POCA: Pareto-Optimal Curriculum Alignment for Visual Text Generation

Supplementary Material

7. Normalized Edit Distance

Edit Distance (ED), also known as Levenshtein distance, measures the minimum number of operations required to transform one string into another. The allowed operations include character insertions, deletions, and substitutions, each contributing a unit cost to the total edit distance. This metric is widely used for text similarity evaluation. Normalized Edit Distance (NED) normalizes the computed ED by dividing it by the maximum length of the two strings, ensuring a value between 0 and 1, where 0 indicates identical strings and 1 represents completely different strings. The formal computation of NED is shown in Algorithm 2, where a two-dimensional dynamic programming (DP) table is used to iteratively compute the minimum edit cost between two input strings. In this work, we apply NED as our OCR reward, since Sen.ACC suffers from reward sparsity, which typically assigns a zero score to partially correct generations, failing to provide the fine-grained feedback necessary for optimization. In contrast, NED offers a continuous measure of character-level alignment. To align this distance metric with the standard RL objective of reward maximization, we formulate the training reward as $1 - \text{NED}$.

Algorithm 2 Normalized Edit Distance (NED)

Given two strings a and b of lengths n and m. Initialize DP table $D \in \mathbb{R}^{(n+1)\times(m+1)}$. Define edit cost c. Initialize $D_{i,0} \leftarrow i$, $D_{0,j} \leftarrow j$ for all i, j.
for $i = 1$ to n **do**
  for $j = 1$ to m **do**
    $D_{i,j} \leftarrow \min(D_{i-1,j} + 1, D_{i,j-1} + 1, D_{i-1,j-1} + I(a_i \neq b_j))$
return $D_{n,m} / \max(n, m)$

8. Pareto Set Comparison

To validate the effectiveness of our bi-directional strategy, we compare it against Parrot [16], which employs one-directional non-dominated sorting to update the policy using only the best samples. Additionally, we investigate the impact of negative samples by evaluating a fully dominated sorting baseline. Following the configuration in [44], we train the base model using each sorting strategy on the POCA-20k dataset for 300 steps. Fig. 10 illustrates that our bi-directional sorting algorithm consistently achieves higher reward curves across all three metrics compared to one-direction baselines. Table 3a further validates the superiority of our approach, demonstrating the overall best performance in Sen.ACC, CLIP score, and HPS score. The combined evidence from the training curves and quantitative metrics confirms that leveraging both positive and negative signals is essential for achieving the best convergence and overall performance in multi-reward policy optimization.

To further assess the upper-bound capability of the learned Pareto sets, we evaluate them using a best@k protocol. In detail, we randomly sample 100 prompts from both English and Chinese benchmarks and generate $k = 8$ candidates per prompt using each method. We then apply a rule-based selection strategy that prioritizes Sen.ACC, followed by CLIP score and finally HPS score, to pick the best image among all candidates. Table 3b indicates that our method has a higher probability of producing the optimal trade-off state in a larger solution space. Finally, we compare all methods in a unified global solution space of their generated solutions. We identify the global Pareto front by calculating the non-dominated points per-image across this combined set and quantify the relative contribution of each method, as shown in Fig. 9. For both English and Chinese subsets, our bi-directional approach dominates the global Pareto front, accounting for the largest proportion of non-dominated solutions.

