# PPM-CLIP: Probabilistic Prompt Modeling for Generalizable AI-Generated Image Detection

## Supplementary Material

## 6. Implementation Details of Patch Scoring

For a given RGB image, we first view it into non-overlapping patches, denoted as $\mathcal{I} = \{p_1, p_2, \ldots, p_n\}$, where $p_i \in \mathbb{R}^{M\times M\times 3}$.

Each patch is then transformed into the frequency domain using the Discrete Cosine Transform (DCT), resulting in corresponding frequency patches:

$$\mathcal{P}_f = \{p_1^{\text{dct}}, p_2^{\text{dct}}, \ldots, p_n^{\text{dct}}\}, \quad p_i^{\text{dct}} \in \mathbb{R}^{M\times M\times 3}.$$

To assess the frequency complexity of each patch, we employ a simple yet effective scoring strategy using $N_f$ distinct band-pass filters defined as:

$$F_{i,j}^{(k)} = \begin{cases} 1, & \text{if } \frac{2M}{N_f}\cdot k \le i+j < \frac{2M}{N_f}\cdot(k+1), \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Here, $F_{ij}^{(k)}$ denotes the weight at position $(i, j)$ in the $k$-th band-pass filter. These filters allow us to differentiate between patches with high and low frequency components. Next, for the $m$-th patch $p_m^{\text{dct}} \in \mathbb{R}^{M\times M\times 3}$, we compute its grade $G_m$ by applying the $N_f$ filters to the logarithm of the absolute DCT coefficients. The grading function is defined as:

$$G_m = \sum_{k=0}^{N_f-1} 2^k \cdot \sum_{c=0}^{2} \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} F_{i,j}^{(k)} \cdot \log\left(\left|p_m^{\text{dct}}(i,j,c)\right| + 1\right). \tag{15}$$

where $c$ denotes the channel index.

## 7. Derivation of the ELBO Objective

Let the training dataset be denoted as $\mathcal{D} = \{\mathbf{I}, \mathbf{Y}\}$, where $\mathbf{I}$ represents the input images and $\mathbf{Y}$ represents the corresponding ground-truth labels. To construct a probabilistic distribution over textual prompts, we model the image-generic, image-specific, and class-indicative text embeddings as latent random vectors $\Phi_g$, $\Phi_s$, and $\Phi_c$, respectively. For compactness, we denote the complete set of latent variables as $\Phi = \{\Phi_g, \Phi_s, \Phi_c\}$. These variables correspond to the distributions parameterized by the Prompt Flow Module described in Section 3.2 of the main text.

Following Bayes' theorem, the posterior distribution over the latent prompts given the dataset is:

$$p(\Phi|\mathcal{D}) = \frac{p(\mathcal{D}|\Phi)p(\Phi)}{p(\mathcal{D})}. \tag{16}$$

Since computing the marginal likelihood $p(\mathcal{D})$ is intractable, we adopt variational inference. We introduce a variational distribution $q_\gamma(\Phi|\mathcal{D})$, parameterized by $\gamma$, to approximate the true posterior.

**Independence Assumption and Factorization.** A core tenet of our PPM framework is the semantic decoupling of prompts. Accordingly, we assume independence among the latent components. The variational distribution is factorized as:

$$q_\gamma(\Phi|\mathcal{D}) \approx q_{\gamma_s}(\Phi_s|\mathbf{I}) \cdot q_{\gamma_g}(\Phi_g) \cdot q_{\gamma_c}(\Phi_c). \tag{17}$$

Crucially, while $\Phi_s$ is conditioned on the specific input image $\mathbf{I}$ , $\Phi_g$ and $\Phi_c$ are modeled as global latent variables shared across the dataset. Although $\Phi_g$ and $\Phi_c$ are not directly conditioned on individual input images, their parameters are optimized jointly with $\Phi_s$ to maximize the total evidence.

**ELBO Derivation.** By applying Jensen's inequality, we derive the Evidence Lower Bound (ELBO) on the log marginal likelihood:

$$\begin{aligned} \log p(\mathcal{D}) &= \log \int p(\mathcal{D}|\Phi)p(\Phi)d\Phi \\ &\ge \mathbb{E}_{q_\gamma(\Phi|\mathcal{D})}\left[\log p(\mathcal{D},\Phi) - \log q_\gamma(\Phi|\mathcal{D})\right] \\ &= -\mathcal{L}_e(\mathcal{D}). \end{aligned} \tag{18}$$

We minimize the negative ELBO, denoted as $\mathcal{L}_e(\mathcal{D})$. Using the chain rule $p(\mathcal{D}, \Phi) = p(\mathcal{D}|\Phi)p(\Phi)$, we decompose the objective:

$$\mathcal{L}_e(\mathcal{D}) = \mathbb{E}_{q_\gamma(\Phi|\mathcal{D})}\left[\log q_\gamma(\Phi|\mathcal{D}) - \log p(\Phi)\right] \tag{19}$$

$$- \mathbb{E}_{q_\gamma(\Phi|\mathcal{D})}[\log p(\mathcal{D}|\Phi)]. \tag{20}$$

**Connection to Loss Functions.** This decomposition reveals the theoretical foundation of our proposed loss functions:

- **KL Divergence (Eq. 19):** Due to the factorization assumption, the joint KL divergence splits into the sum of KL divergences for each component:

$$\mathcal{L}_{KL} = \mathcal{L}_{KL}(\Phi_g) + \mathcal{L}_{KL}(\Phi_s) + \mathcal{L}_{KL}(\Phi_c). \tag{21}$$

This justifies applying the flow-based KL regularization (Eq. 8 in the main text) to each distribution individually.

