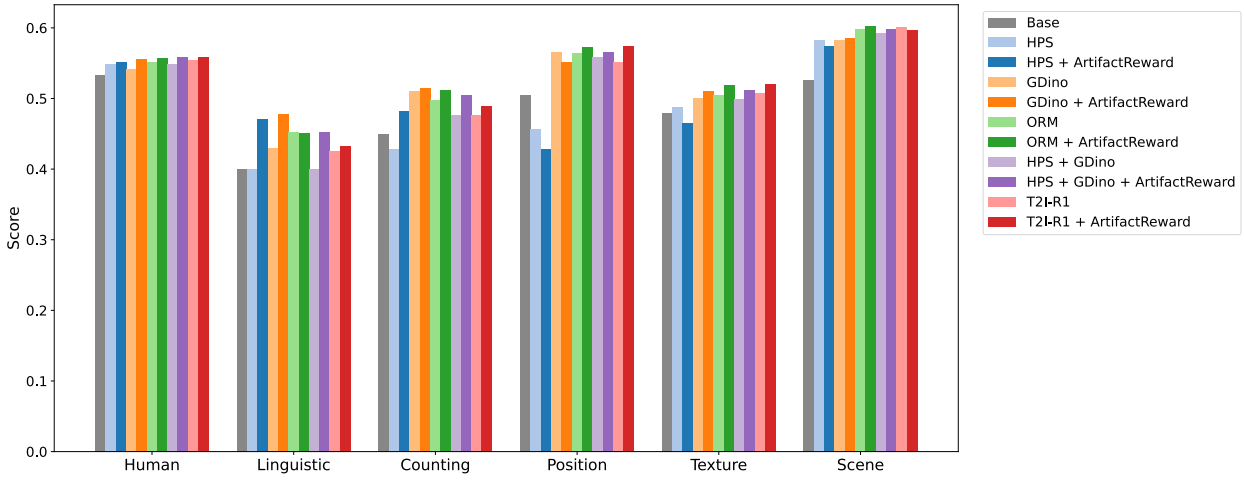
(a) HPS (b) HPS + ArtifactR (c) GDino (d) GDino + ArtifactR

(e) HPS + GDino (f) HPS + GDino + ArtifactR (g) ORM (h) ORM + ArtifactR

Figure 8. Images generated with prompt "*Siblings of identical twins*" in WISE [26] benchmark under different training reward configurations trained on Janus-Pro-1B [4].

