Unified Forensics Preference Learning for Generalizable Synthetic Image Detection

Laixin Zhang[1,*] **Shuaibo Li**[2,*,†] **Wei Ma**[1,‡]
Jianyu Lai[3] **Lei Zhu**[3,4] **Hongbin Zha**[5,6]
[1]Beijing University of Technology [2]The Chinese University of Hong Kong
[3]The Hong Kong University of Science and Technology (Guangzhou)
[4]The Hong Kong University of Science and Technology
[5]Anqing Normal University [6]Peking University

Abstract

Despite rapid progress in synthetic image detection, detectors that perform well on known generators often become brittle when the source model, post-processing pipeline, or scene composition changes. A key reason is that standard binary supervision teaches models to separate training distributions rather than to identify authenticity-relevant evidence: low-level artifacts become shortcuts, while higher-level semantic implausibilities remain weakly grounded. To address this limitation, we propose **UniFPL**, a unified forensics preference learning framework that converts multimodal large language model (MLLM) forensic knowledge into image-specific preference supervision. UniFPL constructs forensic preference data from both controlled artifact synthesis and observed artifact harvesting, then learns from list-level evidence rankings that encode which visual-textual cues are more diagnostic of synthetic manipulation. On top of a CLIP-based discriminative detector, UniFPL jointly optimizes authenticity classification, artifacts-aware preference alignment, and semantic structure regularization, allowing the model to inherit contextual forensic reasoning while retaining the stable and efficient inference of a discriminative encoder. Across eleven curated and in-the-wild benchmarks, UniFPL achieves an average balanced accuracy of 93.4% and improves the worst-case accuracy from 81.4% to 86.1% over the strongest baseline, demonstrating stronger cross-generator generalization and robustness to real-world image degradations.

1 Introduction

Recent advances in generative models [31, 70, 79] have substantially improved the realism and diversity of synthetic images, making high-fidelity content creation increasingly accessible. While these advances enable a wide range of creative and practical applications, they also facilitate sophisticated visual forgeries and misinformation [68, 36, 57], threatening the authenticity and trustworthiness of digital media. As generative techniques continue to evolve, the forensic traces of image synthesis become increasingly subtle and diverse, making reliable AI-generated image detection both more challenging and more critical in visual forensics.

Research on synthetic image detection has made notable progress in recent years [33, 42, 34]. Early detectors relied on frequency-domain statistics [13] and local pixel patterns [59, 32], yet these cues often exhibit limited transferability across different generation methods [41]. Recent

*Equal contribution.
†Project lead.
‡Wei Ma (mawei@bjut.edu.cn) is the corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

studies have increasingly adopted large-scale pretrained foundation models [47, 58], particularly CLIP [52], as backbones for synthetic image detection. For example, UnivFD [47] employs pretrained CLIP representations for authenticity discrimination, achieving strong generalization to unseen generators. Benefiting from large-scale vision-language pretraining, CLIP provides transferable visual representations with broad semantic coverage [47], making it a strong foundation for generalizable synthetic image detection.

Despite these advances, existing detectors remain largely dependent on coarse-grained binary authenticity supervision. By emphasizing the separation of authentic and synthetic training distributions without explicitly supervising the underlying forensic evidence, such supervision may encourage reliance on generator-specific artifacts as shortcuts [78]. Meanwhile, task-specific fine-tuning can distort the rich semantic representations acquired during pretraining, leading to catastrophic forgetting and weakened generalization to unseen generators [63, 71]. As a result, a detector may become adept at identifying familiar forgery patterns while losing even basic visual knowledge, such as what a penguin looks like. Moreover, although CLIP-like representations capture diverse visual concepts [1], they remain limited in spatial and compositional reasoning [22, 61, 75], concept and attribute binding [5, 26], and contextual plausibility [24, 35]. For instance, a detector may recognize both a penguin and a desert, yet fail to understand why a penguin appearing in a hot desert is contextually implausible, let alone interpret this inconsistency as potential evidence of image synthesis. These limitations highlight a fundamental challenge: generalizable synthetic image detection requires not only preserving pretrained semantic knowledge but also establishing a meaningful connection between semantic plausibility and image authenticity.

Multimodal large language models (MLLMs) offer a promising direction for understanding contextual inconsistencies [50, 51, 66]. By combining visual perception with language reasoning, MLLMs demonstrate strong open-ended visual understanding and can assess semantic relationships and contextual plausibility [38, 2]. However, such contextual understanding does not necessarily translate into reliable authenticity discrimination. Since their training primarily emphasizes generative objectives, general-purpose MLLMs often underperform contrastively trained models such as CLIP on discriminative visual tasks [60, 80], with similar limitations observed in image forensics [23, 19]. To bridge this gap, MLLM-based forensic detectors [67, 14, 83] have made encouraging progress through task-specific adaptation and evidence-grounded reasoning. Nevertheless, their reliance on generative reasoning may introduce computational overhead and variability in authenticity judgments. This motivates us to investigate how forensic knowledge from MLLMs can be incorporated into a discriminative feature space to support more reliable authenticity discrimination grounded in semantic plausibility.

A straightforward strategy for leveraging MLLM forensic knowledge is to generate fine-grained evidence descriptions and align them with images through contrastive learning [76]. Such alignment enables the detector to associate visual content with textual evidence, but provides limited guidance on how these associations should contribute to authenticity judgments. A detector may recognize both a penguin and a desert, and even match the image with evidence texts describing their co-occurrence, yet fail to interpret semantic implausibility as potential evidence of image synthesis. Furthermore, multiple forensic descriptions may be relevant to the same image, while offering different degrees of support for authenticity discrimination. This motivates forensic preference learning, where MLLMs assess the relative diagnostic value of candidate evidence and provide image-specific supervision that guides the detector toward more informative forensic cues. To this end, we propose **UniFPL**, a unified forensics preference learning framework that converts MLLM forensic knowledge into image-specific preference supervision. At the data level, UniFPL constructs forensic preference data from two complementary sources: controlled artifact synthesis generates targeted and verifiable artifact evidence, while observed artifact harvesting captures naturally occurring artifact patterns from existing detection data. These image-text candidates are further evaluated through adversarial MLLM rewarding to produce list-level forensic preferences that encode the relative diagnostic value of candidate evidence. At the learning level, UniFPL builds upon a CLIP-based discriminative detector and jointly optimizes authenticity classification, artifacts-aware preference alignment, and semantic structure regularization, with gradient-level coordination to mitigate conflicts among these objectives. This unified design enables the detector to learn from MLLM-derived forensic judgments while preserving pretrained semantic structure and maintaining efficient inference.

Our main contributions are summarized as follows:

- We propose a forensic preference data construction pipeline that combines controlled artifact synthesis and observed artifact harvesting to capture complementary forensic evidence. Through adversarial MLLM rewarding, these candidates are converted into list-level forensic preferences that reflect the relative diagnostic value of different evidence cues.
- We develop a Unified Forensics Preference Learning framework that jointly optimizes authenticity classification, artifacts-aware preference alignment, and semantic structure regularization. With gradient-level coordination, the discriminative detector learns from MLLM-derived forensic preferences while preserving pretrained semantic structure and maintaining efficient inference.
- Extensive experiments across eleven curated and in-the-wild benchmarks demonstrate the strong generalization and robustness of UniFPL, achieving superior detection performance under cross-generator shifts and real-world image degradations over state-of-the-art methods.

2 Related Work

Generalizable Synthetic Image Detection. Research on AI-generated image detection has rapidly evolved as generative models continue to improve in realism and diversity. Early methods exploit frequency-domain artifacts [13], generator fingerprints [73], and local pixel patterns [59, 29], yet these cues often struggle to generalize to unseen generators [43]. Recent works have made notable progress by leveraging large-scale pretrained Vision Foundation Models (VFMs), particularly CLIP [52, 77]. UnivFD [47] employs pretrained CLIP representations for authenticity discrimination, while FatFormer [39] and C2P-CLIP [58] further adapt them through forgery-aware adapters and category-related prompts, respectively. Despite these advances, binary authenticity supervision may encourage reliance on generator-specific artifacts and erode pretrained knowledge [63, 71]. Moreover, recognizing visual concepts alone does not necessarily enable detectors to assess contextual plausibility as evidence of image synthesis [24, 35]. In parallel, recent MLLM-based detectors [30, 14, 83] exploit contextual reasoning to identify forensic inconsistencies, but their reliance on generative inference may introduce unstable authenticity judgments and substantial computational overhead. In contrast, UniFPL leverages MLLM-derived forensic preferences to guide discriminative learning, incorporating contextual forensic evidence while preserving pretrained semantic structure and maintaining efficient inference.

Preference Optimization. Preference optimization enables models to learn from relative judgments rather than isolated labels. Early RLHF [11, 48] methods optimize policies through reward modeling and reinforcement learning, while DPO [53] simplifies this process by directly learning from preferred and rejected responses. Subsequent studies improve optimization stability [45], eliminate reference-model dependence [46, 20], and extend pairwise preferences to listwise ranking [44]. Recent approaches such as CaPO [25] and AMPO [18] further explore richer preference signals for model alignment. While these approaches primarily focus on aligning generative models [49], UniFPL adapts preference supervision to discriminative visual forensics. By constructing forensic preferences from MLLM judgments, UniFPL guides the vision encoder to learn the relative diagnostic value of different forensic cues for authenticity discrimination.