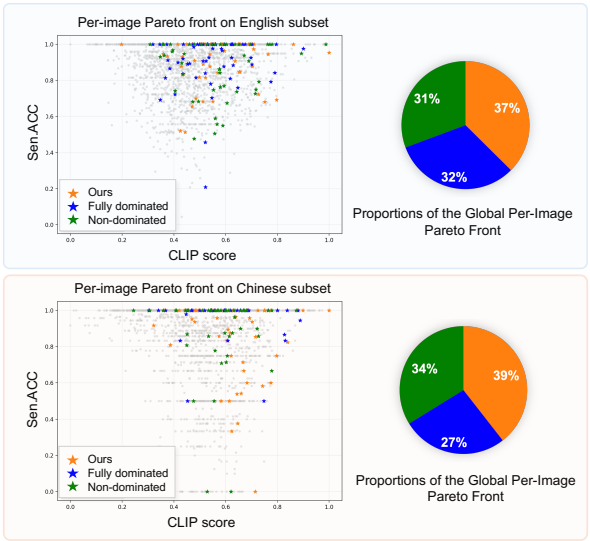

Figure 9. Contribution to the global Pareto front. The figure shows the fraction of globally non-dominated solutions contributed by each method in the unified solution space. Our bi-directional approach accounts for the largest share of optimal solutions, indicating a stronger coverage of the Pareto front than one-direction baselines.

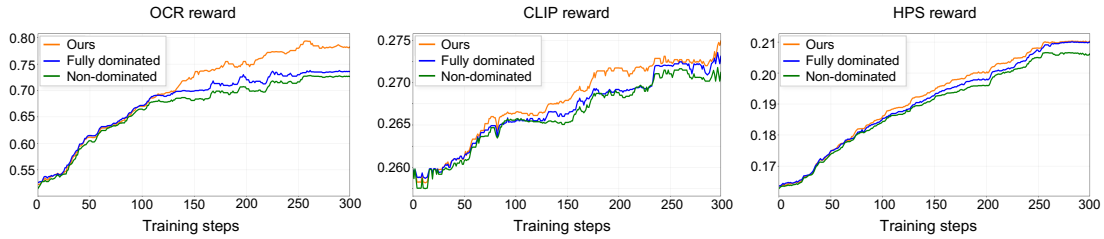

Figure 10. Reward curves for different Pareto selection algorithms. The figure illustrates that our bi-directional approach consistently achieves higher reward curves across all three metrics compared to one-direction baselines. This confirms the superior effectiveness of using both positive and negative signals for stable and efficient policy optimization.

Table 3. Quantitative comparison of different Pareto selection strategies.

(a) Averaged performance of Pareto sorting strategies. The table compares the averaged performance of one-direction methods against our bi-directional approach, validating that using both positive and negative signals leads to superior overall alignment across all metrics.

Methods	English				Chinese			
	Sen.ACC↑	NED↑	CLIP score↑	HPS score↑	Sen.ACC↑	NED↑	CLIP score↑	HPS score↑
GRPO baseline	0.7246	0.8935	0.8970	0.2708	0.6782	0.8663	0.8138	0.2668
Non-dominated set	0.7227	0.8827	0.8990	0.2710	0.6758	0.8655	0.8161	0.2674
Fully dominated set	0.7274	0.8836	0.9019	0.2712	0.6766	0.8656	0.8154	0.2671
Ours	**0.7378**	**0.8923**	**0.8996**	**0.2700**	**0.6867**	**0.8666**	**0.8163**	**0.2650**

(b) Evaluation of upper-bound generation capability. This table demonstrates that our bi-directional selection effectively expands the Pareto frontier, leading to a higher probability of discovering optimal trade-off solutions.

Methods	English				Chinese			
	Sen.ACC↑	NED↑	CLIP score↑	HPS score↑	Sen.ACC↑	NED↑	CLIP score↑	HPS score↑
Non-dominated set	0.8418	0.9220	0.9439	0.2760	0.8769	0.9496	0.8580	0.2702
Fully dominated set	0.8418	0.9223	0.9467	0.2762	0.8769	0.9475	0.8623	0.2725
Ours	**0.8491**	**0.9204**	**0.9438**	**0.2766**	**0.8821**	**0.9495**	**0.8640**	**0.2741**

9. Variance Analysis of Reward Models

To justify our choice of the OCR reward as the difficulty measure in the curriculum, we compare the distributions of all three reward signals (OCR, CLIP, and HPS) over the full training set. For each reward model, we compute scores for every sample and plot the ECDFs together with selected quantiles, as shown in Fig. 11. We observe that CLIP and HPS scores are highly concentrated in a narrow band between the 20th and 80th percentiles, leading to a very limited dynamic range in the central region of the dataset. In contrast, the OCR reward spans a much broader interval over the same percentile range, and its inter-percentile gaps (e.g., at the 20th, 40th, 60th, and 80th percentiles) are consistently larger than those of CLIP and HPS, which is also reflected by a noticeably larger variance. This indicates that OCR provides greater variance and finer ranking resolution across samples. These findings support our design choice that OCR reward provides a more informative and discriminative difficulty signal for curriculum scheduling.

