EVISLOT: EVIDENCE SLOT-GUIDED VISION-LANGUAGE LEARNING FOR GENERALIZABLE AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated images have become increasingly realistic and visually indistinguishable from real images, making generalizable AI-generated image detection a challenging task. Existing vision-language detectors usually rely on global visual representations and deterministic textual prompts, which may overlook localized high-frequency artifacts and provide limited flexibility under cross-generator distribution shifts. To address this issue, we propose **EviSlot**, an evidence slot-guided vision-language learning framework that jointly models global semantics, local forensic evidence, and diverse textual hypotheses. Specifically, we employ DINOv3 as the visual encoder to extract both CLS and patch-level features, and introduce a DCT-guided evidence mining module to identify frequency-abnormal regions with potential artifact cues. A learnable Evidence Slot module then aggregates local artifact cues from patch tokens with frequency-aware attention guidance and fuses them with the global representation for robust visual evidence modeling. On the textual side, frequency-aware probabilistic prompting generates multiple real/fake prototypes from shared, image-specific, and class-specific prompt components. The final prediction is obtained by image-text matching and prompt-ensemble inference. The proposed framework provides a compact, flexible, and robust detection paradigm for generalizable AI-generated image detection.

1 INTRODUCTION

AI-generated images have become highly realistic with the progress of GANs and diffusion models, raising risks in misinformation, copyright misuse, identity abuse, and digital forensics (Mahara & Rishe, 2026; Deng et al., 2025; van Lierop et al., 2026). Reliable detection is therefore important but difficult, as detectors trained on known generators often degrade on unseen generators, diverse scenes, post-processing, and high-resolution in-the-wild images. Large-scale benchmarks further highlight the need for broader generator coverage and more diverse image sources (Zhu et al., 2023; Hong et al., 2025; Yan et al., 2024). Early methods exploit low-level artifacts such as GAN fingerprints, spatial traces, manipulation cues, and abnormal frequency patterns (Yu et al., 2019; Zhang et al., 2019; Wang et al., 2020; Dang et al., 2020; Frank et al., 2020; Schwarz et al., 2021). Recent diffusion and text-to-image detectors use reconstruction errors or benchmark-driven settings (Wang et al., 2023; Chen et al., 2024; Zhu et al., 2023; Sha et al., 2023), but they may still overfit generator-specific artifacts and suffer from domain shifts.

Vision-language models offer transferable representations for universal fake image detection. CLIP-based detectors improve cross-generator generalization (Radford et al., 2021; Ojha et al., 2023; Cozzolino et al., 2024), while prompt-learning methods introduce learnable, multi-modal, category-common, sample-level, probabilistic, or image-adaptive prompts (Zhou et al., 2022; Khattak et al., 2023; Tan et al., 2025; Lin et al., 2025b; Wang et al., 2026; Li et al., 2026). However, many methods rely on global features or deterministic prompts, which limits localized forensic evidence modeling. Recent works show that local patches, DCT-selected regions, texture cues, artifact purification, high-resolution processing, and frequency features can improve robustness (Yan et al., 2025; Lin et al., 2025a; Meng et al., 2024; Mu et al., 2026; Tan et al., 2024a;b; Karageorgiou et al., 2025). These findings motivate local frequency evidence mining with flexible vision-language matching.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

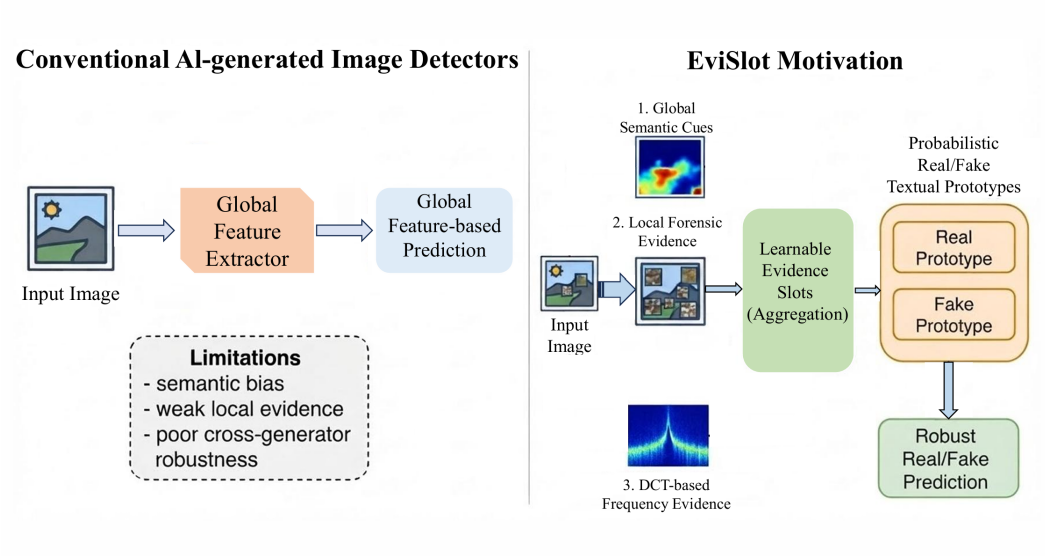


Figure 1: Motivation. EviSlot aggregates global semantic cues, local forensic evidence, and DCT-based frequency evidence with learnable evidence slots for robust real/fake prediction.

