Optimizing Certified Radius of Zero-shot Composed Image Retrieval via Text Guidance

Supplementary Material

8. Proofs of Theorem 1

Theorem 1. *If the Recall@K of a query feature $\langle \hat{v}_s, w_r \rangle$ is 1 and the perturbation δ satisfies the following restrictions:*

$$\begin{aligned} &||v_r^{att} - \hat{v}_s||_2 \\ &< \frac{d_v}{2(||\hat{v}_s - v_k||_2 + ||\hat{v}_s - v_t||_2)} \\ &+ \frac{d_w}{2(||v_r^{att} - v_k||_2 + ||v_r^{att} - v_t||_2)}, \end{aligned} \tag{18}$$

Proof. Let $d_a = ||\hat{v}_s - v_k||_2 + ||\hat{v}_s - v_t||_2$ and $d_f = ||v_r^{att} - v_k||_2 + ||v_r^{att} - v_t||_2$.

According to Definition 2, if the Recall@K of a query feature $\langle \hat{v}_s, w_r \rangle$ is 1, then $||\hat{v}_s - v_t||_2 \leq ||\hat{v}_s - v_k||_2 - \frac{d_v}{d_a}$.

Note that $d_v^{att} = d_f \cdot (||v_r^{att} - v_k||_2 - ||v_r^{att} - v_t||)$, so our goal is to prove: $||v_r^{att} - v_t|| < ||v_r^{att} - v_k||_2 + \frac{d_w}{d_f}$ if $||v_r^{att} - \hat{v}_s||_2 < \frac{d_v}{2d_a} + \frac{d_w}{2d_f}$.

$$\begin{aligned} &||v_r^{att} - v_t||_2 \\ &\leq ||(v_r^{att} - \hat{v}_s) + (-v_t + \hat{v}_s)||_2 \\ &\leq ||v_r^{att} - \hat{v}_s||_2 + ||v_t - \hat{v}_s||_2 \\ &< \frac{d_v}{2d_a} + \frac{d_w}{2d_f} + ||\hat{v}_s - v_k||_2 - \frac{d_v}{d_a} \\ &\leq \frac{d_w}{2d_f} + ||\hat{v}_s - v_k||_2 - \frac{d_v}{2d_a} \\ &< ||\hat{v}_s - v_k||_2 - ||v_r^{att} - \hat{v}_s||_2 + \frac{d_w}{d_f} \\ &\leq ||(\hat{v}_s - v_k) + (v_r^{att} - \hat{v}_s)||_2 + \frac{d_w}{d_f} \\ &\leq ||v_r^{att} - v_k||_2 + \frac{d_w}{d_f} \end{aligned}$$

□

9. Proofs of Lemma 2

Lemma 2. *Leaving $\beta = \ln(\frac{\alpha/2}{k+1})$ and probability at least $1 - \alpha$, we have*

$$\underline{d_\delta}(\hat{v}_s) := d_\delta(\hat{v}_s) - \frac{-\beta + \sqrt{\beta^2 - 18n\beta}}{3n} \tag{19}$$

Proof. Let

$$\kappa = \frac{-\ln(\frac{\alpha}{k+1}) + \sqrt{(\ln(\frac{\alpha}{k+1}))^2 - 18n\ln(\frac{\alpha}{k+1})}}{3n}$$

. According to Eq. 12, with probability at least $1 - \alpha$, we have $\|\hat{v}_s - v_s\|_2 \leq \kappa$. We first bounded the difference between $||\hat{v}_s - v_k||_2$ and $||v_s - v_k||_2$ with probability at least $1 - \alpha$:

$$\begin{aligned} &|||v_s - v_k||_2 - ||\hat{v}_s - v_k||_2| \\ &\leq ||v_s - v_k - \hat{v}_s + v_k||_2 \\ &\leq ||v_s - \hat{v}_s||_2 \\ &= \kappa \end{aligned}$$

Analogously, $|||v_s - v_t||_2 - ||\hat{v}_s - v_t||_2| \leq \kappa$. Thus, we can derive the lower bound of the perturbation margin. With probability at least $1 - 2\alpha$:

$$\begin{aligned} d_\delta(v_s) &= \frac{||v_s - v_k||_2 - ||v_s - v_t||_2}{2} + \frac{d_w}{8} \\ &\geq \frac{||\hat{v}_s - v_k||_2 - \kappa}{2} - \frac{||\hat{v}_s - v_t||_2 + \kappa}{2} + \frac{d_w}{8} \\ &= d_\delta(\hat{v}_s) - \kappa =: \underline{d_\delta}(\hat{v}_s) \end{aligned}$$

