THE *Art* OF RETRIEVAL MEETS LEARNING: RL-GUIDED BEST-OF-N AND AUXILIARY-HEAD FUSION FOR *Style*-TRANSFER DETECTION

Anonymous authors
Paper under double-blind review

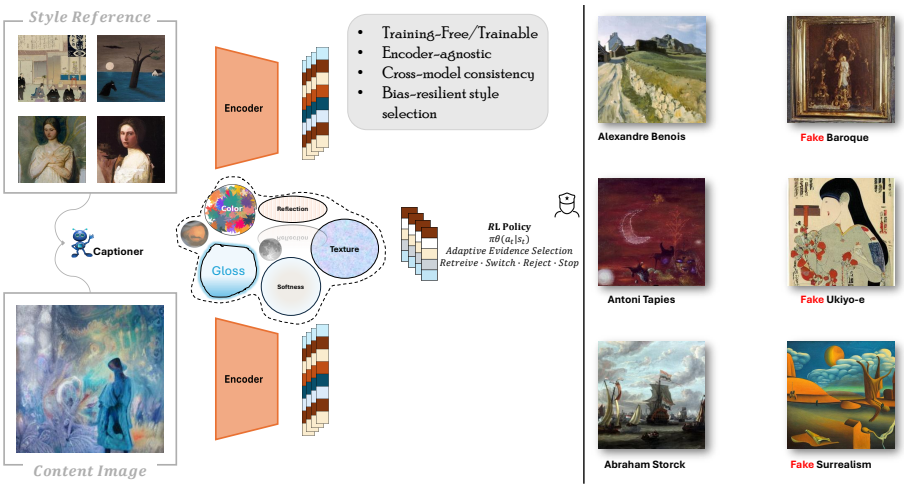

Figure 1: **Style Profiling with Bias-Resilient Reinforcement Retrieval.** We study two alternative pathways for style-aware authenticity detection: a reinforcement-guided retrieval policy that adaptively selects visual-textual evidence, and a trainable auxiliary-head model that combines global style, local artifacts, and cross-region coherence. Given a *Style Reference* set and a *Content Image*, the framework evaluates stylistic consistency and authenticity, identifying compatible style evidence while rejecting inconsistent predictions as **Fake**.

ABSTRACT

The rapid progress of generative imaging has made authenticity detection increasingly difficult, particularly when synthetic content also imitates established artistic styles. We introduce a **reinforcement-guided Best-of-N retrieval framework** that recasts multimodal retrieval as **sequential evidence acquisition**. Rather than aggregating a fixed set of nearest neighbors, a lightweight policy $\pi_\theta(a_t \mid s_t)$ learns when to retrieve additional evidence, switch modality, expand spatial context, reject unreliable candidates, abstain, or stop. Local visual representations are combined with LLM-generated surface descriptors, while the visual and textual encoders remain frozen at inference time. This formulation enables adaptive filtering of noisy or redundant retrieval evidence and improves style-aware representation separation. On AI-ArtBench, the proposed retrieval policy achieves **95.41% mean task accuracy** across authenticity and style prediction, while also showing strong performance on additional real/synthetic and deepfake benchmarks. We further introduce an **alternative trainable Auxiliary-Head Model** that integrates global style, local artifacts, and cross-region coherence through a soft expert-fusion consistency regularizer, achieving **93.96% accuracy**. Together, the two approaches provide distinct operating points for style-aware authenticity detection: **adaptive retrieval with frozen encoders** and **trainable single-pass prediction**.

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

1 INTRODUCTION

The rapid progress of generative models has blurred the boundary between *authentic* and *synthetic* imagery, while modern style-transfer systems can imitate established artistic movements at scale. This creates a broader challenge than binary real/fake detection: applications in digital curation, cultural heritage, and content provenance require reasoning jointly about *authenticity* and *artistic style*. Classical neural style transfer represents style through feature correlations Gatys et al. (2016), while large-scale art-analysis studies have explored style and artist recognition across museum collections Strezoski & Worring (2017); Mensink & Van Gemert (2014). Yet synthetic-image detectors often degrade across generators, post-processing pipelines, and unseen domains Frank et al. (2020); Dong et al. (2022); Yan et al. (2025), and global image descriptors can overlook localized cues such as brushstroke texture, craquelure, and edge irregularities Brendel & Bethge (2019). **This motivates a different question: rather than making a decision from a single embedding or fixed retrieval rule, can a detector learn how to gather, filter, and stop over multimodal evidence?**