To this end, we propose **EviSlot**, an evidence slot-guided vision-language learning framework for generalizable AI-generated image detection. The core idea of EviSlot is to make the detector reason from explicit evidence rather than relying only on a single global representation. Specifically, EviSlot first extracts global CLS features and patch-level tokens with a DINOv3 visual encoder, so that image-level semantics and local visual details can be preserved simultaneously. Since AI-generated images often contain subtle local artifacts that are difficult to capture from global features alone, we further introduce DCT-guided evidence mining to estimate patch-level frequency abnormality. These frequency scores serve as soft guidance for learnable Evidence Slots, which adaptively attend to suspicious patches and aggregate compact local forensic representations.

On the textual side, EviSlot avoids using a single fixed real/fake prompt. Instead, it constructs multiple real/fake textual prototypes through frequency-aware probabilistic prompting, where shared, image-specific, and class-specific prompt components jointly model diverse and complementary decision hypotheses. The final prediction is obtained by matching the fused visual evidence with the prompt ensemble in the CLIP embedding space. In this way, EviSlot unifies global semantic understanding, localized frequency-aware evidence aggregation, and flexible textual hypothesis modeling, leading to more robust detection under challenging cross-generator distribution shifts.

Our major contributions are summarized as follows:

- We propose **EviSlot**, an evidence slot-guided vision-language learning framework for generalizable AI-generated image detection, which jointly models global semantics, local forensic evidence, and probabilistic textual hypotheses.
- We introduce a DCT-guided Evidence Slot module that uses patch-level frequency abnormality as soft attention guidance, enabling the model to mine and aggregate suspicious local artifacts from DINOv3 patch tokens.
- We design a frequency-aware probabilistic prompting strategy that constructs multiple real/fake textual prototypes from shared, image-specific, and class-specific prompt components, improving the flexibility of image-text matching.
- Extensive experiments on GenImage and AIGI-Now demonstrate the effectiveness of EviSlot. Our method achieves strong performance on both benchmarks, with particularly notable cross-domain generalization on AIGI-Now, showing its robustness against unseen generator distributions.

2 Related Work

AI-generated image detection. AI-generated image detection has been widely studied in digital forensics. Early works reveal GAN fingerprints and spatial artifacts (Yu et al., 2019; Zhang et al., 2019), which can be captured by CNN detectors with proper data and augmentation (Wang et al., 2020); related studies also examine face manipulation detection (Dang et al., 2020). Frequency-domain methods identify abnormal spectral patterns (Frank et al., 2020; Schwarz et al., 2021), while recent frequency-aware detectors exploit spectral cues and up-sampling artifacts for better generalization (Tan et al., 2024a;b; Karageorgiou et al., 2025). With diffusion and text-to-image models, detection becomes harder: DIRE uses diffusion reconstruction errors (Wang et al., 2023), DRCT introduces diffusion reconstruction contrastive training (Chen et al., 2024), and De-Fake studies text-to-image detection and attribution (Sha et al., 2023). Large-scale benchmarks and surveys, including GenImage, WildFake, DF40, and recent reviews, stress evaluation across diverse generators and real-world scenarios (Zhu et al., 2023; Hong et al., 2025; Yan et al., 2024; Mahara & Rishe, 2026; Deng et al., 2025; van Lierop et al., 2026). Local evidence further complements global semantics, as shown by AIDE (Yan et al., 2025), dual-path detection (Lin et al., 2025a), artifact purification (Meng et al., 2024), and high-resolution detection (Mu et al., 2026).

Vision-language detection. Vision-language models offer a promising route to generalizable detection. CLIP learns transferable image-text representations (Radford et al., 2021), and CLIP-based detectors improve cross-generator generalization through shared embedding spaces (Ojha et al., 2023; Cozzolino et al., 2024). Prompt learning and reprogramming further adapt vision-language models: CoOp and MaPLe use learnable and multi-modal prompts (Zhou et al., 2022; Khattak et al., 2023); C2P-CLIP introduces category-common prompts (Tan et al., 2025); visual-language reprogramming adapts large VLMs for deepfake detection (Lin et al., 2025b); PPM-CLIP models probabilistic prompts (Wang et al., 2026); and image-adaptive prompt learning builds input-conditioned prompts (Li et al., 2026). Self-supervised visual foundation models such as DINOv2 also provide strong transferable features (Oquab et al., 2024). Unlike methods mainly relying on global image-text matching or deterministic prompts, **EviSlot** uses DCT-guided Evidence Slots to aggregate local frequency evidence and combines them with frequency-aware probabilistic prompting for robust cross-generator detection.