Replace α by $\frac{\alpha}{2}$ we obtain Lemma 2. □

10. Noise Optimization

The optimization problem 16 exists with a simple and straightforward approach to its solution: for each noisy image, we desire to find an optimal noise that minimizes the distance between the noisy image features and the target features. We assume that the nearest optimal noise for the sampled noise ϵ_i is ϵ_i^*. We use the second-order Taylor expansion at ϵ_i^* and the optimization problem Eq. 16 can be approximated as:

$$\mathbb{E}_{\epsilon_i}[L(E_v(I_r + \epsilon_i^*), w_r) + \frac{1}{2}(\epsilon_i^* - \epsilon_i)^\top \mathbf{H}_i(\epsilon_i^* - \epsilon_i)], \tag{20}$$

where $\mathbf{H}_i$ is the Hessian matrix at ϵ_i^*. We find that the key to optimizing Eq. 20 is to reduce $\mathbb{E}[(\epsilon_i^* - \epsilon_i)^\top \mathbf{H}_i(\epsilon_i^* - \epsilon_i)]$. Inspired by Chen et al. [6], we can adjust the optimized noise ϵ^* so that it converges to a point that is near each local optimum noise ϵ_i^*.

Lemma 3 ([6]). *Given the gradient of the loss function $g_i = \nabla_{I_r + \epsilon_i} L(E_v(I_r + \epsilon_i), w_r)$, we can optimize the dot product similarity between the gradients of all models to minimize the upper bound of $\frac{1}{n}\sum_{i=1}^n \|\epsilon_i - \epsilon_i^*\|_2^2$:*

$$\frac{1}{n}\sum_{i=1}^{n} \|\epsilon_i - \epsilon_i^*\|_2^2 \leq -2M\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{i-1} g_i g_j, \tag{21}$$

where $M = \max \|\mathbf{H}_i^{-1}\|_F^2$ and $\|\cdot\|_F$ is the Frobenius norm.

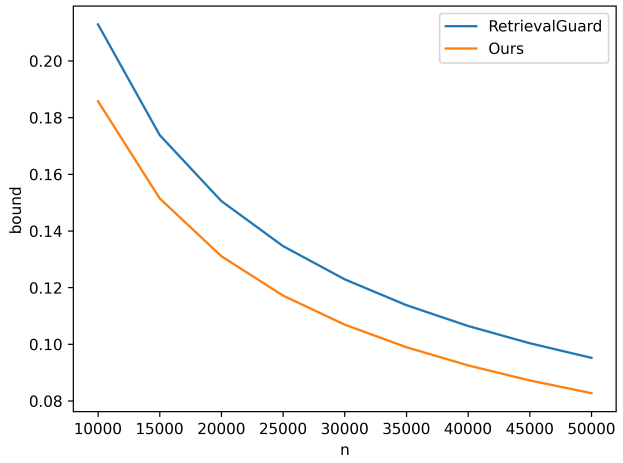

Figure 5. Comparison of Our Boundaries with RetrievalGuard.

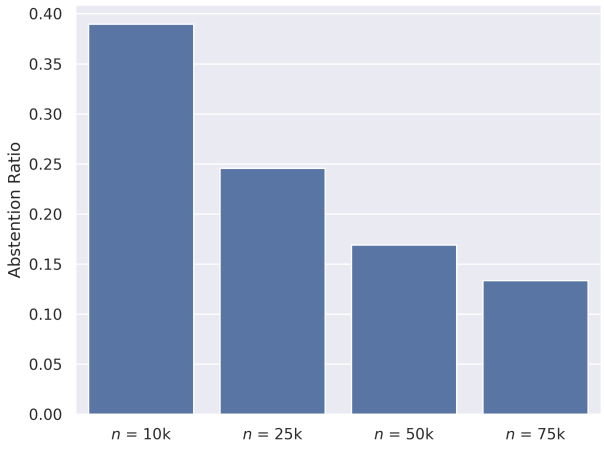

Figure 6. Impact of varying number of Monte Carlo samples n on abstention ratio.