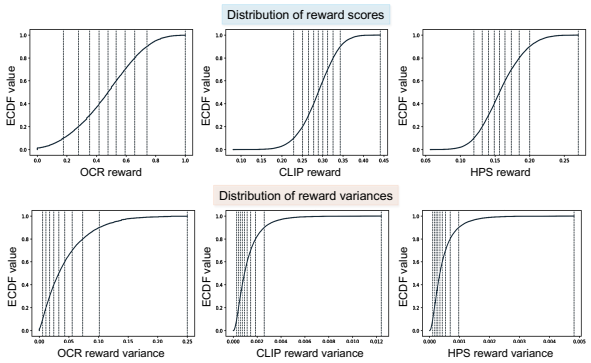

Figure 11. ECDFs for both the mean reward scores and the reward variances of OCR, CLIP, and HPS across the full training set. Vertical dashed lines indicate deciles (10%–100% in steps of 10%).

10. More Details About Dataset Preparation

Our image dataset is randomly sampled from the following datasets:

Table 4. Comparisons between POCA and the counterparts.

Methods	English				Chinese			
	Sen.ACC↑	NED↑	CLIP score↑	HPS score↑	Sen.ACC↑	NED↑	CLIP score↑	HPS score↑
RPO-Harmonic	0.7400	0.8875	**0.9029**	0.2678	0.6908	0.8684	0.8155	**0.2672**
Curriculum-DPO	0.7268	0.8866	0.8962	0.2602	0.6574	0.8435	0.8011	0.2598
POCA	**0.7651**	**0.8983**	0.8985	**0.2694**	**0.6942**	**0.8696**	**0.8170**	0.2653
Glyph-SDXL-v2	0.5950	0.7452	0.8553	0.2131	0.6174	0.7608	0.7796	0.2131
Pareto-guided-SDXL	0.6119	0.7635	0.8700	0.2293	0.6653	0.8281	**0.8042**	0.2294
POCA-SDXL	**0.6218**	**0.7775**	**0.8762**	**0.2302**	**0.6815**	**0.8371**	0.7996	**0.2332**

- SynthText [9] renders synthetic English text onto real-world background images spanning 200 classes, including human portraits, animals, and natural landscapes. Text is composited into these backgrounds using a rule-based rendering engine, producing diverse text images.
- AnyWord-3M [34] is a large-scale multilingual dataset extracted from publicly available images. These images cover a wide range of real-scene images containing text, including street views, advertisements and book covers.
- LeX-10K [51] is a curated collection of 10K English text–image pairs tailored for visual text generation, with a strong emphasis on aesthetics, text fidelity, and stylistic diversity. The images cover a wide range of layouts and themes, such as posters, logos and design-like images.

We sampled 5k images from SynthText [9], which contains all 200 classes, 10k images from AnyText-3M [34] with 5k English text images and 5k Chinese text images, and 5k images from LeX-10K [51]. Examples of collected images are shown in Fig. 12. For training prompts, we leveraged Gemini 2.5 [4] to describe each image and Fig. 13 illustrates the instruction used for prompt generation.

Figure 12. Examples of the diverse image domains in our image dataset.

11. POCA on Larger Model

Focusing on visual text generation, we use the state-of-the-art AnyText for our main experiments. To demonstrate that POCA is model-agnostic, we also evaluate it using the more recent Glyph-SDXL-v2. Table 4 shows the results. Obviously, using Bi-directional Pareto sorting (Pareto-guided-SDXL) can significantly improve the performance on all metrics and introducing curriculum learning (POCA-SDXL) results in an additional improvement. Note that the base models of Glyph-SDXL-v2 and AnyText are trained using different datasets, hence their performances vary. Nevertheless, applying POCA to them leads to remarkable improvements, which indicates that POCA is effective and general.