3 Methodology

We propose **EviSlot**, an evidence slot-guided vision-language framework for generalizable AI-generated image detection. The motivation is that synthetic images may contain both global semantic inconsistencies and localized forensic artifacts. While global image representations are effective for capturing high-level semantics, subtle generation traces often appear in local frequency-abnormal regions. Therefore, our framework combines a DINOv3 visual encoder, a DCT-guided Evidence Slot module, and a probabilistic prompting module. The final prediction is obtained by matching image and text embeddings in the CLIP embedding space.

Given an input image $\boldsymbol{x}$, the model predicts whether it belongs to the real or fake class. The visual branch extracts a global CLS feature and patch-level tokens. The local evidence branch computes DCT-based frequency scores and aggregates suspicious patch cues through Evidence Slots. The text branch constructs multiple real/fake textual prototypes through probabilistic prompt learning. These components are jointly optimized for robust cross-generator detection.The overall framework is shown in Fig. 2.

3.1 Visual Representation with DINOv3

We adopt DINOv3 as the visual encoder due to its strong transferable patch-level representation. Given an image $\boldsymbol{x}$, the encoder produces a global CLS token and a sequence of patch tokens:

$$\boldsymbol{g}, \boldsymbol{P} = f_{vis}(\boldsymbol{x}), \quad \boldsymbol{P} = \{\boldsymbol{p}_l\}_{l=1}^{L}, \tag{1}$$

where $\boldsymbol{g}$ denotes the global representation and $\boldsymbol{P}$ denotes L local patch features. The global token captures image-level semantic information, while patch tokens preserve local visual details that are useful for detecting generation artifacts.

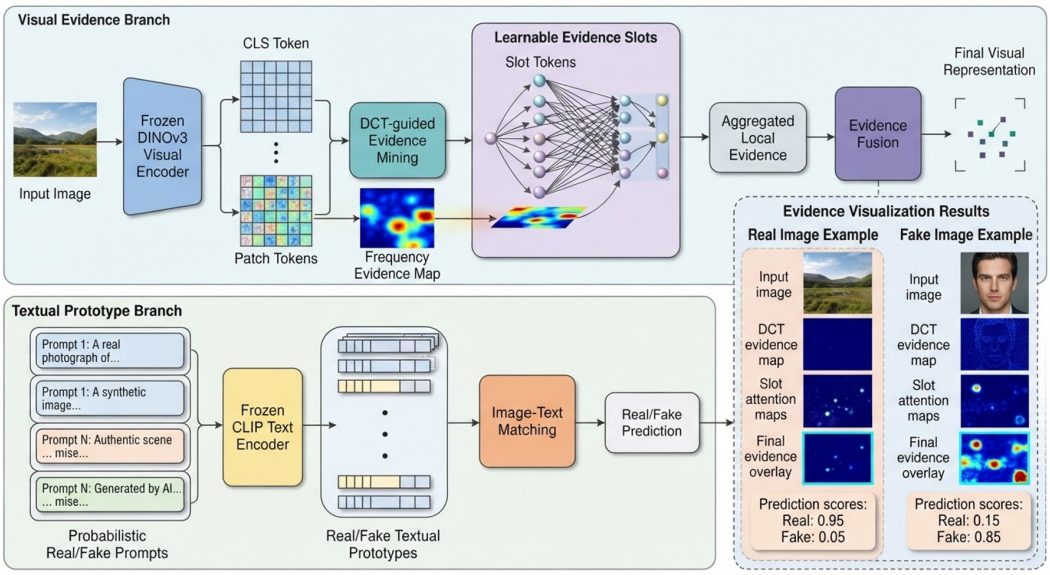


Figure 2: Overview of the proposed EviSlot framework. EviSlot combines DINOv3-based visual evidence mining with CLIP-based textual prototypes, and uses learnable evidence slots to aggregate local artifacts for real/fake prediction.

To reduce training cost, most pretrained parameters are frozen, and lightweight trainable modules are used for adaptation. The final visual representation is later projected into the CLIP embedding space and matched with textual prototypes.

3.2 DCT-Guided Local Evidence Mining

AI-generated images often exhibit abnormal local frequency patterns. To capture such artifacts, we compute a DCT-based evidence score for each image patch. The input image is divided into non-overlapping patches aligned with DINOv3 patch tokens. For each patch $\boldsymbol{x}_l$, we compute its band-weighted frequency energy:

$$s_l = \sum_{g=1}^{G} \omega_g \sum_{c,u,v} \boldsymbol{M}_{u,v}^{(g)} \log\left(1 + |\mathrm{DCT}(\boldsymbol{x}_l)_{c,u,v}|\right), \tag{2}$$

where $\boldsymbol{M}^{(g)}$ is the mask of the g-th frequency band, and ω_g is the corresponding band weight. Higher-frequency bands are assigned larger weights, making the score more sensitive to local high-frequency artifacts.

