F Additional Evaluation Metrics

In this section, we include evaluation metrics beyond the task success rates. Results for aesthetic quality, image diversity, and text-to-image alignment are presented in Figure 8.

Aesthetic Quality. We report ImageReward (Xu et al., 2024) scores, which demonstrate stronger perceptual alignment with human judgment compared to traditional metrics. Higher scores reflect better aesthetic quality. Although human evaluators prioritized task success based on the criteria in Appendix C over aesthetic quality and were not instructed to consider aesthetics, HERO demonstrates comparable aesthetic performance to the baselines, surpassing them in 3 out of 5 tasks.

Image Diversity. Following Section 4.3.3 of von Rütte et al. (2023), we compute "In-Batch Diversity", defined as the complement of the average similarity of CLIP image embeddings (Radford et al., 2021) between pairs of images in a generated batch. Specifically, for a batch of N generated images $I_1, I_2, \ldots, I_N$, and the cosine similarity $\text{CLIPSim}(I_i, I_j)$ of their embeddings in the CLIP feature space, the in-batch diversity is calculated as: $D_{\text{batch}} = 1 - \frac{2}{N(N-1)} \sum_{1 \leqslant i < j \leqslant N} \text{CLIPSim}(I_i, I_j)$, where $1 - \text{CLIPSim}(I_i, I_j)$ represents the dissimilarity between two images. A higher D_{batch} signifies greater diversity. Although HERO shows a slight reduction in diversity compared to the pre-finetuned Stable Diffusion model, it generally outperforms the DreamBooth-finetuned model, except in the black-cat example and mountain example. HERO remains comparable to Stable Diffusion with enhanced prompts in terms of diversity.

Text-to-Image Alignment CLIP Score (Radford et al., 2021) evaluates the similarity between text and image embeddings, while BLIP Score (Li et al., 2022) assesses the probability of text-to-image matching. Together, these metrics provide a quantitative measure of how well the generated images align with the given prompts. Higher scores on both metrics indicate better alignment between the generated images and the prompts. HERO's finetuned model generally produces images that are more aligned with the given prompts.

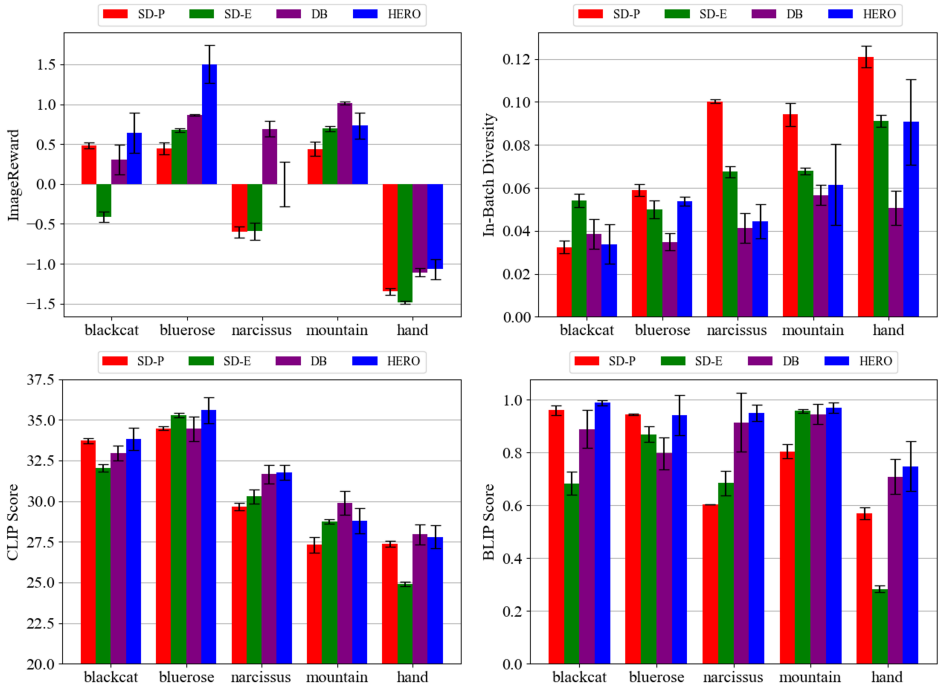

Figure 8: Additional evaluation results. For all metric, higher value indicates better performance. **Top Left.** Aesthetic quality measured with ImageReward (Xu et al., 2024). **Top Right.** In-Batch Diversity computation following Radford et al. (2021). **Bottom.** CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) Text-to-image alignment scores.