12. More Comparisons with Related Works

In this section, we further compare POCA with additional related methods, including 1) the weighted-sum approach RPO [26] and 2) the DPO-based curriculum design, Curriculum-DPO [5].

RPO proposes using the harmonic mean instead of a naive weighted-sum approach to aggregate different rewards in a two-reward setting, so as to place greater emphasis on smaller rewards. In other words, a high final reward is obtained only when both rewards are relatively large. We extend the harmonic reward function in RPO to support three rewards: $r = \frac{1}{\frac{\lambda}{r_{ocr}}+\frac{\alpha}{r_{clip}}+\frac{\beta}{r_{hps}}}$, where $\lambda = \alpha = \beta = \frac{1}{3}$. As shown in Table 4, POCA outperforms RPO on multiple metrics, especially Sen.ACC. Unlike RPO and other weighted-sum methods, POCA avoids the difficulty of balancing aggregation hyperparameters.

While Curriculum-DPO builds an easy-to-hard learning path by ranking candidate samples with a single reward model and progressively training on preference pairs from coarse, easily distinguishable pairs to finer, harder ones, POCA is designed to address multiple conflicting rewards, leading to distinct sample ranking strategies. We compared POCA with Curriculum-DPO by using OCR rewards for pair ranking. As shown in Table 4, Curriculum-DPO surpasses the AnyText baseline and performs similarly to GPRO, but remains inferior to POCA, particularly across multiple metrics.

13. Assessment of computational overhead:

Generating the 20k training prompts using Gemini 2.5 took ∼40 hours. Performing inference on the entire set of prompts for difficulty measurement requires ∼15 hours with 8 GPUs.

14. More Visual Examples

We show additional visual examples of POCA in this section. We first provide more examples for comparison with

Instructions for Prompt Generation

Input: <Image>, <Text>

You are an expert prompt engineer for Stable Diffusion. Analyze the provided image and generate a high-quality that works for SD1.5 based model. **Critical Rule:**

1. The description **MUST ONLY** be about the visual elements (objects, background, style, lighting, colors). **IGNORE** the following text elements found in the image: <Text>. Do not mention them.
2. The description should be a short descriptive sentence, followed by comma-separated keywords.
3. The detailed caption should be human-readable and fluent.
4. The entire rewritten prompt must be a single line and should not exceed 150 words.

Now, generate a new prompt for the provided image.

Figure 13. Rule-based instructions utilized with the Gemini 2.5 model to generate high-fidelity prompts for the POCA-20k dataset.

other baselines in Fig. 14. These images illustrate that POCA maintains image coherence while generating accurate text, allowing the text to seamlessly integrate with the background. Fig. 15 shows the comparison with the standard weighted-sum GRPO method. While the GRPO baseline brings some degree of improvement in accuracy and aesthetics to the base model, it struggles to properly balance multiple reward objectives. For example, it often generates semantically inconsistent superfluous elements and extra text. Conversely, our Pareto-guided method improves image aesthetics by increasing the level of visual detail while remaining faithful to the semantics. POCA further improves the accuracy and clarity of the text.

Finally, we use longer and more complex prompts in Fig. 16 to test POCA, AnyText, and AnyText2 for their instruction-following capability. The results show that POCA generates the images most faithful to the instructions, such as the ”gold holly motifs and red berry clusters” (column one), the ”shadowy cloaked figure” (column three), and the ”gradient lighting from the upper left” (column five). This large number of visual examples demonstrates that the images generated from POCA are improved across multiple reward criteria.