The scores are normalized within each image and used as soft frequency guidance for Evidence Slots. In addition, the highest-scoring patches and the remaining patches are used to form a patch-level contrastive regularization. The patch with the highest DCT score is treated as the anchor, encouraging high-evidence patches to form compact local artifact representations while separating them from low-evidence patches.

3.3 Evidence Slot Aggregation

Directly using selected patches may be unstable because local frequency responses can be noisy. To obtain a compact and robust local representation, we introduce learnable Evidence Slots. These slots act as adaptive queries that attend to DINOv3 patch tokens and aggregate manipulation-related cues.

Let $\boldsymbol{E}^0 \in \mathbb{R}^{S\times d}$ denote S learnable evidence slots. Cross-attention is performed between slots and patch tokens. We inject normalized DCT scores as a soft attention bias:

$$\boldsymbol{A} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}} + \eta \cdot \mathrm{Norm}(\boldsymbol{s})\right), \tag{3}$$

where $\boldsymbol{Q}$ is projected from Evidence Slots, $\boldsymbol{K}$ is projected from patch tokens, $\boldsymbol{s}$ denotes the DCT evidence scores, and η is a learnable guidance scale. The attended patch features are then updated through residual projection and feed-forward layers to obtain slot features.

The slot features are pooled into a compact local evidence vector and fused with the global CLS feature:

$$\boldsymbol{e} = \sum_{r=1}^{S} \pi_r \boldsymbol{E}_r, \quad \boldsymbol{v} = \boldsymbol{g} + \lambda_e \boldsymbol{e}, \tag{4}$$

where π_r denotes the learnable pooling weight of the r-th slot, and λ_e controls the contribution of local evidence.

Following the implementation, we further apply a lightweight residual MLP adapter after evidence fusion:

$$\widetilde{\boldsymbol{v}} = \boldsymbol{v} + \alpha\, \mathrm{MLP}(\boldsymbol{v}), \quad \boldsymbol{z}^{img} = W_{img}\, \mathrm{LN}\left(\widetilde{\boldsymbol{v}}\right). \tag{5}$$

Here, $\boldsymbol{z}^{img}$ is the final image embedding in the CLIP space. To avoid all slots focusing on the same region, we apply a diversity regularization on the slot attention maps, encouraging different slots to capture complementary local evidence.

3.4 Probabilistic Prompting

Instead of relying on a single deterministic textual prompt, we construct multiple real/fake textual prototypes through probabilistic prompting. The prompt repository contains three learnable components: shared context tokens, private tokens, and class-specific tokens. For the real and fake classes, the base prompts are defined as:

$$\boldsymbol{C}_{real} = [\boldsymbol{P}^{sha} \parallel \boldsymbol{P}^{pri} \parallel \boldsymbol{P}^{real}], \quad \boldsymbol{C}_{fake} = [\boldsymbol{P}^{sha} \parallel \boldsymbol{P}^{pri} \parallel \boldsymbol{P}^{fake}]. \tag{6}$$

Here, $\boldsymbol{P}^{sha}$ models class-shared textual context, $\boldsymbol{P}^{pri}$ provides image-adaptive prompt capacity, and $\boldsymbol{P}^{real}$ and $\boldsymbol{P}^{fake}$ encode class-specific semantics.

We further generate dynamic prompt biases with probabilistic flow modules. Given the image embedding $\boldsymbol{z}^{img}$, an image-conditioned flow produces the private bias $\boldsymbol{b}^{pri}$. In parallel, learnable probabilistic flows generate the shared bias $\boldsymbol{b}^{sha}$ and the class-specific biases $\boldsymbol{b}^{real}$ and $\boldsymbol{b}^{fake}$. These biases are added to the corresponding prompt tokens:

$$\widetilde{\boldsymbol{C}}_c = [\boldsymbol{P}^{sha} + \boldsymbol{b}^{sha} \parallel \boldsymbol{P}^{pri} + \boldsymbol{b}^{pri} \parallel \boldsymbol{P}^{c} + \boldsymbol{b}^{c}], \quad c \in \{real, fake\}. \tag{7}$$

The biased prompts are inserted into the CLIP text encoder between the start and end tokens to obtain textual prototypes:

$$\boldsymbol{z}_c^{txt} = f_{txt}\left(\widetilde{\boldsymbol{C}}_c\right). \tag{8}$$

During training, one prompt hypothesis is sampled for efficient optimization. During inference, multiple prompt samples are generated for each class, producing an ensemble of real/fake textual prototypes. This probabilistic design improves robustness by allowing the text branch to represent diverse decision hypotheses instead of a fixed prompt template.

3.5 Image-Text Matching and Optimization

The image embedding and textual prototypes are first mapped and normalized in the CLIP embedding space. For class c, the image-text matching logit is computed by cosine similarity with a learnable temperature:

$$\ell_c = \exp(\tau) \cdot \left(\hat{\boldsymbol{z}}^{img}\right)^{\top} \hat{\boldsymbol{z}}_c^{txt}, \tag{9}$$

where τ is the learnable temperature parameter, and $\hat{\boldsymbol{z}}$ denotes the normalized embedding.