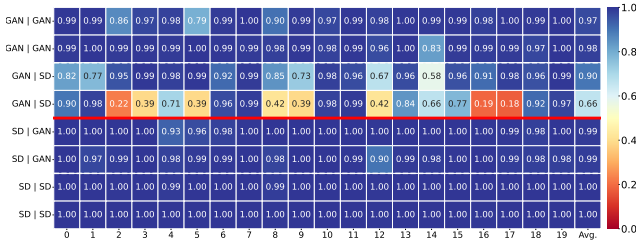


Figure 6. Heatmap of ensemble confidence scores demonstrating cross-generator robustness. The Y-axis denotes the domain shift scenarios ("Train Model | Test Source"). The X-axis represents the confidence scores produced by the $B \times N$ ($B = 2, N = 10$) individual generated prompt pairs (columns 0–19) alongside their final ensemble average (the "Avg" column).

- **Reconstruction (Eq. 20):** The term $-\log p(\mathcal{D}|\Phi)$ represents the negative log-likelihood of the data observation. In our framework, the prompt generation is conditioned on the visual feature $\mathbf{X}_{\text{cls}}$ extracted from $\mathbf{I}$. To ensure the image-specific latent $\Phi_s$ retains fine-grained visual fidelity, we maximize the likelihood of the visual features $p(\mathbf{X}_{\text{cls}}|\Phi_s)$. Assuming a Gaussian likelihood, minimizing this term is equivalent to the Mean Squared Error (MSE) loss:

$$-\log p(\mathbf{X}_{\text{cls}}|\Phi_s) \propto \|\mathbf{X}_{\text{cls}} - \text{Dec}(\Phi_s)\|_2^2, \tag{22}$$

  which corresponds to $\mathcal{L}_{rec}$ (Eq. 9 in the main text).

Consequently, maximizing the ELBO is equivalent to minimizing the weighted sum of $\mathcal{L}_{KL}$ and $\mathcal{L}_{rec}$. In our full objective (Eq. 13), these terms serve as a generative regularization to the discriminative classification task.

# 8. Robustness Evaluation.

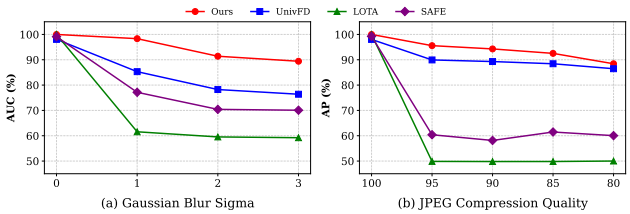


Figure 7. Robustness evaluation against common image Degradations. (a) AUC scores under different Gaussian blur sigma values. (b) AP scores under different JPEG compression qualities.

## 8.1. Insights into Cross-Generator Robustness.

To further elucidate the mechanism behind our model's superior cross-generator generalization, we visualize the confidence scores of individual generated prompts in Figure 6. The heatmap illustrates various training and testing scenarios, including challenging domain shifts such as training on GANs and evaluating on Stable Diffusion (denoted as "GAN | SD"). As observed in the heatmap, when the model encounters unseen artifacts from a novel generator, relying on a single static prompt is highly risky. This is evidenced by the high variance and occasional low confidence scores (e.g., values dropping to 0.18 or 0.22, shown in warm colors) among individual prompt hypotheses (columns 0–19). However, our probabilistic ensemble approach effectively neutralizes this unreliability. By aggregating the predictions across all $B \times N$ adaptive hypotheses, the final ensemble consensus (the "Avg" column) consistently maintains a high and stable confidence level. This visualization explicitly validates our core motivation: shifting from a single deterministic boundary to a probabilistic distribution of hypotheses is the key to mitigating generator-specific biases and achieving robust cross-domain detection.

## 8.2. Robustness against Image Degradations.

To assess the model's resilience to common image transmission and processing artifacts, we evaluate its robustness against Gaussian blur and JPEG compression across the GenImage benchmark. For the baseline comparisons, SAFE and LOTA utilize their officially provided models. As illustrated in Figure 7, our PPM-CLIP demonstrates exceptional resilience against visual degradations. Under increasing Gaussian blur (Figure 7a, up to a sigma of 3), our method maintains a robust AUC of approximately 90%, whereas competing methods (e.g., LOTA and SAFE) experience severe performance drops. Similarly, against JPEG compression (Figure 7b), our model sustains high AP scores across varying quality levels. While methods like SAFE and LOTA exhibit a sharp decline in performance even at mild compression levels (Quality 95), PPM-CLIP consistently outperforms all baselines, highlighting the superior robustness of our probabilistic reasoning approach against low-level visual distortions.