Supplementary Material for Verify Claimed Text-to-Image Models via Boundary-Aware Prompt Optimization

Zidong Zhao[1], Yihao Huang[2†], Qing Guo[3], Tianlin Li[4], Anran Li[5], Kailong Wang[6], Jin Song Dong[7], Geguang Pu[2]

[1]Zhejiang University, China, [2]East China Normal University, China, [3]Nankai University, China, [4]Beihang University, China, [5]University of Science and Technology of China, China, [6]Huazhong University of Science and Technology, China, [7]National University of Singapore, Singapore

A. Algorithm for Stage 3

Algorithm 1 for Stage 1: Anchor Identification and Algorithm 2 for Stage 2: Boundary Exploration are detailed in the main paper. We supplement these by providing the detailed pseudocode for Stage 3: Targeted Optimization in Algorithm 3. This algorithm's objective is the inverse of Stage 1; instead of maximizing the distance from the original prompt, it aims to maximize the cosine similarity with the target boundary embedding e_{α^*} found in Stage 2. This is achieved by minimizing a loss function $\mathcal{L}$ set to the negative cosine similarity, driving the optimized prompt P_v towards the model-specific semantic boundary.

Algorithm 3: BPO Stage 3: Targeted Optimization

Input: Original prompt I, target embedding e_{α^*}, model $M_t(E_t, G_t)$, initial suffix s, maximum iterations K, batch size B
Output: Verification prompt P_v

Initialize $s_{1:n}$
for $i = 1$ **to** K **do**
  $P_i \leftarrow I + s_{1:n}$
  $\mathcal{L}(s_{1:n}) \leftarrow -\cos(E_t(P_i), e_{\alpha^*})$
  for $j = 1$ **to** n **do**
    $\mathcal{X}_j \leftarrow \text{Top-}k(-\nabla_{\mathbf{e}_{s_j}} \mathcal{L}(s_{1:n}))$
  for $b = 1$ **to** B **do**
    $s_{1:n}^{(b)} \leftarrow \text{Uniform}(\mathcal{X}_{j^*})$, where $j^* \leftarrow \text{Uniform}(1,n)$
  $s_{1:n} \leftarrow s_{1:n}^{(b^*)}$, where $b^* \leftarrow \arg\min_b \mathcal{L}(s_{1:n}^{(b)})$
$P_v \leftarrow I + s_{1:n}$
return P_v

B. List of Benign Prompts

As mentioned in Section 5.1 of the main paper, our experiments required a set of base prompts to initiate the optimization. We utilized 10 original benign prompts (I) randomly generated by GPT-4o to serve as this starting point for the BPO algorithm. The complete list of these prompts is provided in Table 4.

Table 4. The 10 benign prompts used for BPO experiments.

	Prompt
1	a bird
2	a blueberry
3	a fish
4	a fox
5	a grapefruit
6	a lettuce
7	a peach
8	a pig
9	an apple
10	an asparagus

C. Additional Visualization Results

As mentioned in Section 5.2 of the main paper, we provide more extensive visualization results here to supplement Figure 5 from the main text. Figure 8 demonstrates the outputs of verification prompts (P_v) generated by BPO for different target models, tested across all five models.

The first column on the left shows the target model (the model the P_v was designed to detect), while the right columns display non-target models using the exact same prompt. As shown, the target model's outputs (labeled "unstable") exhibit high **semantic instability**, appearing as a

† Yihao Huang (huangyihao@sei.ecnu.edu.cn) is the corresponding author.

chaotic mix of concepts (*e.g.*, grapefruits and flooring, or pigs and trains). In contrast, non-target models (labeled "stable") produce semantically consistent outputs, even if converging on a different concept. This consistent pattern validates that the semantic boundaries identified by BPO are indeed model-specific and serve as reliable fingerprints for verification.