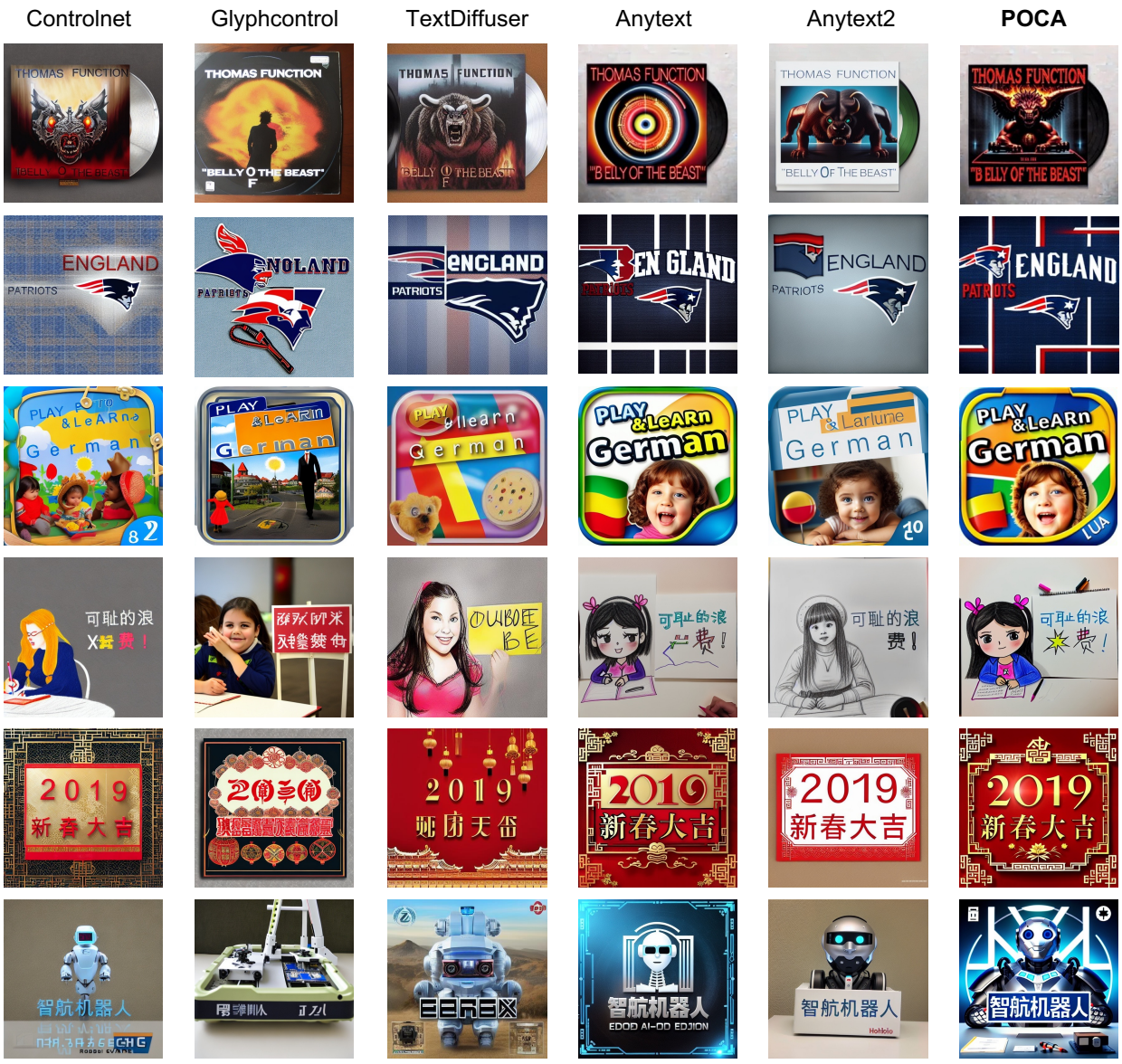

Figure 14. General qualitative comparison of POCA and other baselines. The figure demonstrates POCA's superior overall balance between text accuracy, image coherence, and aesthetic appeal compared to state-of-the-art methods.