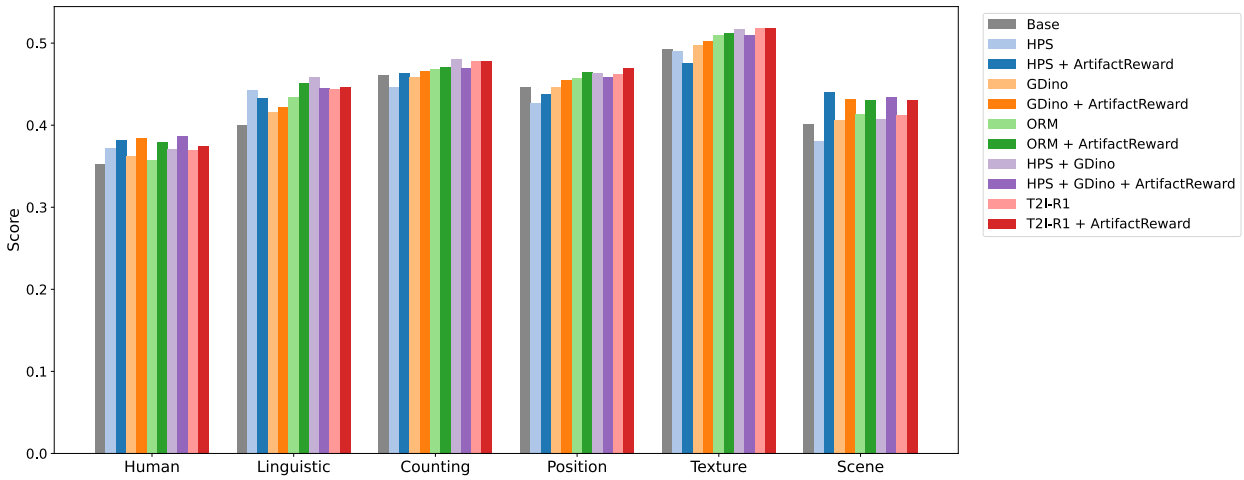


(a) Perception

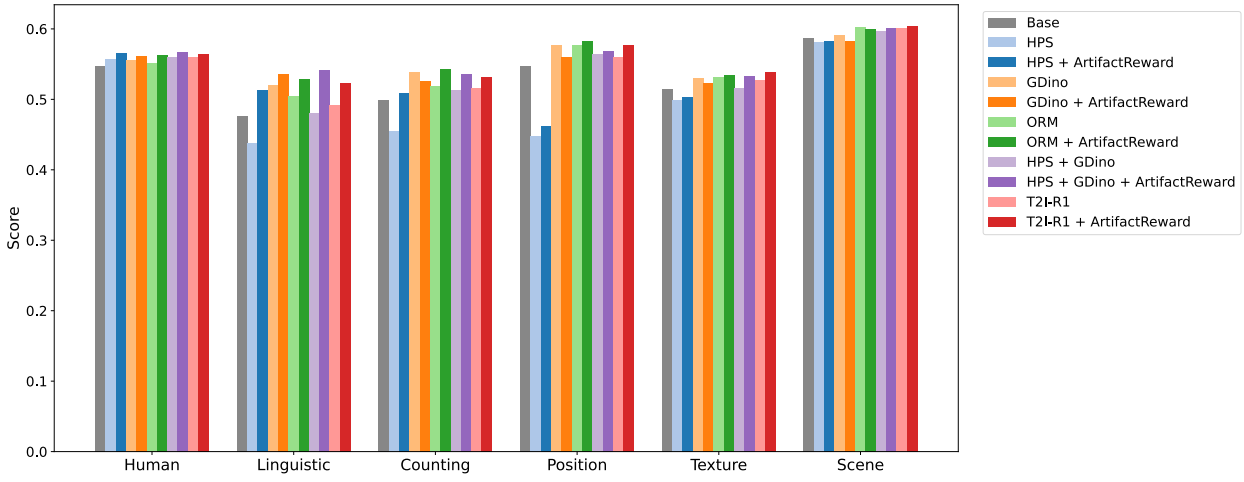


(b) Correspondence

Figure 9. Performance on LLM4LLM [37] benchmark across different categories trained on Janus-Pro-1B [4]



(a) Perception



(b) Correspondence

Figure 10. Performance on LLM4LLM [37] benchmark across different categories trained on Janus-Pro-7B [4]

(a) HPS

(b) HPS + ArtifactR

(c) GDino

(d) GDino + ArtifactR

(e) HPS + GDino

(f) HPS + GDino + ArtifactR

(g) ORM

(h) ORM + ArtifactR

Figure 11. Images generated with prompt "*A teacher in a white blouse stands at the blackboard, her curly brown hair tied back in a ponytail.*" in LLM4LLM [37] benchmark under different training reward configurations trained on Janus-Pro-1B [4].

(a) HPS

(b) HPS + ArtifactR

(c) GDino

(d) GDino + ArtifactR

(e) HPS + GDino

(f) HPS + GDino + ArtifactR

(g) ORM

(h) ORM + ArtifactR

Figure 12. Images generated with prompt "*a dog is smiling with happy emotion. He find a lot of delicious food.*" in LLM4LLM [37] benchmark under different training reward configurations trained on Janus-Pro-1B [4].