We answer this question through two alternative pathways. First, we formulate authenticity and style prediction as *reinforcement-guided Best-of-N retrieval*. Local visual regions and LLM-generated textual descriptors retrieve labeled training exemplars, while a lightweight policy $\pi_\theta(a_t \mid s_t)$ sequentially decides whether to retrieve additional evidence, switch modality, expand context, reject inconsistent candidates, abstain, or stop. The policy is trained using rewards for prediction correctness, visual-textual agreement, retrieval consistency, confidence improvement, and retrieval efficiency. At inference time, no query or test labels are used; predictions rely only on labels attached to exemplars in the fixed training retrieval bank, and the encoders remain frozen. This turns retrieval from a static aggregation mechanism into an adaptive evidence-selection process. Second, we introduce a *trainable auxiliary-head model* for settings where end-to-end supervised adaptation is preferred. Specialized heads model global style, local artifacts, and cross-region coherence, while a learned fusion classifier combines their outputs. A **soft expert-fusion consistency regularizer** encourages agreement with the aggregate expert response without collapsing head specialization. The two approaches therefore represent distinct operating points: adaptive retrieval with frozen encoders versus trainable single-pass fusion. We further provide an expert visual-arts assessment in B.8, offering a qualitative perspective. **Contributions.**

Reinforcement-guided Best-of-N retrieval. We introduce a learned sequential policy $\pi_\theta(a_t \mid s_t)$ that **adaptively gathers, filters, and stops** over visual and textual retrieval evidence. The encoders remain frozen, while a lightweight policy is trained with rewards for correctness, multimodal agreement, retrieval consistency, confidence gain, and retrieval cost. This turns Best-of-N retrieval from a **fixed threshold rule** into a **learned decision process**.

Theoretical bias reduction through dual visual-textual evidence. We combine local visual regions with LLM-generated textual descriptors and provide theoretical and empirical evidence that cross-modal agreement **reduces visual encoder bias**, improves style-aware separability, and strengthens robustness under distribution shift.

Auxiliary-head for calibrated fusion. A trainable auxiliary-head model with specialized **global style**, **local artifact**, and **cross-region coherence** heads. A learned fusion classifier, supported by a **soft expert-fusion consistency loss**, improves calibration while preserving head specialization.

2 RELATED WORK

Style transfer and artistic analysis. Classical neural style transfer Gatys et al. (2016) formalized the separation of *content* and *style* by matching second-order feature statistics (Gram matrices) over convolutional activations. Specifically, $G^{(l)} = F^{(l)}F^{(l)\top}$, where $F^{(l)} \in \mathbb{R}^{C_l \times (HW)}$ denotes activations at layer l. This perspective, viewing style as second-order correlations, motivated pipelines for style recognition and artist attribution that emphasize mid-level texture and palette cues rather than object identity. Museum-scale resources and multi-task formulations enabled joint prediction of style, artist, period, and genre Strezoski & Worring (2017); Mensink & Van Gemert (2014); Van Noord et al. (2015); Kazoom et al. (2025a). Recent systems adopt transformer backbones and ensemble strategies to improve generalization across collections and digitization conditions Joshi et al. (2020); Menis-Mastromichalakis et al. (2020). Our grid-level retrieval formulation is motivated by the observation that discriminative style cues often reside in *local* brushstroke micro-textures and pigment