To avoid second-order derivatives, we use a first-order derivative approximation algorithm [28] for optimization. Convergence to the final optimized noise ϵ^* is achieved by updating the gradient sequentially for each noise ϵ_i sampled from the normal distribution. Formally,

$$\epsilon_i^* = \epsilon_{i-1}^* - \nabla_{I_r+\epsilon_{i-1}^*+\epsilon_i} L(E_v(I_r + \epsilon_{i-1}^* + \epsilon_i), w_r), \quad (22)$$

where $\epsilon_0^* = 0$. When the n noises are optimized, we obtain the final noise $\epsilon^* = \epsilon_n^*$. Then, we employ the optimized noise ϵ^* to compute the smooth embedding. Finally, we compute the robustness radius r with the optimized smooth embedding.

11. Analysis of Randomized Smoothing Boundaries

As shown in Figure 5, the Lipschitz bound of RetrievalGuard is much worse than that of ours with all hyperparameters being equal.

12. Analysis of Abstention Ratio with the Different Monte-Carlo Samples.

Figure 6 shows the abstention ratio for different Monte Carlo samples n. The abstention rate is the fraction of samples where the smooth model can retrieve the target image ($d_\delta(\hat{v}_s) > 0$) but cannot guarantee robustness ($\underline{d_\delta}(\hat{v}_s) < 0$). It is evident that the abstention ratio decreases steadily as n increases. This suggests that by increasing the Monte Carlo sample size, the model becomes increasingly confident in its predictions, allowing it to provide stronger robustness guarantees. However, it should be noted that the rate of decline in the abstention rate slows down as n escalates. For instance, when n advances from 10,000 to 25,000, the abstention rate drops by roughly 0.14 (from 0.38 to 0.24).

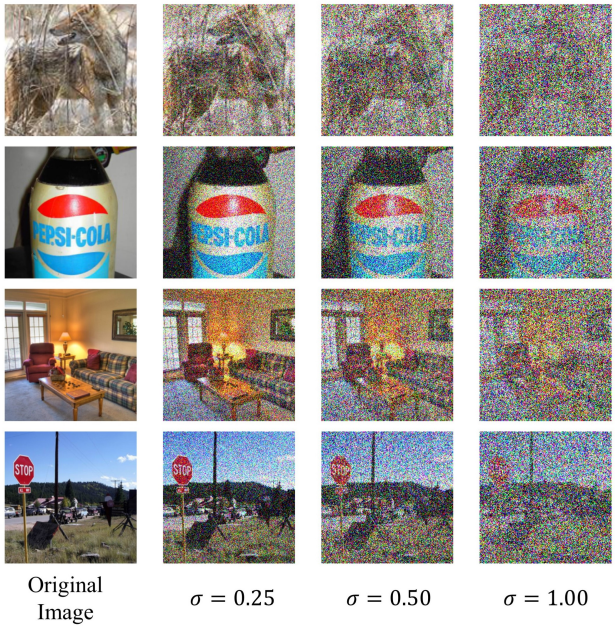

Figure 7. Example visualization of Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ being added to an image. The images in the top two rows are from CIRR, the rest are from GeneCIS. We clipped pixel values greater than 1.0 (=255) to 1.0 and pixel values less than 0.0 (=0) to 0.0.

Conversely, when n moves from 50,000 to 75,000, the reduction in the abstention rate is only roughly 0.04 (from 0.17 to 0.13). By prudently adjusting n, we can attain a favorable equilibrium between robustness guarantees and efficacy.

13. Visualization of Noisy Images

Our method constructs a smoothing model by adding Gaussian noise to the reference image. We show examples of CIRR and GeneCIS images disturbed by varying degrees of noise in Figure 7.

14. Derivation of Margin Function for Recall@K (Definition 2)

Given that all encoder outputs $(\hat{v}_s, w_r, v_t, v_k)$ are ℓ_2 normalized (i.e., $\|v_t\|_2 = \|v_k\|_2 = 1$).