3 Methodology

We introduce **UniFPL**, a unified forensics preference learning framework for generalizable synthetic image detection. Binary authenticity supervision provides only image-level class labels, while conventional image-text alignment captures semantic correspondence without explicitly modeling the relative diagnostic value of forensic evidence. To address this limitation, UniFPL transfers MLLM-derived relative forensic preferences into a discriminative detector through two stages. **Forensic Preference Data Construction** establishes image-specific evidence preferences, while **Unified Forensics Preference Learning** aligns the detector with these preferences and coordinates complementary learning objectives.

3.1 Forensic Preference Data Construction

An image may contain multiple authenticity-related cues, yet these cues have image-dependent and unequal diagnostic value. As illustrated in Fig. 1, we construct image-text candidate sets through two complementary branches: the **Controlled Artifact Branch** provides pre-specified and verifiable

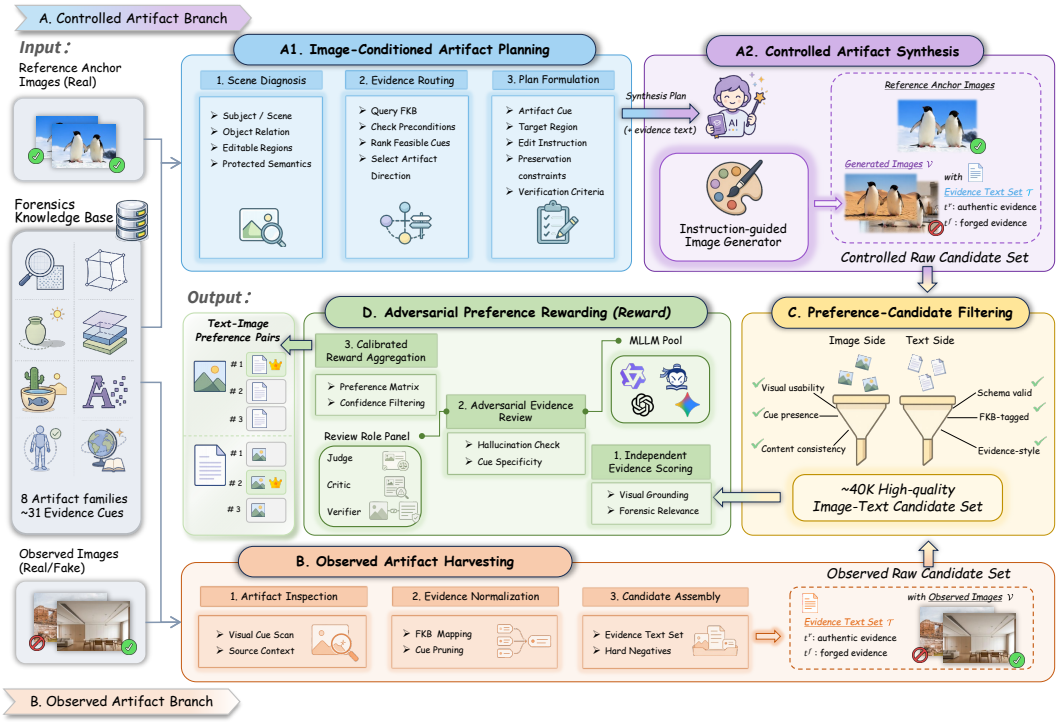


Figure 1: **Overview of Forensic Preference Data Construction.** Guided by a shared Forensics Knowledge Base, the Controlled Artifact Branch constructs image-text candidates through image-conditioned planning and synthesis, while the Observed Artifact Branch harvests naturally occurring forensic traces from existing detection data. After image- and text-side filtering, adversarial MLLM rewarding evaluates and ranks the retained candidates according to their relative diagnostic value.

artifact evidence, while the **Observed Artifact Branch** captures naturally occurring generator traces. After filtering, adversarial preference rewarding converts these candidates into listwise rankings according to their forensic diagnosticity.

To guide artifact selection and evidence normalization, UniFPL constructs a **Forensics Knowledge Base (FKB)** shared by the two branches. We first identify relevant evidence dimensions from studies on image forensics [67, 72, 23] and image quality assessment [81, 64, 15], and then refine them using observations from in-the-wild synthetic images. The resulting taxonomy comprises 8 artifact families and 31 evidence cues. Given the taxonomy $\mathcal{F}$ and a knowledge-base prompt $\mathcal{P}_{know}$, an MLLM organizes these cues into structured knowledge entries:

$$\mathcal{K} = \mathrm{MLLM}(\mathcal{F}, \mathcal{P}_{know}) = \{\kappa_t\}_{t=1}^{31}. \tag{1}$$

Each entry κ_t contains an artifact family, evidence cue, forensic rule, entity precondition, and applicable context. During preference data construction, we employ retrieval-augmented generation (RAG) [27] to mitigate hallucinated evidence and improve the reliability of forensic cue selection. Specifically, using the image context $\mathtt{I}_i^{ctx}$ extracted from a reference image x_i^r as the query, we retrieve the top-k relevant forensic evidence from the FKB:

$$\mathtt{E}_i = \mathrm{RetrieveTopK}(\mathtt{I}_i^{ctx}, \mathcal{K}). \tag{2}$$

3.1.1 Controlled Artifact Branch

Given a real image as the anchor, the Controlled Artifact Branch generates counterfactual images with targeted forensic artifacts through Image-Conditioned Artifact Planning and Controlled Artifact Synthesis. These images, together with the real anchor and corresponding evidence texts, form a controlled raw candidate set with explicit correspondences between visual cues and textual evidence.

Image-Conditioned Artifact Planning. Since feasible artifact cues vary across images, fixed editing templates may introduce mismatched evidence or shortcut signals. We therefore perform image-conditioned planning in three steps. First, given a real reference image x_i^r from the source training

dataset, the MLLM-based planner analyzes its scene content, object relationships, editable regions, and protected semantics to obtain the image context $\mathtt{I}_i^{ctx}$. Second, based on the RAG-retrieved evidence $\mathtt{E}_i$, the planner checks the applicability of forensic cues against the image context and selects H feasible artifact directions, denoted as $\mathtt{E}_i^{sel}$. Third, the planner formulates executable synthesis plans $\mathtt{P}_i$ and corresponding evidence texts $\mathcal{T}_i$ based on the selected artifact directions:

$$\begin{aligned}(\mathtt{P}_i, \mathcal{T}_i) &= \mathrm{MLLM}(x_i^r, \mathtt{I}_i^{ctx}, \mathtt{E}_i^{sel}),\\ \mathtt{P}_i = \{(u_{i,h}, \rho_{i,h}, d_{i,h}, \gamma_{i,h}, v_{i,h})\}_{h=1}^{H}, &\quad \mathcal{T}_i = \{t_i^r\} \cup \{t_{i,h}^f\}_{h=1}^{H}.\end{aligned} \tag{3}$$

where $u_{i,h}$, $\rho_{i,h}$, $d_{i,h}$, $\gamma_{i,h}$, and $v_{i,h}$ denote the target artifact cue, target region, editing instruction, preservation constraints, and verification criteria, respectively. The text t_i^r describes the plausibility of the original image, while $t_{i,h}^f$ specifies the intended forensic violation associated with the h-th counterfactual edit.

Controlled Artifact Synthesis. Given the synthesis plans $\mathtt{P}_i$ and the real reference image x_i^r, the MLLM-based executor employs an instruction-guided image generator to synthesize H images $\{x_{i,h}^f\}_{h=1}^{H}$ with pre-specified forensic artifacts. Each generated image is then checked for the presence of the target artifact and compliance with the preservation constraints. The generated images, together with the real reference image and corresponding evidence texts $\mathcal{T}_i$, form a controlled raw candidate set $\mathcal{G}_i^{\mathrm{crc}} = (\mathcal{V}_i, \mathcal{T}_i)$, where $\mathcal{V}_i = \{x_i^r\} \cup \{x_{i,h}^f\}_{h=1}^{H}$.

3.1.2 Observed Artifact Branch

To further enhance the diversity of forensic evidence, we introduce the Observed Artifact Branch, which mines naturally occurring artifact cues from existing detection datasets and organizes them into image-text candidate sets. Given an observed image group containing authentic and synthetic samples, $\mathcal{V}_j = \{x_{j,m}^r\}_{m=1}^{M} \cup \{x_{j,m}^f\}_{m=1}^{M}$, an MLLM-based curator first inspects the synthetic images to identify visible artifact cues and their source context. These cues are then mapped to the FKB and pruned to remove ambiguous or visually unsupported evidence, yielding the normalized evidence set $\mathtt{E}_j^{\mathrm{obs}}$. Finally, guided by an evidence-assembly prompt $\mathcal{P}_{\mathrm{asm}}$, the curator generates corresponding evidence texts for both authentic and synthetic images:

$$\mathcal{T}_j = \mathrm{MLLM}(\mathcal{V}_j, \mathtt{E}_j^{\mathrm{obs}}, \mathcal{P}_{\mathrm{asm}}) = \{t_{j,m}^r\}_{m=1}^{M} \cup \{t_{j,m}^f\}_{m=1}^{M}. \tag{4}$$

Here, $t_{j,m}^r$ describes the plausibility of the authentic image, while $t_{j,m}^f$ explains the observed forensic violation in the synthetic image. Together, the observed images and their evidence texts form an observed raw candidate set $\mathcal{G}_j^{\mathrm{orc}} = (\mathcal{V}_j, \mathcal{T}_j)$.