During inference, all sampled real/fake prompt pairs are evaluated, and their probabilities are averaged:

$$\hat{\boldsymbol{p}} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{softmax}\left([\ell_{real}^n, \ell_{fake}^n]\right). \tag{10}$$

Although the implementation contains several auxiliary regularization terms, we group them into two compact objectives for clarity. The full training objective is written as:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{evi}\mathcal{L}_{evi} + \lambda_{prompt}\mathcal{L}_{prompt}. \tag{11}$$

Here, $\mathcal{L}_{cls}$ is the binary classification loss. The evidence regularization term $\mathcal{L}_{evi}$ summarizes DCT-guided patch contrastive learning and Evidence Slot diversity regularization, encouraging the visual branch to focus on discriminative local forensic cues. The prompt regularization term $\mathcal{L}_{prompt}$ summarizes the reconstruction, KL-divergence, and prompt orthogonality constraints used in probabilistic prompt learning. This grouped formulation keeps the objective concise while remaining consistent with the implementation.

4 Experiments

4.1 Settings

Datasets. Following the common cross-generator evaluation protocol in AI-generated image detection, we restrict the accessible training data to a single source generator domain and evaluate the detector on both in-domain and out-of-domain data. Specifically, EviSlot is trained on the **train** split of GenImage-SDv1.4 Zhu et al. (2023). The training set contains real images from the `nature` category and fake images from the `ai` category. For in-domain evaluation, we use the **val** split of GenImage-SDv1.4.

To evaluate cross-generator generalization, we further test the model on the **pixel** subset of AIGI-Now. This subset contains images from nine recent generators, including Nano, flux-dev, flux-kera, flux-kontext, flux-pro, gpt4o, jimeng, keling, and minimax. For all datasets, natural images are labeled as real, while AI-generated images are labeled as fake.

Evaluation metric. Following previous AI-generated image detection works Wang et al. (2020); Ojha et al. (2023); Wang et al. (2026), we report accuracy (Acc) and average precision (AP) as the main evaluation metrics. Acc measures the overall binary classification performance, while AP evaluates the ranking quality of predicted fake probabilities. In addition, we report real-class accuracy (r_acc) and fake-class accuracy (f_acc) to evaluate the performance on real and fake images separately. For AIGI-Now, we report both per-generator accuracy and the mean performance over all generators.

Implementation details. EviSlot adopts DINOv3 ViT-L/16 Siméoni et al. (2025) as the visual encoder to extract global and patch-level image representations. The text branch uses a ViT-L/14 backbone. The input images are resized to 256×256, and ImageNet normalization is applied. The DCT-guided evidence mining module is used to compute patch-level frequency evidence scores, which guide the Evidence Slot aggregation module.

We use 4 Evidence Slots, 8 attention heads in the evidence module, an evidence dropout rate of 0.1, and set the evidence fusion weight to 0.2. For parameter-efficient tuning, LoRA is applied to both encoders with rank 2, alpha 2, and q, k, v, o projections. The LoRA position is set to half-up. The prompt module uses 2 prompts, with a class length of 10, private length of 7, and shared length of 3. The sample number is set to 10.

The model is trained with AdamW optimizer. The batch size is set to 128, the learning rate is 1×10^{-4}, and the weight decay is 1×10^{-4}. The number of data loading workers is set to 12. The loss weights are set as $\lambda_{\text{contra}} = 0.5$, $\lambda_{\text{rec}} = 0.5$, $\lambda_{\text{ort}} = 1.0$, $\lambda_{\text{kl}} = 0.001$, and $\lambda_{\text{slot}} = 0.001$. We train the model for up to 10 epochs and apply early stopping with a patience of 3 epochs.

4.2 Main Results

Table 1 reports the accuracy comparison on GenImage-SDv1.4 and AIGI-Now pixel subsets. GenImage-SDv1.4 is used for in-domain validation, while AIGI-Now is used to evaluate cross-generator generalization across recent image generators. Although several baselines obtain strong in-domain performance on GenImage-SDv1.4, their accuracy drops significantly on AIGI-Now, showing the difficulty of generalizing to unseen generators.

On AIGI-Now, EviSlot achieves the best mean accuracy of 88.62%, outperforming PPM-CLIP by 12.85 percentage points, NPR by 16.65 percentage points, and LGrad by 25.11 percentage points.

Table 1: Accuracy comparison on GenImage-SDv1.4 and AIGI-Now pixel subsets. The best result in each column is highlighted in bold.