Let $\mathbf{a} = \hat{v}_s + w_r$ (simplify combined feature). The Margin Function is defined as two expressions:

$$d\left(\hat{v}_s, w_r\right) = 2\|\mathbf{a}\|_2 \cdot \left(\cos(\mathbf{a}, v_t) - \cos(\mathbf{a}, v_k)\right) \tag{23}$$

By $\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2}$, and $\|v_t\|_2 = \|v_k\|_2 = 1$: $\cos(\mathbf{a}, v_t) = \frac{\mathbf{a}\cdot v_t}{\|\mathbf{a}\|_2}$, $\cos(\mathbf{a}, v_k) = \frac{\mathbf{a}\cdot v_k}{\|\mathbf{a}\|_2}$. Substitute into Eq. 23; cancel $\|\mathbf{a}\|_2$: $d\left(\hat{v}_s, w_r\right) = 2\|\mathbf{a}\|_2 \cdot \frac{\mathbf{a}\cdot(v_t - v_k)}{\|\mathbf{a}\|_2} = 2\mathbf{a} \cdot (v_t - v_k)$

Replace $\mathbf{a} = \hat{v}_s + w_r$, and expand the dot product: $d\left(\hat{v}_s, w_r\right) = 2\left[\hat{v}_s \cdot (v_t - v_k) + w_r \cdot (v_t - v_k)\right]$ Split into two terms:

$$d\left(\hat{v}_s, w_r\right) = 2\hat{v}_s \cdot (v_t - v_k) + 2w_r \cdot (v_t - v_k) \tag{24}$$

For any x, y, $\|x - y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2x \cdot y$. Rearrange to: $2x \cdot (y_1 - y_2) = \left(\|x - y_2\|_2^2 - \|x - y_1\|_2^2\right)$ (Note: $\|y_1\|_2^2 = \|y_2\|_2^2 = 1$ cancels out.) Apply to Eq. 24:

- $2\hat{v}_s \cdot (v_t - v_k) = \|\hat{v}_s - v_k\|_2^2 - \|\hat{v}_s - v_t\|_2^2$
- $2w_r \cdot (v_t - v_k) = \|w_r - v_k\|_2^2 - \|w_r - v_t\|_2^2$

Substitute the two results back into Eq. 24: $d\left(\hat{v}_s, w_r\right) = \underbrace{\left(\|\hat{v}_s - v_k\|_2^2 - \|\hat{v}_s - v_t\|_2^2\right)}_{\text{Image term}} + \underbrace{\left(\|w_r - v_k\|_2^2 - \|w_r - v_t\|_2^2\right)}_{\text{Text term}}$

15. Derivation of Equation 8 to Equation 9

All encoder outputs are ℓ_2-normalized ($\forall x, \|x\|_2 = 1$). By the triangle inequality: $\|x - y\|_2 \leq \|x\|_2 + \|y\|_2 = 2$. For $x = v_r^{att}$ and $y \in \{v_k, v_t\}$, this gives: $\|v_r^{att} - v_k\|_2 \leq 2$, $\|v_r^{att} - v_t\|_2 \leq 2$. Summing these bounds: $\|v_r^{att} - v_k\|_2 + \|v_r^{att} - v_t\|_2 \leq 2 + 2 = 4$. Substitute this into the d_w-term of Equation (8). Since the denominator is lower-bounded by 4, the term is upper-bounded by: $\frac{d_w}{2(\|v_r^{att} - v_k\|_2 + \|v_r^{att} - v_t\|_2)} \leq \frac{d_w}{2\cdot 4} = \frac{d_w}{8}$.

Here we assume that $d_w > 0$. When $d_w \leq 0$, text contributes nothing to retrieval, differing from the CIR setting and not related to our method design.

16. Experimental Setup for the GME Model

Experiments were conducted on the GME-Qwen2-VL 2B model, with all tests performed on a 24GB NVIDIA 3090 GPU. To fit the model and accommodate the randomized smoothing pipeline within the 24GB GPU memory constraint, we adopted "bfloat16" precision for all computations. For consistency with general ZS-CIR experimental settings, the image processing pipeline remains fully aligned with the CLIP architecture. We also preserve gradient flow through the entire preprocessing pipeline: we modified the Qwen2VLProcessor to replace non-differentiable operations with differentiable alternatives during gradient computation, ensuring effective backpropagation of adversarial perturbations for robustness evaluation. Given that the computational cost of GME-Qwen2-VL 2B is substantially higher than that of the CLIP model (used in general settings), a key adjustment was made to balance efficiency without compromising the reliability of the result: The number of Monte Carlo samples for randomized smoothing was set to $n = 1000$. This adjustment reduced the inference time per query to approximately 1 minute, representing a 90% reduction in latency compared to $n = 10,000$. Importantly, the GME-Qwen2-VL 2B model still achieved robust empirical robustness under this setting (see Table 2 for detailed results).