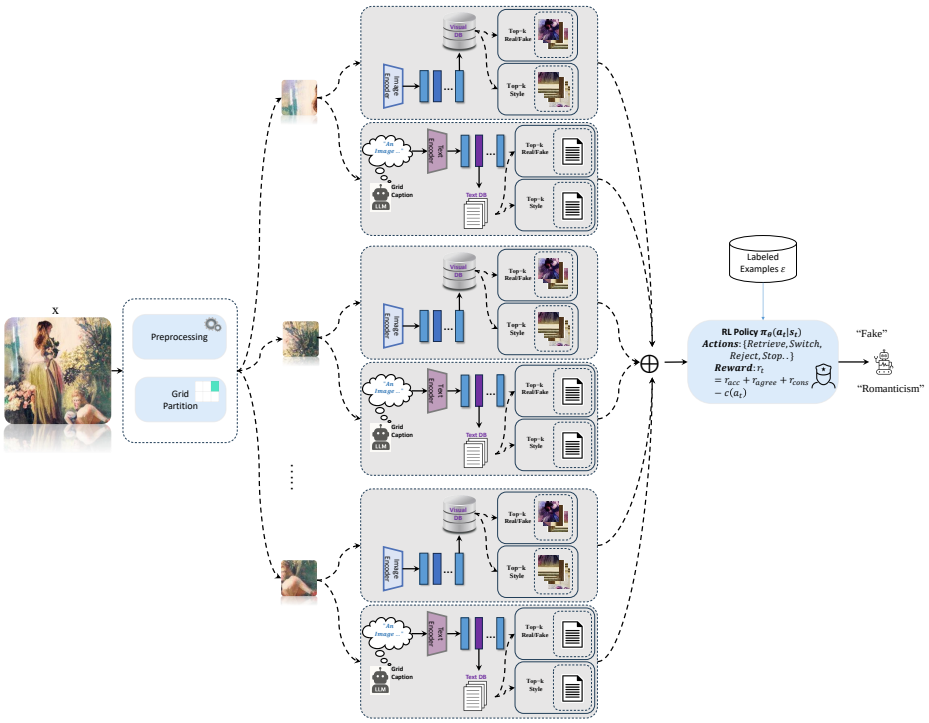

Figure 2: Retrieval-augmented authenticity and style attribution pipeline. The input image is partitioned into a $g \times g$ grid, where local visual embeddings query a labeled region-level database, while a parallel caption pathway retrieves textual evidence. A learned policy $\pi_\theta(a_t \mid s_t)$ adaptively gathers, filters, and stops over multimodal evidence to predict authenticity and style.

patterns. Aggregating evidence from subregions is therefore more discriminative than relying on a single global descriptor, consistent with findings that bag-of-patches models can match or surpass full-context CNNs when local statistics dominate Brendel & Bethge (2019).

Best-of-N selection and reinforcement-guided retrieval. Recent alignment literature has highlighted the effectiveness of *Best-of-N (BoN)* inference strategies, where a model generates N candidate outputs and a reward model selects the maximizer $\hat{y} = \arg\max_{i \le N} R(y_i)$. Sessa et al. (2025) show that *BoN* distillation can transfer the benefits of reinforcement-trained reward models into deterministic policies. Jinnai et al. (2025) formulate *BoN* under a minimum Bayes risk objective and propose regularized selection to stabilize reward-based rules. Verdun et al. (2025) extend this view with *soft BoN*, replacing hard maximization with a reward-weighted distribution that mitigates overconfidence in noisy reward landscapes. Ichihara et al. (2025); Kazoom et al. (2025c) provide a systematic evaluation of *BoN* strategies, demonstrating that inference-time selection can rival or surpass RL-trained policies when the reward model is strong. In contrast to standard *BoN*, where candidates are sampled model outputs, our candidates are retrieved visual and textual exemplars from a labeled bank. Rather than selecting the highest-scoring candidate with a fixed rule, we train a sequential policy $\pi_\theta(a_t \mid s_t)$ that adaptively gathers, filters, and stops over multimodal retrieval evidence. This connects *BoN* selection with RL: the reward is defined over prediction correctness during training, visual-textual agreement, retrieval consistency, confidence improvement, and retrieval cost, while the action space controls evidence acquisition and candidate filtering.

Training-based detection, robustness, and multi-branch fusion architectures. GAN- and diffusion-generated imagery increasingly relies on frequency and spectrum-domain artifacts Frank et al. (2020); Dong et al. (2022); Marra et al. (2019); Yan et al. (2025); Kazoom et al. (2025b). Frequency-based detectors analyze magnitude spectra $|\mathcal{F}(I)|$ in the DCT/FFT domains to identify deviations from natural-image statistics Dong et al. (2022), motivating sanity-check paradigms Yan et al. (2025). Fingerprint-style analyses trace generator-specific patterns Marra et al. (2019) but remain fragile under compression and resizing Kazoom et al. (2024). Recent approaches employ adap-

tive multi-branch networks Meng et al. (2025), local-global feature integration Ding et al. (2024), and RGB-text fusion with MFF-Net Zhao et al. (2021). Our auxiliary-head design aligns with these directions, specializing heads for global style, local artifacts, and cross-region coherence, and fusing them through a calibrated model.

Retrieval-guided and policy-guided paradigms. Recent detection approaches reduce reliance on end-to-end retraining by exploiting signal-level irregularities, retrieval cues, or frozen pretrained representations. Frequency-based methods Frank et al. (2020); Dong et al. (2022) and calibration-style tests Yan et al. (2025) demonstrate that detectors can operate without updating large visual backbones by checking natural-image statistics or consistency constraints. Retrieval-based approaches ground decisions in evidence from labeled banks rather than relying only on learned discriminative boundaries, improving robustness to distribution shift. Prior work highlights both the promise and pitfalls of these methods: frequency detectors often struggle under processing and cross-generator transfer Dong et al. (2022), motivating sanity-check protocols to avoid spurious correlations. In parallel, retrieval-guided decision rules extend RAG principles to vision, where embedding-based retrieval supports decision making through evidence aggregation. In forgery and synthetic media detection, retrieval-driven reasoning can improve robustness over fully end-to-end models.

Our method follows this retrieval-guided direction but replaces a fixed aggregation rule with a learned sequential policy. The visual and textual encoders remain frozen, while the policy $\pi_\theta(a_t \mid s_t)$ is trained to decide how much evidence to gather, which modality to trust, which candidates to reject, and when to stop. This preserves the interpretability and flexibility of retrieval-based inference while allowing adaptive evidence selection under uncertainty.