Method	GenImage	Nano	flux-dev	flux-kera	flux-kontext	flux-pro	gpt4o	jimeng	keling	minimax	AIGI-Now Mean
ResNet-50	82.76	49.40	52.30	50.15	52.60	53.15	50.95	49.65	49.50	49.80	50.83
EfficientNet-B0	89.47	48.20	51.25	52.35	54.65	57.65	55.15	47.25	47.65	48.25	51.38
ViT-B/16	86.41	47.65	50.75	50.15	53.05	55.35	54.15	48.90	51.40	48.70	51.12
LGrad	89.46	62.05	68.90	60.40	63.15	58.10	73.65	52.80	70.45	61.55	63.51
NPR	97.53	85.65	86.95	59.00	72.40	47.60	92.65	46.75	91.80	64.95	71.97
UnivFD	63.76	50.95	50.20	50.95	57.50	51.85	53.85	50.65	51.40	50.45	51.98
PPM-CLIP	**99.76**	**99.40**	**99.50**	51.20	82.95	50.00	**99.45**	50.00	**98.85**	50.60	75.77
EviSlot	98.51	94.00	93.45	**94.00**	**87.00**	**83.15**	94.25	**63.70**	95.00	**93.05**	**88.62**

This indicates that EviSlot has stronger overall generalization ability under the cross-generator setting. In particular, EviSlot achieves the best results on five AIGI-Now subsets, including flux-kera, flux-kontext, flux-pro, jimeng, and minimax.

The improvement is especially clear on challenging generators. On flux-kera, EviSlot reaches 94.00% accuracy, improving over the best previous result by 33.60 percentage points. On flux-pro, EviSlot achieves 83.15%, outperforming the best baseline by 25.05 percentage points. On minimax, EviSlot obtains 93.05%, which is 28.10 percentage points higher than the best previous result. EviSlot also improves the accuracy on jimeng to 63.70%, while all baselines remain around 46.75%–52.80%.

For some subsets such as Nano, flux-dev, gpt4o, and keling, PPM-CLIP achieves the highest single-generator accuracy. However, its performance drops sharply on flux-kera, flux-pro, jimeng, and minimax, leading to a lower mean accuracy of 75.77%. In contrast, EviSlot maintains more balanced performance across different generators. These results suggest that DINOv3-based visual representations, DCT-guided evidence mining, and Evidence Slot aggregation help EviSlot capture more transferable local forensic cues instead of overfitting to generator-specific artifacts.

Moreover, the per-generator results show that high in-domain accuracy does not necessarily lead to stable cross-generator detection. For example, PPM-CLIP obtains the highest accuracy on GenImage-SDv1.4, Nano, flux-dev, gpt4o, and keling, but its performance is close to random guessing on flux-kera, flux-pro, jimeng, and minimax. This indicates that some detectors may learn generator-dependent cues that transfer well to only a subset of target domains. By contrast, EviSlot does not achieve the best result on every single generator, but it consistently avoids severe performance collapse and achieves the best mean accuracy. This more stable behavior is important for practical AI-generated image detection, where the target generator is often unknown.

4.3 Qualitative Visualization

Fig. 3 visualizes the evidence mining process of EviSlot. The DCT evidence map provides a frequency-domain prior that highlights suspicious local regions. Guided by this prior, different Evidence Slots attend to complementary local patterns, including object boundaries, texture-rich regions, and high-frequency areas. The final evidence overlay aggregates slot-level responses and shows the localized cues used for prediction.

For fake images, the evidence responses are stronger and more concentrated around regions with potential forensic inconsistencies. For real images, although some local responses may still appear, the predicted fake probability remains lower than the real probability. This visualization demonstrates that EviSlot does not rely only on global semantic representations, but also mines interpretable local evidence for real/fake decisions.

The visualization further supports the motivation of using Evidence Slots rather than a single global representation. Different slots often focus on different regions, suggesting that they can capture complementary local cues instead of repeatedly attending to the same image area. Such localized evidence is useful for cross-generator detection because unseen generators may produce artifacts in different semantic regions or frequency bands.

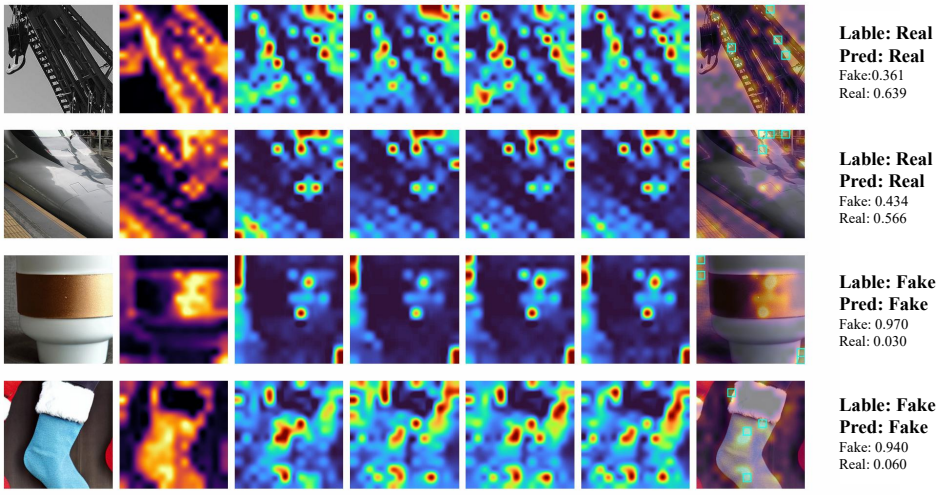


Figure 3: Evidence mining visualization of EviSlot. For each image, we show the input image, DCT-based evidence map, four Evidence Slot attention maps, the final evidence overlay, and the prediction scores. Brighter regions indicate stronger evidence responses, and cyan boxes mark the top-attended patch regions.