3.1.3 Preference-Candidate Filtering

Raw candidate sets from both branches may contain ambiguous visual artifacts, unintended content changes, or weakly grounded evidence texts. To ensure reliable preference supervision, we apply modality-specific filtering before preference rewarding. On the image side, we assess visual usability, cue presence, and content consistency. On the text side, we require evidence texts to be schema-valid, FKB-tagged, and written in a forensic evidence style rather than as generic descriptions. After filtering, we obtain approximately 40K high-quality image-text candidate sets.

3.1.4 Adversarial Preference Rewarding

To capture the relative diagnostic value of forensic evidence, we introduce an adversarial MLLM rewarding strategy that converts filtered image-text candidates into list-level preferences. We denote the n-th retained candidate set from the two branches as $\mathcal{G}_n = (\mathcal{V}_n, \mathcal{T}_n)$. For each image or text anchor a, let $\mathcal{C}_n(a)$ denote its opposite-modality candidates. An MLLM pool first independently scores each anchor-candidate pair based on visual grounding and forensic relevance. These assessments are further refined through adversarial review, where the critic challenges hallucinated or generic evidence, the verifier examines cue specificity and visual support, and the judge aggregates the reviewed assessments into calibrated forensic rewards $\alpha_n(a, c)$. After confidence filtering, candidates are ranked according to their rewards to obtain an anchor-conditioned preference list:

$$\mathcal{R}_n(a) = \{(c, \alpha_n(a, c)) \mid c \in \mathcal{C}_n(a)\}, \tag{5}$$

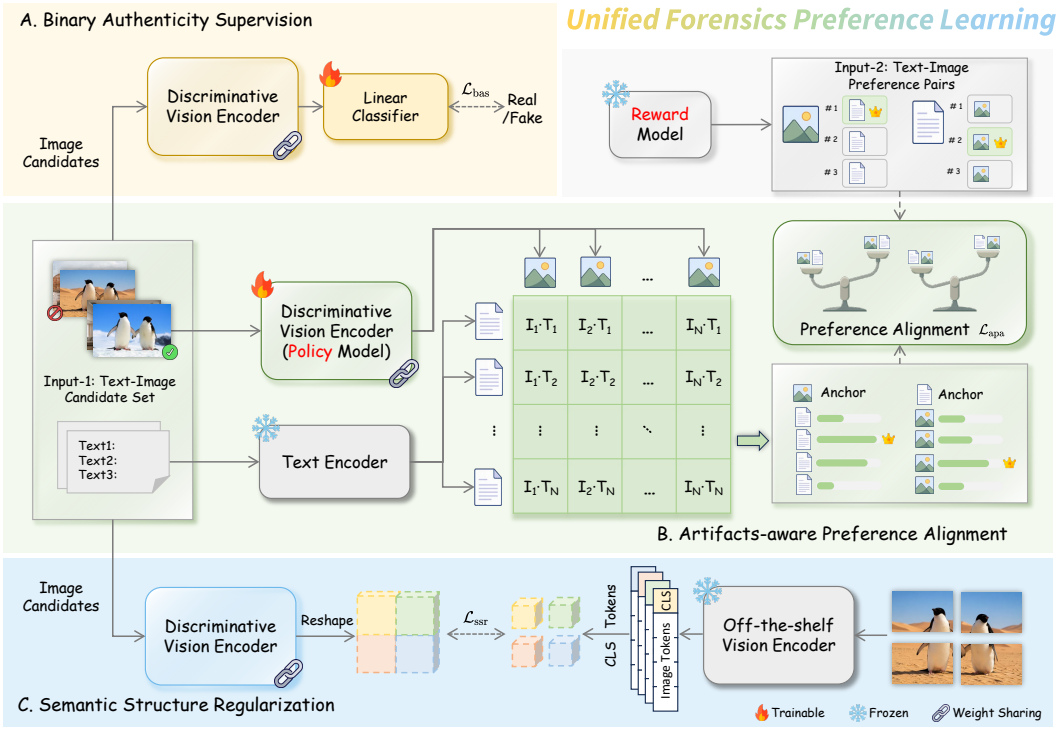


Figure 2: **Overview of Unified Forensics Preference Learning.** UniFPL jointly optimizes three objectives: (A) Binary Authenticity Supervision (BAS) ensures stable authenticity discrimination; (B) Artifacts-aware Preference Alignment (APA) aligns bidirectional image-text preferences with MLLM-derived forensic preferences; and (C) Semantic Structure Regularization (SSR) preserves pretrained representations through region-level self-distillation.

where the preference order is determined by the descending forensic rewards. Collecting these preferences from both image and text anchors yields the final **Forensic Preference Dataset** $\mathcal{D}$:

$$\mathcal{R}_n = \{\mathcal{R}_n(a)\}_{a \in \mathcal{V}_n \cup \mathcal{T}_n}, \qquad \mathcal{D} = \{(\mathcal{G}_n, \mathcal{R}_n)\}_{n=1}^{N}. \tag{6}$$

3.2 Unified Forensics Preference Learning

To transfer MLLM-derived forensic knowledge into discriminative representations, UniFPL jointly optimizes three complementary objectives, as illustrated in Fig. 2. Specifically, **Binary Authenticity Supervision** provides stable supervision for authenticity discrimination; **Artifacts-aware Preference Alignment** aligns visual representations with forensic preferences; and **Semantic Structure Regularization** preserves generalizable pretrained representations. To mitigate potential conflicts among these objectives, we further coordinate their gradients during training.

Binary Authenticity Supervision. To ground forensic preference learning in stable authenticity discrimination, we attach a lightweight classifier to the shared discriminative vision encoder $\mathcal{E}_v$. For each image x in the sampled preference sets, the classifier predicts its authenticity probability $\hat{y}$, which is supervised by its label y through the binary cross-entropy loss:

$$\mathcal{L}_{\text{bas}}(x, y) = -\left[y \log \hat{y} + (1 - y) \log(1 - \hat{y})\right]. \tag{7}$$

At inference, only the vision encoder and lightweight classifier are retained, enabling efficient and reliable authenticity detection.

Artifacts-aware Preference Alignment. To incorporate semantic plausibility into authenticity discrimination, we introduce Artifacts-aware Preference Alignment (APA). Inspired by DPO, APA treats the discriminative vision encoder as a policy model, where image-text similarity scores reflect its preferences over candidate evidence, while MLLM-derived forensic rewards provide sophisticated

preference signals. By aligning the preference distributions in both image-to-text and text-to-image directions, APA guides the visual representations to distinguish more diagnostic forensic cues from less relevant ones.

Given a candidate set $\mathcal{G}_n = (\mathcal{V}_n, \mathcal{T}_n)$, we consider each image or text as an anchor a, with $\mathcal{C}_n(a)$ denoting its opposite-modality candidates. For each anchor-candidate pair (a, c), $\alpha_n(a, c)$ denotes the MLLM-derived forensic reward, while $s_\theta(a, c)$ denotes the detector's image-text cosine similarity. We apply softmax over the same candidates to obtain the corresponding preference distributions:

$$p_\Phi(c \mid a) = \frac{\exp(\alpha_n(a, c))}{\sum_{c' \in \mathcal{C}_n(a)} \exp(\alpha_n(a, c'))}, \quad p_\theta(c \mid a) = \frac{\exp(s_\theta(a, c))}{\sum_{c' \in \mathcal{C}_n(a)} \exp(s_\theta(a, c'))}. \tag{8}$$

Here, p_Φ and p_θ denote the MLLM-derived target and detector-induced preference distributions, respectively. We adopt soft distributional alignment to preserve the relative preference strengths among candidates and avoid over-penalizing ambiguous evidence. For image anchors, the preference loss is defined as:

$$\mathcal{L}_{\mathrm{i2t}} = \frac{1}{B} \sum_{n=1}^{B} \frac{1}{|\mathcal{V}_n|} \sum_{a \in \mathcal{V}_n} \mathrm{KL}\left(p_\Phi(\cdot \mid a) \,\|\, p_\theta(\cdot \mid a)\right). \tag{9}$$

where B denotes the number of preference sets in a mini-batch. The text-to-image direction is defined symmetrically as $\mathcal{L}_{\mathrm{t2i}}$, and the final APA loss is:

$$\mathcal{L}_{\mathrm{apa}} = \frac{1}{2}\left(\mathcal{L}_{\mathrm{i2t}} + \mathcal{L}_{\mathrm{t2i}}\right). \tag{10}$$