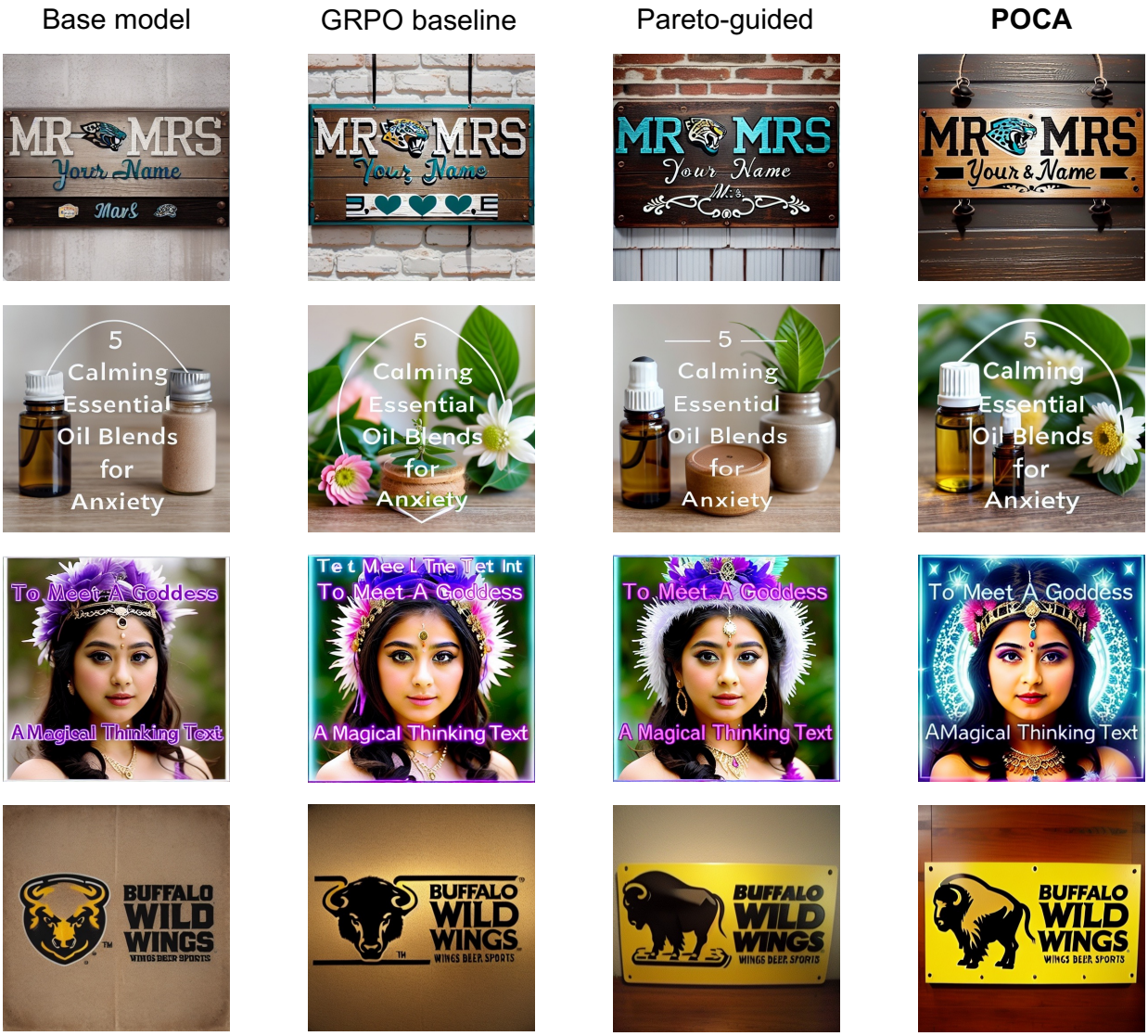

Figure 15. Qualitative comparison with standard GRPO baseline. Although the GRPO baseline improves both aesthetics and text accuracy to some extent, the inconsistent reward signals lead to visually unbalanced images with excessive text and semantic inconsistencies.