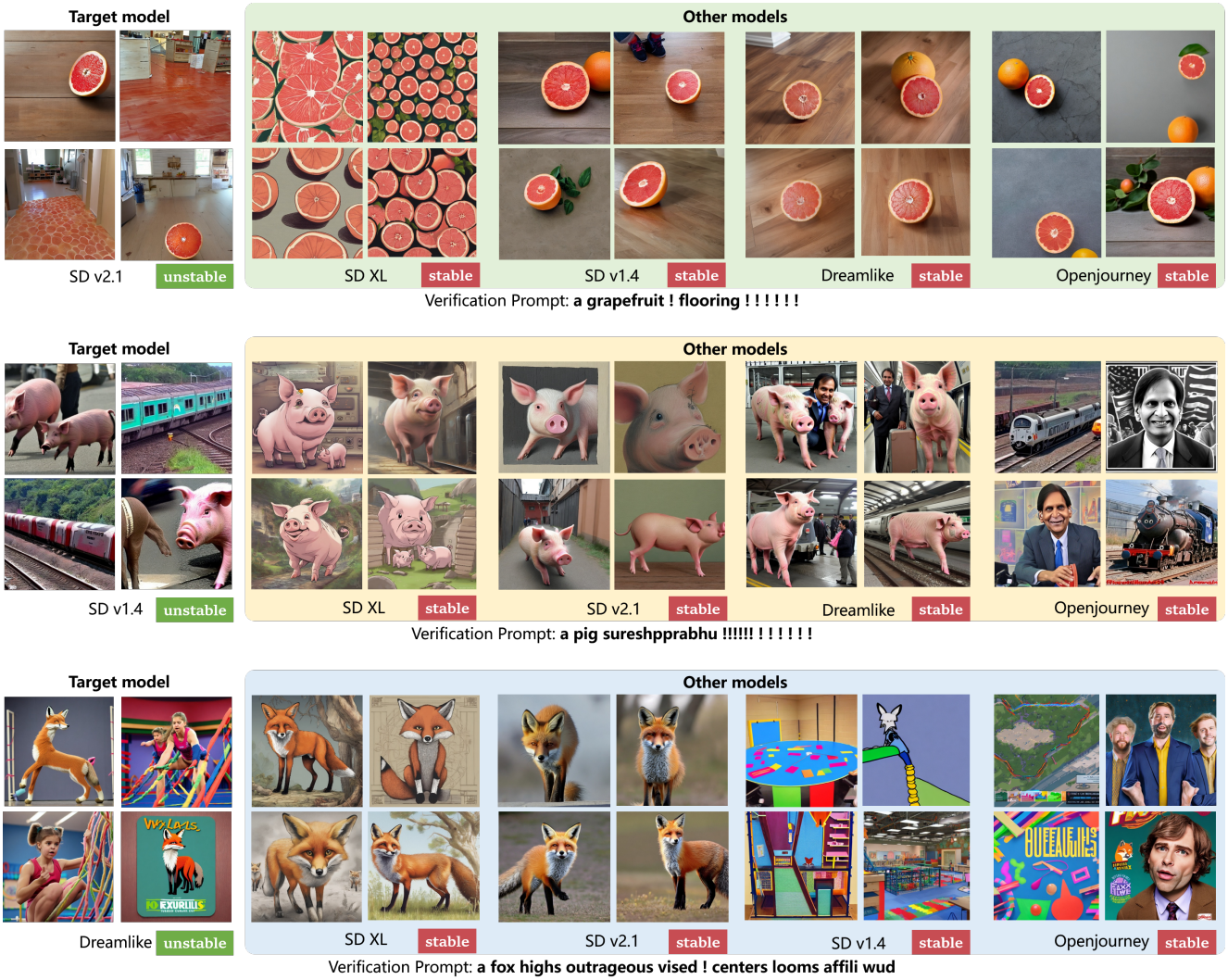

Figure 8. Comparison of generated images between the target model and other models using the same verification prompt. Each row shows a verification prompt generated by BPO targeting a specific model (*e.g.*, SD v1.4, SD v2.1, SDXL, Dreamlike, Openjourney). As described in the main paper, the target model exhibits "unstable" outputs while other models remain "stable".