Semantic Structure Regularization. Adapting a pretrained encoder to authenticity discrimination may enhance its sensitivity to forensic artifacts but also induce catastrophic forgetting of the generalizable knowledge acquired during pretraining. To mitigate this effect, we adopt region-level self-distillation [65] with a frozen pretrained vision encoder as the teacher. Given an image x, we partition it into Q regions and obtain the corresponding crops $\{x^{(q)}\}_{q=1}^{Q}$. The shared discriminative vision encoder processes the full image and pools a student feature f_q^s from its feature map for each region, while the frozen teacher extracts the corresponding feature f_q^t from the crop $x^{(q)}$. We minimize the cosine distance between these representations to preserve the pretrained semantic knowledge during forensic adaptation:

$$\mathcal{L}_{\mathrm{ssr}}(x) = \frac{1}{Q} \sum_{q=1}^{Q} \left(1 - \frac{f_q^s \cdot f_q^t}{\|f_q^s\|_2 \|f_q^t\|_2}\right). \tag{11}$$

Gradient-Level Preference Coordination. The three objectives provide complementary supervision for the shared Discriminative Vision Encoder, but their gradients may differ in magnitude and direction. Consequently, fixed loss weighting may favor objectives with larger gradients while failing to resolve directional conflicts. Inspired by ConFIG [40], we coordinate these gradients directly at the gradient level. For the trainable parameters θ of the shared Discriminative Vision Encoder, we compute the objective-specific gradients:

$$g_{\mathrm{bas}} = \nabla_\theta \mathcal{L}_{\mathrm{bas}}, \quad g_{\mathrm{apa}} = \nabla_\theta \mathcal{L}_{\mathrm{apa}}, \quad g_{\mathrm{ssr}} = \nabla_\theta \mathcal{L}_{\mathrm{ssr}}. \tag{12}$$

To reduce the influence of gradient magnitudes on the update direction, we normalize the gradients and construct the gradient matrix:

$$\mathbf{G} = [U(g_{\mathrm{bas}}), U(g_{\mathrm{apa}}), U(g_{\mathrm{ssr}})], \qquad U(g) = \frac{g}{\|g\|_2 + \epsilon}. \tag{13}$$

We then seek a common update direction with equal positive projections onto the normalized gradients:

$$\bar{g} = U\left(\mathbf{G}^{-\top} \mathbf{1}_3\right), \tag{14}$$

where $\mathbf{G}^{-\top}$ denotes the pseudoinverse of $\mathbf{G}^\top$, and $\mathbf{1}_3$ is a three-dimensional all-ones vector. We further adaptively scale this direction using the sum of the projections of the original gradients onto $\bar{g}$:

$$g_{\mathrm{cf}} = \sum_{\ell \in \mathcal{O}} \left(g_\ell^\top \bar{g}\right) \bar{g}, \qquad \mathcal{O} = \{\mathrm{bas}, \mathrm{apa}, \mathrm{ssr}\}. \tag{15}$$

The resulting gradient g_{cf} is passed to the optimizer to update the shared vision encoder, allowing the three objectives to be jointly optimized without manually tuning fixed loss weights.

Table 1: **Overall comparison across 11 benchmarks.** We report balanced accuracy (%) across all benchmarks except EvalGEN, which contains only synthetic images and is evaluated using synthetic-image detection accuracy. For datasets with multiple generators, results are reported as mean ± standard deviation. We follow the same evaluation protocol as DDA [8] and adopt its reported baseline results. The best and second-best results are marked in **bold** and underline, respectively.

Dataset	NPR CVPR'24	UnivFD CVPR'23	FatFormer CVPR'24	SAFE KDD'25	C2P-CLIP AAAI'25	AIDE ICLR'25	DRCT ICML'24	AlignedForensics ICLR'25	DDA NeurIPS'25	**Ours** —
Manually Curated Datasets										
GenImage	51.5 ± 6.3	64.1 ± 10.8	62.8 ± 10.4	50.3 ± 1.2	74.4 ± 8.4	61.2 ± 11.9	84.7 ± 2.7	79.0 ± 22.7	91.7 ± 7.8	**94.9** ± 4.3
DRCT-2M	37.3 ± 15.0	61.8 ± 8.9	52.2 ± 5.7	59.3 ± 19.2	59.2 ± 9.9	64.6 ± 11.8	90.5 ± 7.4	95.5 ± 6.1	**98.1** ± 1.4	**98.1** ± 0.5
DDA-COCO	42.2 ± 5.4	52.4 ± 1.5	51.7 ± 1.5	49.9 ± 0.3	51.3 ± 0.6	50.0 ± 0.4	60.2 ± 4.3	86.5 ± 19.1	92.2 ± 10.6	**94.9** ± 3.5
EvalGEN	2.9 ± 2.7	15.4 ± 14.2	45.6 ± 33.1	1.1 ± 0.6	38.9 ± 31.2	19.1 ± 11.1	77.8 ± 5.4	68.0 ± 20.7	**97.2** ± 4.2	95.3 ± 3.9
Synthbuster	50.0 ± 2.6	67.8 ± 14.4	56.1 ± 10.7	46.5 ± 20.8	68.5 ± 11.4	53.9 ± 18.6	84.8 ± 3.6	77.4 ± 25.0	90.1 ± 5.6	**94.8** ± 5.5
ForenSynths	47.9 ± 22.6	77.7 ± 16.1	90.0 ± 11.8	49.7 ± 2.7	**92.0** ± 10.1	59.4 ± 24.6	73.9 ± 13.4	53.9 ± 7.1	81.4 ± 13.9	87.5 ± 3.4
AIGCDetection	53.1 ± 12.2	72.5 ± 17.3	85.0 ± 14.9	50.3 ± 1.1	81.4 ± 15.6	63.6 ± 13.9	81.4 ± 12.2	66.6 ± 21.6	87.8 ± 12.6	**92.4** ± 4.9
In-the-Wild Datasets										
Chameleon	59.9	50.7	51.2	59.2	51.1	63.1	56.6	71.0	82.4	**86.1**
SynthWildx	49.8 ± 10.0	52.3 ± 11.3	52.1 ± 8.2	49.1 ± 0.7	57.1 ± 4.2	48.8 ± 0.8	55.1 ± 1.8	78.8 ± 17.8	90.9 ± 3.1	**92.8** ± 2.8
WildRF	63.5 ± 13.6	55.3 ± 5.7	58.9 ± 8.0	57.2 ± 18.5	59.6 ± 7.7	58.4 ± 12.9	50.6 ± 3.5	80.1 ± 10.3	90.3 ± 3.5	**93.1** ± 2.0
BFree-Online	49.5	49.0	50.0	50.5	50.0	53.1	55.7	68.5	95.1	**97.5**
Average	46.1	56.3	59.6	47.6	62.1	54.1	70.1	75.0	90.7	**93.4**
Minimum	2.9	15.4	45.6	1.1	38.9	19.1	50.6	53.9	81.4	**86.1**

4 Experiments

In this section, we conduct a comprehensive empirical evaluation of UniFPL. We first describe the experimental setup and then present extensive results and analyses.

4.1 Experimental Setup

Datasets. To thoroughly evaluate the detection performance and generalization ability of UniFPL, we assess all compared methods on eleven diverse datasets, including two categories: *(1) Manually Curated Datasets*: GenImage [84], DRCT-2M [7], Synthbuster [4], DDA-COCO [8], EvalGEN [8], AIGCDetection [82], and ForenSynths [62]; and *(2) In-the-Wild Datasets*: Chameleon [69], WildRF [6], SynthWildx [12], and BFree-Online [17]. These datasets encompass diverse image sources and generation models, including GANs, diffusion models, autoregressive models, and unknown generators. They also vary in image content, format, and resolution, providing a comprehensive evaluation of generalization across different data distributions.

Evaluation Metrics and Baselines. Following prior work [8], we adopt balanced accuracy as the primary evaluation metric. To provide a comprehensive comparison, we evaluate UniFPL against a diverse set of state-of-the-art methods, including NPR [59], SAFE [28], AIDE [69], UnivFD [47], FatFormer [39], C2P-CLIP [58], DRCT [7], AlignedForensics [54], and DDA [8]. To ensure a fair comparison, we follow a consistent evaluation protocol.

Implementation Details. The Discriminative Vision Encoder, Off-the-shelf Vision Encoder, and Text Encoder are all based on OpenAI CLIP ViT-L/14 [52]. The off-the-shelf vision and text encoders remain frozen, while the discriminative vision encoder is fine-tuned with LoRA [21] to effectively capture forensic evidence while mitigating overfitting. For Semantic Structure Regularization, each image is partitioned into $Q = m \times n$ regions, where m and n are randomly sampled from $[1, 6]$. Student region features are pooled from the full-image feature map, while the frozen teacher extracts corresponding representations from the cropped regions. We train UniFPL using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$, a batch size of 16 preference sets, and up to 6 epochs on NVIDIA A800 GPUs. The learning rate follows a cosine annealing schedule from 1×10^{-3} to 1×10^{-4}. The forensic preference data are constructed from the SD v1.4 subset of DRCT-2M [7].