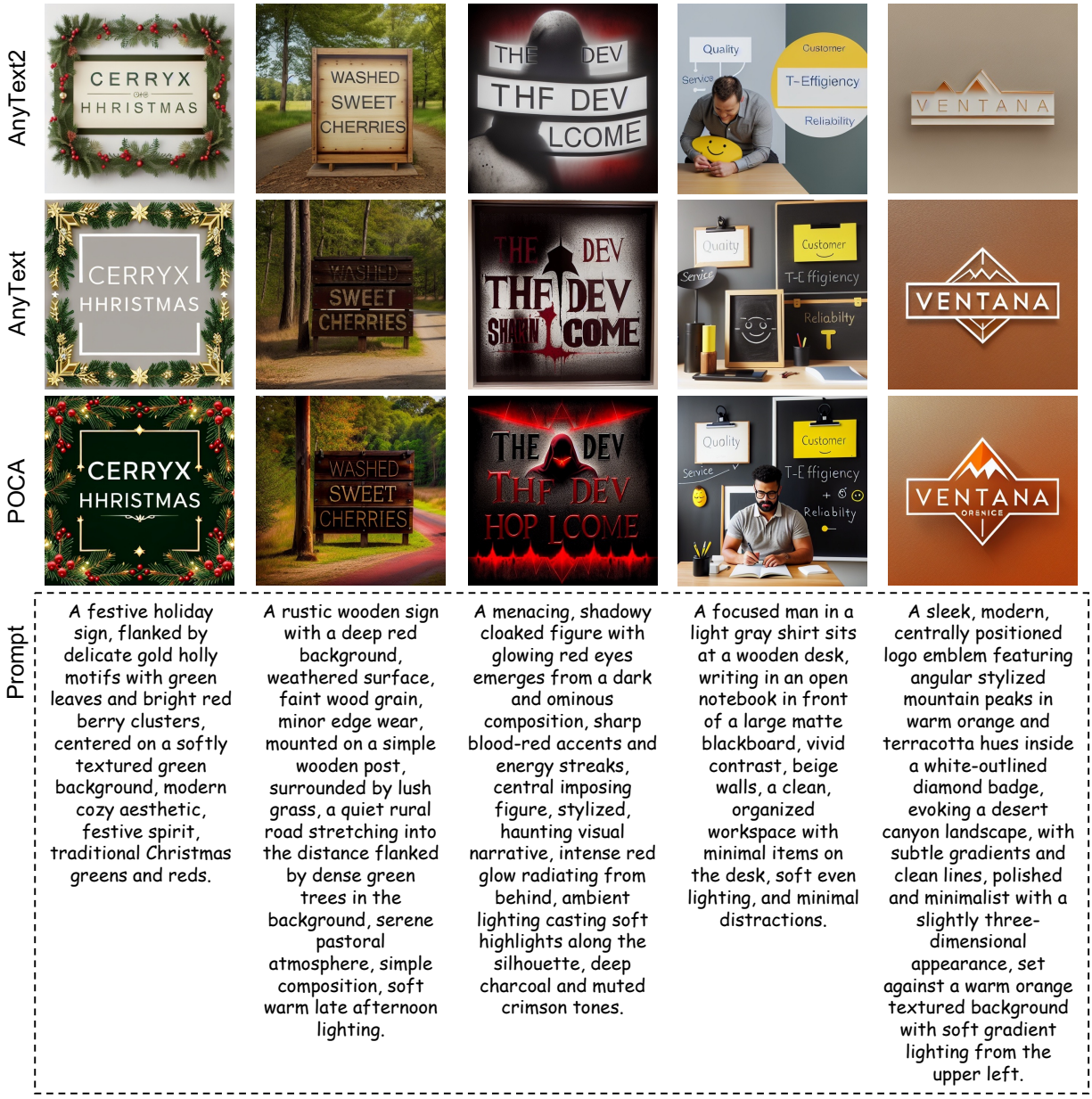

Figure 16. Evaluation of complex instruction-following capability. POCA demonstrates fidelity to instructions and higher control over fine-grained details under complex prompts, outperforming baselines on these challenging tasks.