4.4 Ablation Study

Table 2: Ablation study of EviSlot-CLIP on GenImage and AIGI-Now. We report Acc, AP, real-class accuracy (r_acc), and fake-class accuracy (f_acc).

Variant	GenImage				AIGI-Now			
	Acc	AP	r_acc	f_acc	Acc	AP	r_acc	f_acc
Full model	**97.58**	99.88	**95.45**	99.72	**88.62**	93.86	89.79	87.46
w/o Evidence Slots	92.89	99.85	85.82	**99.97**	85.26	91.46	86.11	84.41
w/o DCT Evidence	77.98	99.69	55.97	100.00	80.97	**95.85**	63.22	**98.71**
w/o Evidence Fusion	94.56	99.84	89.15	**99.97**	86.21	92.32	**90.72**	81.69

Table 2 reports the ablation results of EviSlot on GenImage-SDv1.4 and AIGI-Now. We compare the full model with three variants: removing Evidence Slots, removing DCT Evidence, and removing Evidence Fusion. The full model achieves the best overall balance, with 97.58% Acc and 99.88% AP on GenImage-SDv1.4, and 88.62% Acc and 93.86% AP on AIGI-Now.

Removing Evidence Slots reduces the GenImage-SDv1.4 accuracy from 97.58% to 92.89%, and the AIGI-Now accuracy from 88.62% to 85.26%. The real-class accuracy also drops from 95.45% to 85.82% on GenImage-SDv1.4 and from 89.79% to 86.11% on AIGI-Now. This shows that explicit Evidence Slot aggregation helps the model mine useful local evidence and improves the robustness of real/fake discrimination.

Removing DCT Evidence causes the largest degradation. The accuracy drops by 19.60 percentage points on GenImage-SDv1.4 and by 7.65 percentage points on AIGI-Now. More importantly, the real-class accuracy decreases sharply from 95.45% to 55.97% on GenImage-SDv1.4 and from 89.79% to 63.22% on AIGI-Now, while the fake-class accuracy remains very high. This indicates that without DCT-guided frequency evidence, the model becomes biased toward predicting images as fake, leading to poor balance between real and fake classes.

Removing Evidence Fusion also weakens the model. The accuracy decreases from 97.58% to 94.56% on GenImage-SDv1.4 and from 88.62% to 86.21% on AIGI-Now. On AIGI-Now, the

fake-class accuracy drops from 87.46% to 81.69%, showing that evidence fusion is important for integrating global semantic representations and local forensic evidence. Overall, these ablation results demonstrate that DCT Evidence, Evidence Slots, and Evidence Fusion all contribute to the final performance, with DCT Evidence being the most critical component for balanced and robust detection.

4.5 More Analysis

We further analyze two key hyperparameters in EviSlot: the evidence fusion weight and the number of evidence slots, as shown in Fig. 4. For the evidence fusion weight, smaller values are more stable, while an overly large weight causes a clear performance drop under fixed-threshold prediction. On AIGI-Now, increasing the fusion weight from 0.1 to 0.2 improves the accuracy from 83.07% to 88.62%, showing that moderate evidence fusion helps transfer local forensic cues to unseen generators. However, when the weight is further increased to 0.3 or 0.5, the AIGI-Now accuracy drops to 50.00%, although AP remains relatively high. This indicates that excessive evidence fusion can shift the prediction scores and make the fixed threshold poorly calibrated. Therefore, 0.2 provides the best overall balance and is used as the default evidence fusion weight.

For the number of Evidence Slots, using 4 slots gives the best overall performance. On AIGI-Now, increasing the number of slots from 2 to 4 improves the accuracy from 84.94% to 88.62%, suggesting that more slots can capture more diverse local evidence patterns. When the number of slots is increased to 6, the accuracy decreases to 85.89%, which may be caused by redundant or unstable slot representations. Using 8 slots recovers the accuracy to 88.31%, but it is still slightly lower than using 4 slots. These results suggest that a moderate number of slots is sufficient for robust evidence aggregation, while too few slots limit evidence diversity and too many slots may introduce unnecessary redundancy.