4.2 Quantitative Analysis

Table 1 presents a comprehensive comparison of UniFPL with existing methods across eleven datasets, including seven manually curated benchmarks and four in-the-wild datasets.

Evaluation on Manually Curated Datasets. Table 1 presents results on seven manually curated benchmarks, covering diverse generation paradigms and settings that pose substantial challenges to

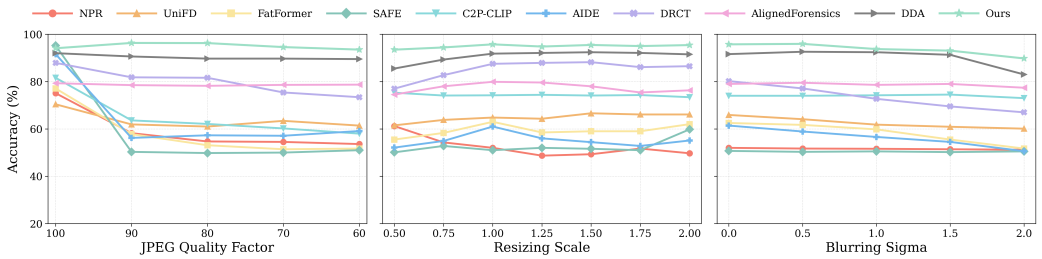


Figure 3: **Robustness comparison under image degradations.** Balanced accuracy (%) on GenImage under varying levels of JPEG compression, resizing, and Gaussian blurring.

Table 2: **Ablation study of UniFPL.** We evaluate the contributions of key design choices.

(a) Objective Components

$\mathcal{L}_{\text{bas}}$	$\mathcal{L}_{\text{apa}}$	$\mathcal{L}_{\text{ssr}}$	mAcc
✓			88.4
✓	✓		93.2
✓		✓	90.7
✓	✓	✓	**94.9**

(b) Coordination Strategy

Strategy	mAcc
Fixed Weighting	92.4
GradNorm	93.6
PCGrad	93.8
Ours	**94.9**

(c) Preference Optimization Design

Design	mAcc
Shuffled Preference	87.2
Pair-wise	92.2
Hard List-wise	93.4
Ours	**94.9**

cross-generator generalization. UniFPL achieves the highest average score of 94.0% and outperforms the strongest baseline DDA by 2.8 percentage points. Notably, UniFPL achieves 94.9% on GenImage and surpasses DDA by 3.2 percentage points across its diverse generator families. Moreover, UniFPL achieves a minimum score of 87.5% across these benchmarks compared with 81.4% for DDA, indicating more consistent detection performance across heterogeneous synthetic sources. These results highlight the effectiveness of combining forensic preference alignment with pretrained representation preservation in learning transferable authenticity cues.

Evaluation on In-the-Wild Datasets. Table 1 further presents results on four in-the-wild benchmarks, featuring high-fidelity synthetic images, commercial generators, and content collected from social platforms. These less controlled conditions pose substantial challenges to existing detectors. UniFPL achieves the highest average balanced accuracy of 92.4% and outperforms the strongest baseline DDA by 2.7 percentage points, with consistent improvements across all four benchmarks. Notably, on Chameleon, where most competing detectors struggle with highly realistic synthetic images, UniFPL achieves 86.1%, surpassing DDA by 3.7 percentage points. These gains stem from forensic preference learning, which guides the detector toward more informative authenticity cues. By coordinating authenticity supervision, forensic preference alignment, and semantic structure regularization at the gradient level, UniFPL strengthens forensic discrimination while preserving generalizable pretrained representations, enabling reliable detection across diverse real-world conditions.

4.3 Robustness Evaluation

Following the evaluation protocol of prior works [8], we further assess the robustness of UniFPL on GenImage under three common image degradations: JPEG compression, resizing, and Gaussian blurring. As shown in Fig. 3, UniFPL consistently achieves the best performance across all three settings, maintaining relatively stable accuracy even under stronger perturbations. In contrast, several competing methods exhibit substantial performance degradation. These results suggest that the forensic representations learned by UniFPL remain discriminative under image degradations, supporting reliable authenticity detection even in practical scenarios involving post-processing distortions.

4.4 Ablation Study

We perform extensive ablation studies on GenImage, with mean balanced accuracy (mAcc) reported in Table 2.

Effectiveness of Training Objectives. As shown in Table 2(a), the contributions of our training objectives are evident from the substantial performance degradation observed when individual

objectives are removed. Eliminating APA reduces mAcc to 90.7%, revealing the critical role of forensic preferences in guiding the detector toward discriminative forensic cues. Removing SSR lowers performance to 93.2%, reflecting the benefit of preserving pretrained semantic structure during forensic adaptation. When both objectives are excluded, mAcc falls to 88.4%, indicating the limitations of relying solely on authenticity supervision. The complete model achieves 94.9%, demonstrating that forensic preference learning and pretrained knowledge preservation complement authenticity supervision to strengthen generalization.

Analysis of Objective Balancing. As shown in Table 2(b), our full model achieves an mAcc of 94.9% with gradient-level coordination. In comparison, fixed weighting achieves 92.4%, while GradNorm [9] and PCGrad [74] reach 93.6% and 93.8%, respectively. These results indicate that conventional weighting and gradient adjustment strategies are less effective at coordinating the three objectives. In contrast, our method constructs a common update direction from the normalized objective-specific gradients to mitigate directional conflicts during joint optimization.

Analysis of Preference Design. Table 2(c) examines the effectiveness of different preference optimization designs. Shuffling forensic preferences reduces mAcc to 87.2%, demonstrating that reliable preference rankings are essential for learning discriminative forensic cues. Pair-wise and hard list-wise optimization achieve 92.2% and 93.4%, respectively, but both remain below our full model. Pair-wise comparisons do not explicitly model relationships across the entire candidate set, while hard list-wise targets discard information about relative preference strengths. In contrast, our soft list-wise formulation preserves these graded preferences by aligning the detector with the complete preference distribution, achieving the best overall performance.

5 Conclusion

We present **UniFPL**, a Unified Forensics Preference Learning framework for generalizable synthetic image detection. By combining controlled artifact synthesis, observed artifact mining, and adversarial MLLM rewarding, UniFPL constructs image-specific, list-level forensic preferences that capture the relative diagnostic value of candidate evidence. Through gradient-level coordination of authenticity supervision, artifacts-aware preference alignment, and semantic structure regularization, UniFPL enables a discriminative vision encoder to learn from MLLM-derived forensic preferences while preserving pretrained semantic structure. The resulting detector maintains efficient inference. Extensive experiments across eleven curated and in-the-wild benchmarks demonstrate its strong generalization and robustness to common image degradations.

Limitations and Future Work. Despite its strong performance and efficient inference, UniFPL requires a computationally demanding offline process for forensic preference construction. This process involves artifact synthesis, candidate filtering, and repeated MLLM evaluations, increasing the cost of preparing preference data at scale. Future work will focus on improving the efficiency and scalability of preference construction while exploring more effective ways to leverage MLLM forensic knowledge.

6 Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62672013 and 62176010), the Beijing Natural Science Foundation (No. 4252029), the Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515140021), and the Guangdong Science and Technology Department (No. 2024ZDZX2004).

References

[1] Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Clip under the microscope: A fine-grained analysis of multi-object representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9317, 2025.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2023.

[5] Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. *Advances in Neural Information Processing Systems*, 37:113436–113460, 2024.

[6] Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024.

[7] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024.

[8] Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025.

[9] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[11] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[12] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024.

[13] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3247–3258, 2020.

[14] Yueying Gao, Dongliang Chang, Bingyao Yu, Haotian Qin, Lei Chen, Kongming Liang, and Zhanyu Ma. Fakereasoning: Towards generalizable forgery detection and reasoning. *arXiv e-prints*, pages arXiv–2503, 2025.

[15] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

[16] Google DeepMind. Gemini 3 pro. `https://deepmind.google/models/gemini/pro/`, 2025. Accessed: 2026-05-06.

[17] Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. *arXiv preprint arXiv:2412.17671*, 2024.

[18] Taneesh Gupta, Rahul Madhavan, Xuchao Zhang, Chetan Bansal, and Saravan Rajmohan. Ampo: Active multi-preference optimization for self-play preference selection. *arXiv preprint arXiv:2502.18293*, 2025.

[19] Yiran He, Yun Cao, Bowen Yang, and Zeyu Zhang. Can gpt tell us why these images are synthesized? empowering multimodal large language models for forensics. In *Proceedings of the 2025 ACM Workshop on Information Hiding and Multimedia Security*, pages 24–34, 2025.

[20] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, 2024.

[21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[22] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, 2023.

[23] Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, et al. Legion: Learning to ground and explain for synthetic image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18937–18947, 2025.

[24] Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is clip ideal? no. can we fix it? yes! In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22436–22446, 2025.

[25] Kyungmin Lee, Xiahong Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18465–18475, 2025.

[26] Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1487–1500, 2024.

[27] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[28] Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. *arXiv preprint arXiv:2408.06741*, 2024.

[29] Shuaibo Li, Wei Ma, Jianwei Guo, Shibiao Xu, Benchong Li, and Xiaopeng Zhang. Unionformer: Unified-learning transformer with multi-view representation for image manipulation detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12523–12533, 2024.