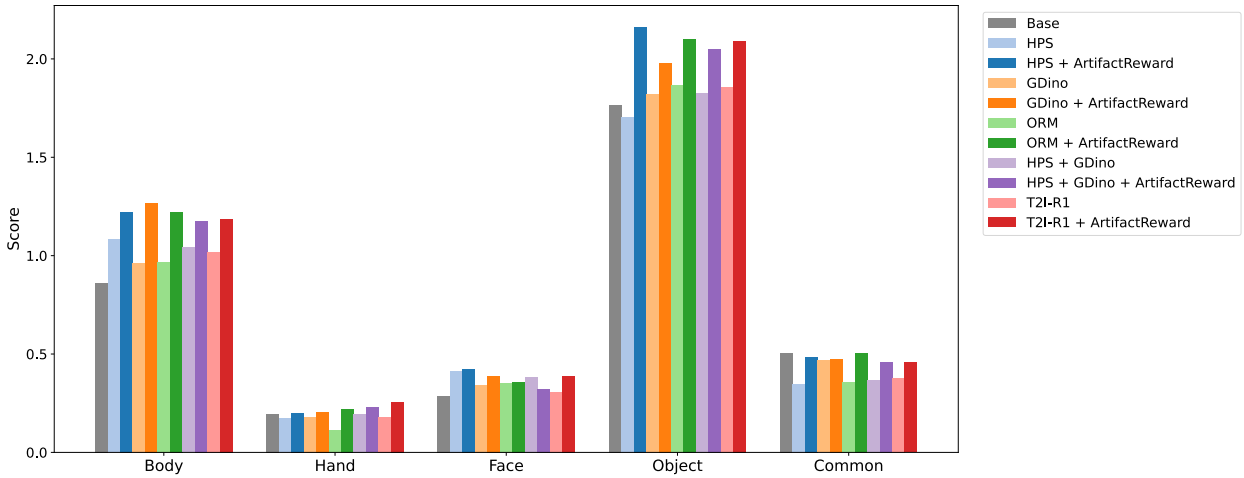


Figure 13. Performance on EvalAlign [35] benchmark across different categories trained on Janus-Pro-1B [4].

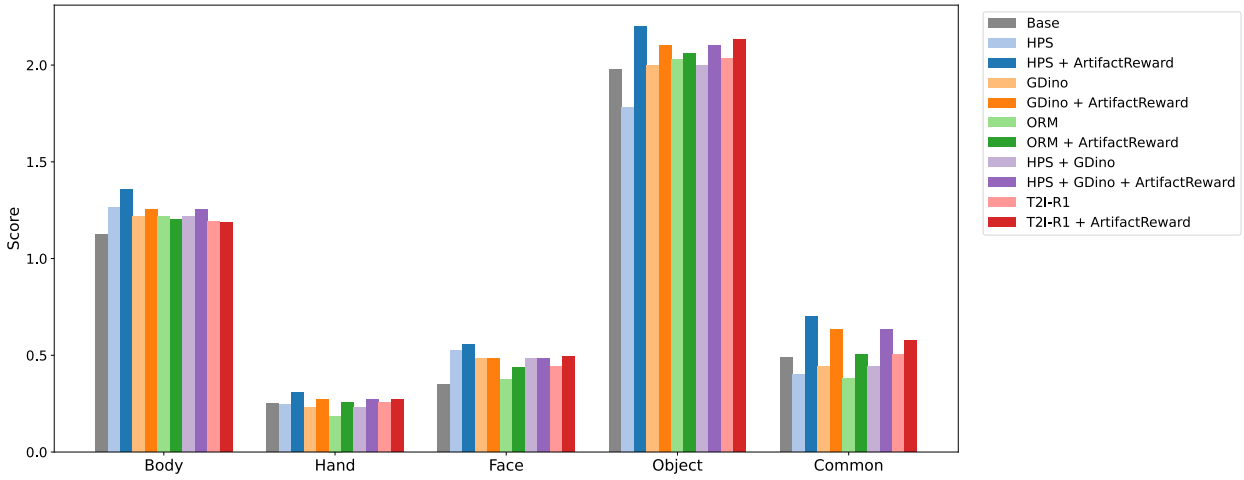


Figure 14. Performance on EvalAlign [35] benchmark across different categories trained on Janus-Pro-7B [4].