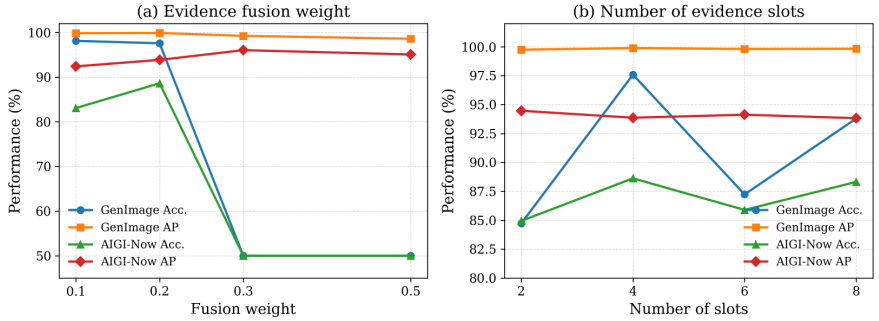


Figure 4: Parameter analysis of EviSlot.

5 Conclusion

This paper presented EviSlot, an evidence slot-guided vision-language framework for generalizable AI-generated image detection. Unlike CLIP-based detectors that mainly rely on global features and fixed prompts, EviSlot integrates global semantics, DCT-guided local frequency evidence, and probabilistic textual prototypes in a unified image-text matching framework. It uses DINOv3 features, mines frequency-abnormal patches, aggregates them with learnable Evidence Slots, and performs frequency-aware probabilistic prompting for stable real/fake prediction. Experiments on GenImage and AIGI-Now show strong in-domain and out-of-domain performance, while ablation studies confirm the effectiveness of the main components. Further analysis shows that moderate evidence fusion and a compact number of slots are important for maintaining both accuracy and robustness. These results indicate that explicit local evidence modeling can effectively complement global vision-language representations under cross-generator shifts. Overall, EviSlot provides a compact and robust paradigm for generalizable AI-generated image detection.

AI USE STATEMENT

AI-based assistants were used during the preparation of this work, limited to editorial, organisational, and routine programming support. Specifically, they were used for language editing, improving clarity, arranging material, refining presentation order, identifying redundant content, and assisting with code cleaning or routine refactoring. AI systems were not used to formulate the research questions, develop the analytical model, design the experiments, construct the formalisation, interpret the results, or generate, select, or modify any experimental result. All reported numerical results were computed from fixed experimental artefacts using the corresponding scripts and are reproducible from them. The authors reviewed and verified all AI-assisted text and code, and take full responsibility for the final manuscript and all results reported in it.

REFERENCES

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *ICML*, volume 2, pp. 6, 2024.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4356–4366. IEEE, 2024.

Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5780–5789. IEEE, 2020.

Jingyi Deng, Chenhao Lin, Zhengyu Zhao, Shuai Liu, Zhe Peng, Qian Wang, and Chao Shen. A survey of defenses against ai-generated visual media: Detection, disruption, and authentication. *ACM Computing Surveys*, 58(5):1–35, 2025.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pp. 3247–3258. PMLR, 2020.

Yan Hong, Jianming Feng, Haoxing Chen, Jun Lan, Huijia Zhu, Weiqiang Wang, and Jianfu Zhang. Wildfake: A large-scale and hierarchical dataset for ai-generated images detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3500–3508, 2025.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution ai-generated image detection by spectral learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18706–18717. IEEE, 2025.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122. IEEE, 2023.

Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable ai-generated image detection via image-adaptive prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21262–21272, 2026.

Jiapeng Lin, Qihao Zhou, and Jie Li. Dual-path ai-generated image detection: Leveraging texture-rich and texture-poor patches with global semantic features. *Digital Signal Processing*, pp. 105840, 2025a.

Kaiqing Lin, Yuzhen Lin, Weixiang Li, Taiping Yao, and Bin Li. Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5262–5270, 2025b.

Arpan Mahara and Naphtali Rishe. Methods and trends in detecting ai-generated images: A comprehensive review. *Computer Science Review*, 60:100908, 2026.

Zheling Meng, Bo Peng, Jing Dong, Tieniu Tan, and Haonan Cheng. Artifact feature purification for cross-domain detection of ai-generated images. *Computer Vision and Image Understanding*, 247: 104078, 2024.

Lianrui Mu, Haoji Hu, Xingze Zou, Jianhong Bai, and Jiaqi Hu. No pixel left behind: A detail-preserving architecture for robust high-resolution ai-generated image detection. In *International Conference on Learning Representations*, volume 2026, pp. 80332–80355, 2026.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489. IEEE, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*, 34:18126–18136, 2021.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3418–3432, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28130–28139, 2024b.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

Stijn van Lierop, Ivona Najdenkoska, Marcel Worring, and Zeno Geradts. Interpol review of detection of ai-generated image and video deepfakes, 2022-2025. *Forensic Science International: Synergy*, 13:100704, 2026.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8692–8701. IEEE, 2020.

Xinyuan Wang, Yingxin Lai, Zhiming Luo, and Zhihui Liu. Ppm-clip: Probabilistic prompt modeling for generalizable ai-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21316–21325, 2026.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22388–22398. IEEE, 2023.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37:29387–29434, 2024.

Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7556–7566, 2019.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE, 2019.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36:77771–77782, 2023.