[30] Shuaibo Li, HAO Pengfei, Hongtao Wu, Jianfeng Dong, Ping Li, Xiaohong Liu, and Lei Zhu. Hermes: An evidence-driven agentic framework for trustworthy and explainable ai-generated video detection. In *Forty-third International Conference on Machine Learning*.

[31] Shuaibo Li, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, and Lei Zhu. Toward medical deepfake detection: A comprehensive dataset and novel method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 626–637. Springer, 2025.

[32] Shuaibo Li, Shibiao Xu, Wei Ma, and Qiu Zong. Image manipulation localization using attentional cross-domain cnn features. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5614–5628, 2021.

[33] Shuaibo Li, Yijun Yang, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, Qing Zhang, and Lei Zhu. Synerdetect: Hierarchical synergistic learning for generalizable ai-generated image detection. 2026.

[34] Shuaibo Li, Laixin Zhang, Wei Ma, Jianwei Guo, Shibiao Xu, Zhijie Qiu, and Hongbin Zha. Realnet: Efficient and unsupervised detection of ai-generated images via real-only representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38889–38898, 2026.

[35] Siting Li, Pang Wei Koh, and Simon Shaolei Du. Exploring how generative mllms perceive more than clip with the same vision encoder. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10101–10119, 2025.

[36] Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? *arXiv preprint arXiv:2505.12335*, 2025.

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, 2014.

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[39] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10770–10780, 2024.

[40] Qiang Liu, Mengyu Chu, and Nils Thuerey. Config: Towards conflict-free training of physics informed neural networks. *arXiv preprint arXiv:2408.11104*, 2024.

[41] Ruiqi Liu, Manni Cui, Ziheng Qin, Zhiyuan Yan, Ruoxin Chen, Yi Han, Zhiheng Li, Junkai Chen, ZhiJin Chen, Kaiqing Lin, et al. Mirror: Manifold ideal reference reconstructor for generalizable ai-generated image detection. *arXiv preprint arXiv:2602.02222*, 2026.

[42] Ruiqi Liu, Yi Han, Boyi Sun, Houxin He, Yuxuan Zhou, and Yan Wang. M 2 ea: Multi-vae manifold envelope alignment for challenging ai-generated image detection. In *ICASSP 2026*, pages 9147–9151. IEEE, 2026.

[43] Ruiqi Liu, Yi Han, Zhengbo Zhang, Liwei Yao, Zhiyuan Yan, Jialiang Shen, ZhiJin Chen, Manni Cui, Boyi Sun, Lubin Weng, et al. Beyond artifacts: Real-centric envelope modeling for reliable ai-generated image detection. In *European Conference on Computer Vision*, pages 225–243. Springer, 2026.

[44] Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, et al. Lipo: Listwise preference optimization through learning-to-rank. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2404–2420, 2025.

[45] Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer. *Advances in Neural Information Processing Systems*, 37:138663–138697, 2024.

[46] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

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

[48] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[49] Weiwei Qi, Chongyu Wang, Tianhang Zheng, Zefeng Wu, Zhilin Guo, Xiaojun Jia, Zhongjie Ba, and Kui Ren. Ascent: First-order optimal fine-tuning with recalibration for safety–utility co-enhancement, 2026.

[50] Weiwei Qi, Zefeng Wu, Zhilin Guo, Tianhang Zheng, Chaochao Lu, Liang He, Zhan Qin, and Kui Ren. Darwin: Evolving jailbreak adversary and guardrail for llm safety evaluation and protection. *arXiv preprint arXiv:2607.19829*, 2026.

[51] Weiwei Qi, Zefeng Wu, Tianhang Zheng, Zikang Zhang, Xiaojun Jia, Zhan Qin, and Kui Ren. Towards identification and intervention of safety-critical parameters in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 32293–32312, 2026.

[52] Alec Radford et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[53] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[54] Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images, 2025.

[55] Bytedance Seed. Seed1. 8 model card: Towards generalized real-world agency. *arXiv preprint arXiv:2603.20633*, 2026.

[56] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[57] Yiren Song, Shengtao Lou, Xiaokang Liu, Hai Ci, Pei Yang, Jiaming Liu, and Mike Zheng Shou. Anti-reference: Universal and immediate defense against reference-based generation. *arXiv preprint arXiv:2412.05980*, 2024.

[58] 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*, 2024.

[59] 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*, pages 28130–28139, 2024.

[60] Piotr Teterwak, Ximeng Sun, Bryan A Plummer, Kate Saenko, and Ser-Nam Lim. Clamp: contrastive language model prompt-tuning. *arXiv preprint arXiv:2312.01629*, 2023.

[61] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578, 2024.

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

[63] Yabin Wang, Zhiwu Huang, Zhou Su, Adam Prugel-Bennett, and Xiaopeng Hong. Penny-wise and pound-foolish in ai-generated image detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[64] Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, et al. Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation. *arXiv preprint arXiv:2510.18701*, 2025.

[65] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023.

[66] Zefeng Wu, Weiwei Qi, Jielong Chen, Tianhang Zheng, Di Hong, Chaochao Lu, Liang He, Zhan Qin, and Kui Ren. Datashield: Uncovering risky fine-tuning data across llms through consensus subspace alignment. *arXiv preprint arXiv:2607.15081*, 2026.

[67] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*, 2024.

[68] Jiazhen Yan, Fan Wang, Weiwei Jiang, Ziqiang Li, and Zhangjie Fu. Ns-net: Decoupling clip semantic information through null-space for generalizable ai-generated image detection. *arXiv preprint arXiv:2508.01248*, 2025.

[69] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024.

[70] Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Zhendong Wang, Hao Liu, Bin Lin, Hao Li, Xue Xu, et al. Can understanding and generation truly benefit together–or just coexist? *arXiv preprint arXiv:2509.09666*, 2025.

[71] Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024.

[72] Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, et al. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. *arXiv preprint arXiv:2410.09732*, 2024.

[73] 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*, pages 7556–7566, 2019.

[74] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020.

[75] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

[76] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024.

[77] Haifeng Zhang, Qinghui He, Xiuli Bi, Bo Liu, Chi-Man Pun, and Bin Xiao. Towards universal AI-generated image detection by variational information bottleneck network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[78] Haifeng Zhang, Qinghui He, Xiuli Bi, Bo Liu, Chi-Man Pun, and Bin Xiao. Combating pattern and content bias: Adversarial feature learning for generalized AI-generated image detection. *IEEE Transactions on Image Processing*, 2026.

[79] Laixin Zhang, Shuaibo Li, Wei Ma, and Hongbin Zha. Truemoe: Dual-routing mixture of discriminative experts for synthetic image detection. *arXiv preprint arXiv:2509.15741*, 2025.

[80] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *Advances in Neural Information Processing Systems*, 37:51727–51753, 2024.

[81] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

[82] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. PatchCraft: Exploring Texture Patch for Efficient AI-generated Image Detection. *arXiv e-prints*, page arXiv:2311.12397, 2023.

[83] Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18746–18758, 2025.

[84] 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. In *Advances in Neural Information Processing Systems*, volume 36, pages 77771–77782, 2023.

This supplementary document provides additional information to complement the main paper. Section A details the construction of the Forensic Preference dataset. Section B provides more implementation details for data construction. Section C introduces the evaluation benchmarks used in our experiments. Section D presents qualitative visualizations to support our design motivation. Section E provides additional ablation studies.

A More Details about Preference Dataset

We construct the Forensic Preference Dataset using the SD v1.4 training subset of DRCT-2M [7]. The authentic images originate from MSCOCO [37], which contains approximately 118K images across 80 object categories. Following B-Free [17], we retain only Creative Commons-licensed images and discard samples where the target objects are absent, yielding an initial pool of 51,517 authentic images and 118,287 SD v1.4-generated synthetic images. After branch-specific construction and filtering, the Controlled Artifact Branch retains 31,064 candidate sets, each containing one authentic image and three controlled counterfactual images, yielding 93,192 synthetic images. The Observed Artifact Branch retains 12,054 candidate sets, each comprising two authentic and two synthetic images, yielding 24,108 images per class. The two branches share 13,135 authentic images, resulting in 42,037 unique authentic images and 117,300 synthetic images, totaling 159,337 images. The resulting dataset comprises $N = 43{,}118$ image-text preference sets.

We compare UniFPL with four representative training-data construction methods, including DRCT [7], AlignedForensics [54], B-Free [17], and DDA [8]. Table 3 summarizes the number of authentic and synthetic training images and their corresponding generation strategies. Rather than simply increasing the volume of training data, UniFPL emphasizes the quality of forensic supervision by combining controlled and observed artifact evidence with MLLM-derived preference rankings. Although our dataset contains fewer images than these large-scale training sets, it provides richer evidence-level supervision by capturing the relative diagnostic value of forensic cues beyond binary authenticity labels.

B More Implementation Details

For forensic preference data construction, we employ Qwen3-VL-8B [3] as the primary MLLM. The Forensics Knowledge Base (FKB) comprises eight artifact families and 31 evidence cues derived from prior image forensics and image quality assessment studies, as illustrated in Fig. 5. Both branches employ RAG [27] to retrieve the top five relevant forensic evidence entries from the FKB. In the Controlled Artifact Branch, the MLLM-based planner selects three feasible artifact directions for each authentic image ($H = 3$). The corresponding counterfactual images are synthesized using an SD-based instruction-guided generator. In the Observed Artifact Branch, samples are organized into category-wise groups, each containing two authentic and two synthetic images ($M = 2$). We employ Seed 1.8 [55] to filter raw candidate sets before preference rewarding. The adversarial rewarding pool consists of Qwen3-VL-8B [3], InternVL2-8B [10], GPT-5 [56], and Gemini 3 [16]. The models independently score each anchor-candidate pair, and the resulting assessments undergo adversarial review, calibrated reward aggregation, and confidence filtering to obtain reliable list-level forensic preferences.

C Details of the Test Dataset

To comprehensively evaluate the generalization ability, we conduct experiments on 11 representative synthetic image detection datasets, covering mainstream curated benchmarks and in-the-wild evaluation sets. These datasets span diverse generators, image sources, post-processing pipelines, and scene compositions, with different emphases on cross-generator transfer, real-world robustness, and distributional bias.

GenImage [84] is a million-scale synthetic image detection benchmark built on ImageNet, covering 1,000 object categories with 1,331,167 authentic images and 1,350,000 synthetic images. Its synthetic images are generated by eight representative generators, including Midjourney, Stable Diffusion V1.4/V1.5, ADM, GLIDE, Wukong, VQDM, and BigGAN.

Table 3: Comparison of training data scales and construction strategies across different methods. "Number" denotes the numbers of real and synthetic training images, respectively.

Method	Number	Generation Method
DRCT	118K / 354K	Diff. Rec.
AlignedForensics	179K / 179K	VAE Rec.
B-Free	51K / 309K	Diff. Rec. + Inpaint.
DDA	118K / 118K	VAE Rec. + DDA
Ours	42K / 117K	Instruction-guided Image Gen.

DRCT-2M [7] is a large-scale benchmark built from MSCOCO real images and synthetic images generated by 16 diffusion-based generation pipelines, including text-to-image, ControlNet-based image-to-image, and diffusion-reconstruction variants. It contains about two million generated images, with each generator subset containing roughly 120K images.

DDA-COCO [8] is a bias-reduced evaluation benchmark constructed from the MSCOCO validation set. It contains authentic images and their synthetic counterparts generated using different reconstruction configurations, including XL, EMA, MSE, SD21, SD35, and FLUX.1. These synthetic images are aligned with their real counterparts in both pixel and frequency domains, reducing spurious discrepancies while preserving image content. DDA-COCO thus provides a challenging evaluation setting for assessing whether detectors capture intrinsic authenticity cues rather than relying on dataset-specific biases.

EvalGEN [8] is a challenging benchmark designed to test generalization to newly evolved text-to-image generators. It is constructed from prompts aligned with GenEval and includes images generated by five recent models: FLUX, GoT, Infinity, OmniGen, and NOVA, covering both diffusion-based and auto-regressive generation paradigms.

Synthbuster [4] is a diffusion-focused benchmark built from 1K real images in the RAISE-1k dataset and 9K synthetic images generated by nine recent text-to-image models, including Stable Diffusion 1.3/1.4/2/XL, Midjourney v5, DALL·E 2/3, Glide, and Adobe Firefly. The prompts are manually written from descriptions of RAISE images to produce photorealistic synthetic images with similar category diversity rather than pixel-level alignment.

ForenSynths [62] is a classic benchmark for synthetic image detection. It contains synthetic images produced by 13 synthesis methods, covering unconditional GANs, conditional GANs, perceptual-loss-based synthesis, low-level vision tasks, and DeepFake face manipulation. This benchmark is widely used to evaluate generalization across diverse synthesis methods.

AIGCDetection [82] is a comprehensive benchmark introduced by PatchCraft for evaluating AI-generated image detectors across diverse generation sources. It covers 17 evaluation subsets, including GAN-based models such as ProGAN, StyleGAN, BigGAN, CycleGAN, StarGAN, GauGAN, and StyleGAN2; diffusion-based models such as ADM, GLIDE, Stable Diffusion v1.4/v1.5, VQDM, Wukong, and SDXL; commercial generators such as Midjourney and DALL-E 2; and synthetic facial images from WhichFaceIsReal. This diversity makes the benchmark suitable for evaluating broad cross-generator generalization.

Chameleon [69] is a challenging in-the-wild benchmark designed to evaluate whether detectors can recognize AI-generated images that are genuinely difficult for human perception. It contains 14,863 real images and 11,170 AI-generated images, covering four categories: human, animal, object, and scene. The fake images are collected from online AI-art communities such as ArtStation, Civitai, and Liblib, while real images are collected from Unsplash using the same search queries. After resolution, safety, duplication, and CLIP-based filtering, each remaining synthetic image is independently evaluated by two human annotators, and only those misclassified as real by both annotators are retained.

SynthWildx [12] is an in-the-wild benchmark collected from X/Twitter, containing 500 real images and 1,500 synthetic images. The synthetic images are generated by three online text-to-image systems, DALL·E 3, Firefly, and Midjourney, while the real images are collected from the same platform to match the distribution of online visual content. Compared with manually curated benchmarks,

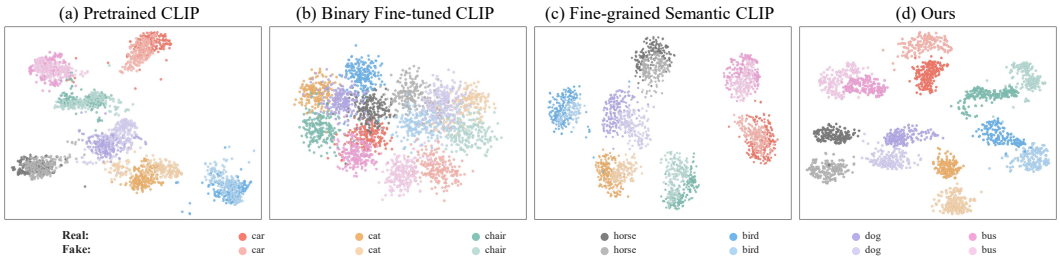


Figure 4: t-SNE visualization of feature distributions under different training strategies on unseen synthetic datasets. Colors denote object categories, while marker styles distinguish authentic and synthetic samples.

Table 4: Ablation study of forensic preference data construction on GenImage.

(a) Effect of image quantity.

Quantity	Category	Acc.
10K	80	93.5
20K	80	94.1
30K	80	94.6
40K	80	**94.9**

(b) Effect of category number.

Category	Quantity	Acc.
20	20K	91.8
40	20K	92.9
60	20K	93.8
80	20K	**94.1**

(c) Effect of two-stream image construction.

Controlled	Observed	Acc.
✓		93.7
	✓	92.2
✓	✓	**94.9**

SynthWildx contains images that have undergone real-world platform-level processing, such as resizing, compression, and format conversion.

WildRF [6] is a real-world deepfake detection benchmark constructed from images shared on Reddit, X, and Facebook. Authentic and synthetic images are manually collected using relevant keywords and hashtags, resulting in diverse visual content, resolutions, formats, and processing histories. The benchmark is particularly suited to assessing detector generalization across social platforms and naturally occurring distribution shifts.

BFree-Online [17] is an in-the-wild benchmark associated with B-Free, comprising authentic and synthetic images that have circulated widely on the Internet. It includes multiple web-scraped versions of the same images collected at different stages of online dissemination. These versions reflect variations in image quality associated with repeated sharing and reposting, providing a realistic setting for evaluating detection robustness to online image degradations.

D Qualitative Analysis

To investigate how different training strategies affect feature discrimination and generalization, we visualize their feature distributions on unseen synthetic datasets using t-SNE. As shown in Fig. 4(a) and (b), binary fine-tuning introduces a certain degree of separation between authentic and synthetic samples but substantially disrupts the semantic structure of pretrained CLIP, causing features from different object categories to become entangled. Fine-grained evidence-text alignment in Fig. 4(c) better preserves semantic distinctions, yet authentic and synthetic samples still exhibit considerable overlap. In contrast, UniFPL maintains well-structured semantic clusters while achieving clearer separation between authentic and synthetic samples, as illustrated in Fig. 4(d). By preserving pretrained semantic structure while establishing the connection between semantic plausibility and image authenticity, our unified training design enhances feature discrimination, leading to stronger detection performance and generalization to unseen generators.

Figure 6 visualizes the MLLM-derived preference scores of randomly sampled image-text candidate sets, where columns represent image candidates and rows denote forensic evidence texts. The heatmaps reveal distinct evidence-specific patterns, with high scores concentrated on corresponding image-text pairs rather than uniformly distributed across candidates. Notably, individual synthetic images may receive high scores across multiple evidence descriptions, suggesting that multiple forensic cues can be relevant to the same image. Meanwhile, the varying score magnitudes reflect differences in the relative diagnostic value of these cues. Such graded evidence relationships motivate

our list-level preference learning design, which preserves the relative strengths of multiple forensic cues to provide more informative supervision for authenticity discrimination.

E More Ablation Results

Table 4 investigates the effects of data quantity, category coverage, and two-branch construction on forensic preference learning. As shown in Table 4(a), increasing the training data from 10K to 40K improves accuracy from 93.5% to 94.9% under fixed category coverage, although the gains gradually diminish as data scale increases. More notably, expanding category coverage from 20 to 80 with the same 20K training samples raises accuracy from 91.8% to 94.1%, indicating that diverse visual content provides substantial benefits beyond simply increasing data volume. We further examine the complementary roles of the two construction branches in Table 4(c). Using the Controlled or Observed branch alone achieves 93.7% and 92.2%, respectively, whereas combining both improves performance to 94.9%. Controlled artifact synthesis provides explicit and verifiable forensic evidence, while observed artifact mining captures artifact patterns present in existing synthetic images. Their combination offers complementary forensic supervision, highlighting the importance of both data diversity and evidence complementarity in effective preference learning.

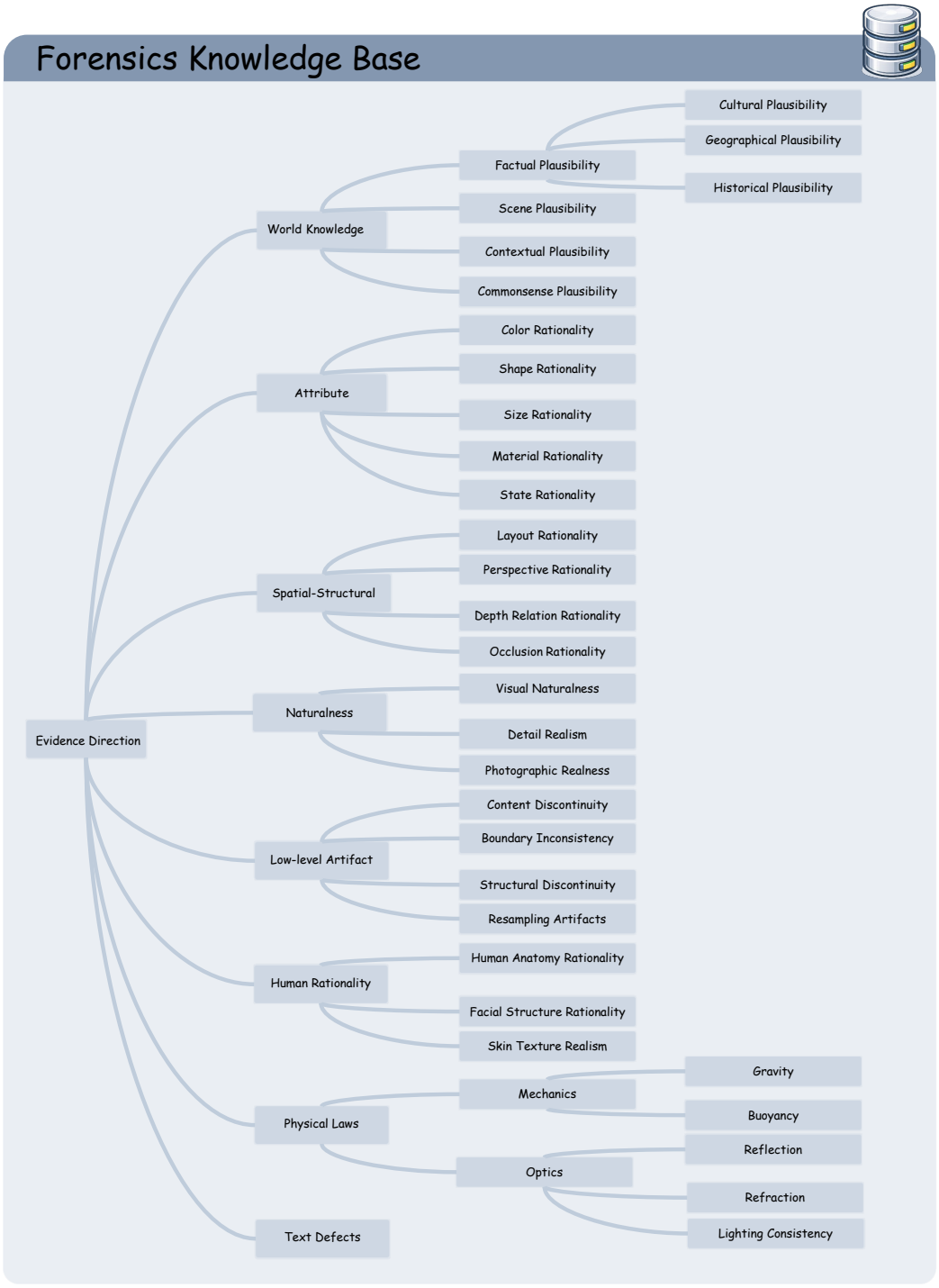


Figure 5: Hierarchical taxonomy of the Forensics Knowledge Base (FKB), comprising eight evidence families and 31 forensic cues used to guide forensic preference data construction.

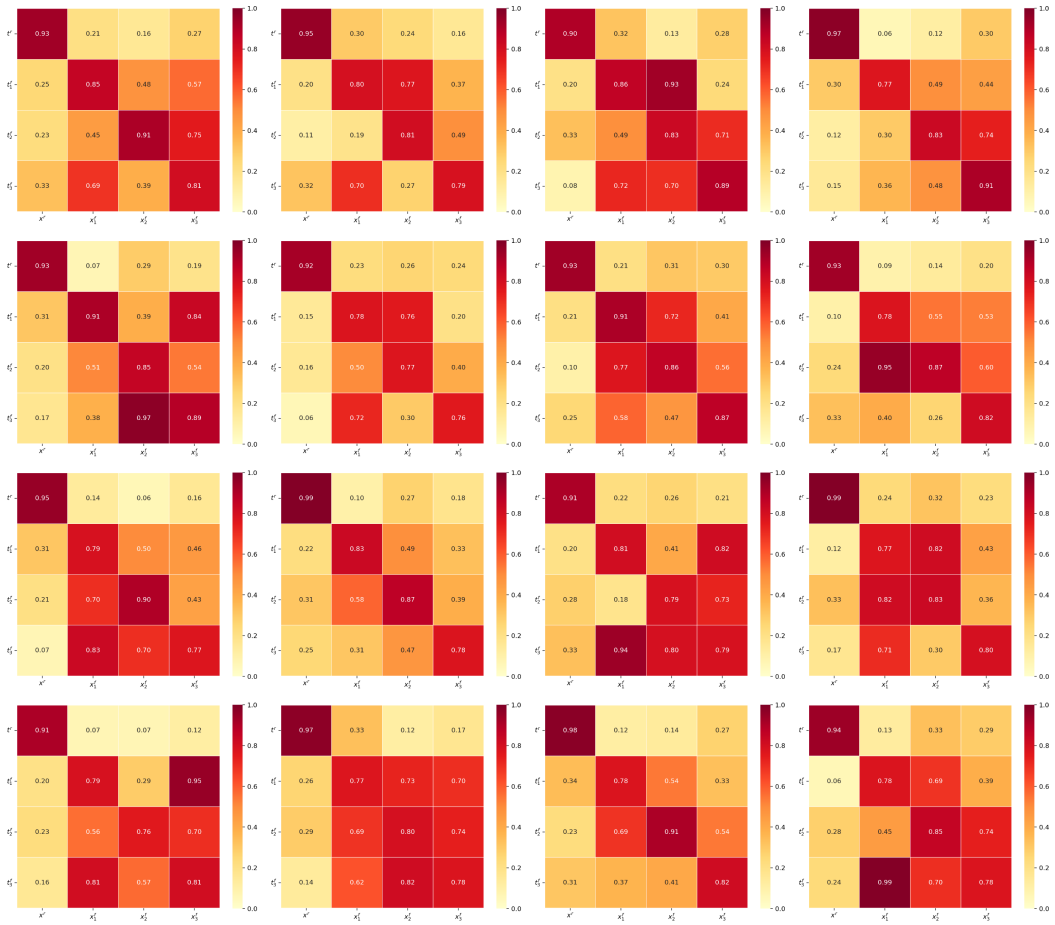


Figure 6: Visualization of MLLM-derived preference scores for randomly sampled image-text preference sets. Columns denote image candidates, and rows denote forensic evidence texts.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have clearly outlined our contributions and scope in both the abstract and the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have provided the limitations of this work in the conclusion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We have provided theoretical derivations and proofs in the methods section of the paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided the data used and implementation details of our method in the experimental section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We commit to releasing all code and data upon paper acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: In the Implementation Details section of the experiments, we provided the training and testing details of our method.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report statistical significance in our experimental results; see Section 4.2 for details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the experimental section, we provided information about the graphics cards and computing frameworks we used.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research adheres to the NeurIPS Code of Ethics in all aspects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the Introduction section, our paper discusses the positive societal impact of our research, namely its ability to mitigate the harm caused by false information.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper does not entail such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have provided a detailed reference list.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have provided detailed documentation in the supplementary materials.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We have detailed the usage of LLMs in the supplementary material.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.