Backbone Encoders for Multimodal Style and Authenticity Retrieval. Modern retrieval pipelines rely on embedding functions that map inputs to a metric space where semantically related samples are close and unrelated ones are well separated. Contrastive learning provides a principled framework by aligning *positives* and repelling *negatives* via temperature-scaled softmax losses Oord et al. (2018); Chen et al. (2020); He et al. (2020); Khosla et al. (2020). Given encoder $f(\cdot)$ and projection head $g(\cdot)$, the InfoNCE loss is $\mathcal{L}_{\text{InfoNCE}}(i,j) = -\log \frac{\exp\left(s(\mathbf{z}_i,\mathbf{z}_j)/\tau\right)}{\sum_{k\neq i}\exp\left(s(\mathbf{z}_i,\mathbf{z}_k)/\tau\right)}$, where $s(\mathbf{z}_a,\mathbf{z}_b) = \mathbf{z}_a^\top \mathbf{z}_b$ and $\tau > 0$. SimCLR exploits large-batch negatives Chen et al. (2020), MoCo employs a momentum encoder He et al. (2020), and supervised contrastive learning sharpens intra-class compactness Khosla et al. (2020). Beyond instance discrimination, CLIP introduces symmetric cross-modal contrastive learning to align images and text Radford et al. (2021), enabling strong zero-shot retrieval. Self-distillation methods such as DINO further structure representations under multi-view augmentations Caron et al. (2021), while angular-margin objectives enhance inter-class separation Deng et al. (2019). Recent multimodal backbones include **CLIP** Radford et al. (2021) for joint image-text embedding, **DINOv2** Oquab et al. (2023) for self-supervised structure-aware visual features, and **SigLIP2** Tschannen et al. (2025), improves localization and dense feature quality via sigmoid-based contrastive objectives. In our framework, these encoders are kept frozen and used to construct visual and textual retrieval states for the RL policy. Paired text encoders map captions into the joint embedding space, enabling cross-modal retrieval and reducing visual encoder bias.

3 METHODOLOGY

The two approaches target different operating settings: adaptive retrieval with frozen encoders versus end-to-end trainable single-pass prediction. First, we introduce a reinforcement-guided retrieval framework based on *Best-of-N* evidence selection (Section 3.1), where frozen visual and textual encoders provide candidate evidence and a lightweight policy adaptively gathers, filters, and aggregates retrieved training exemplars. Second, we introduce a supervised auxiliary-head model (Section 3.2) that jointly learns specialized global-style, local-artifact, and cross-region-coherence representations through a trainable fusion mechanism.

3.1 REINFORCEMENT-GUIDED BEST-OF-N RETRIEVAL

We partition each image into a $g \times g$ grid, extract localized visual embeddings, and retrieve the top-k nearest labeled exemplars from a reference bank. In parallel, we enrich the retrieval bank with *textual* descriptors: for every image, an LLM-generated caption is computed and stored along-

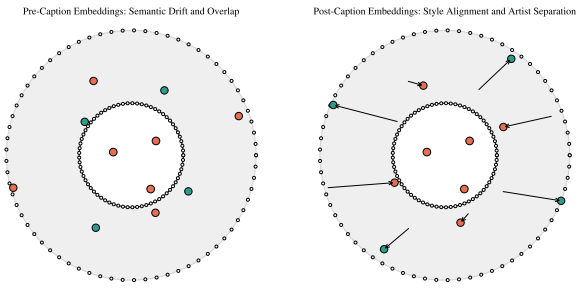

Figure 3: **Embedding geometry before and after style-aware captioning. Left:** Pre-caption embeddings exhibit semantic drift and overlap, with same-artist samples scattered and mixed with others. **Right:** After applying texture captioning, same-artist embeddings realign toward a consistent core while different-artist embeddings are pushed outward, improving style and artist separation.

side its region embeddings. During inference, the query image is captioned in the same manner, enabling text-to-text retrieval in parallel with image-to-image retrieval. The resulting visual and textual neighbors provide complementary evidence for authenticity and style prediction. Unlike a fixed retrieval rule that directly aggregates all retrieved labels by majority or similarity-weighted voting, we formulate evidence aggregation as a finite-horizon sequential decision problem. For each query image, the agent observes a retrieval state s_t containing the current visual and textual evidence, modality-specific label distributions, confidence margins, agreement scores, and retrieval history. It then samples an action from a learned policy $\pi_\theta(a_t \mid s_t)$. The action space includes retrieving additional neighbors, switching between visual and textual evidence, expanding local patch context, rejecting unstable candidates, abstaining, or stopping with a prediction. After each action, the evidence set is updated and a new state s_{t+1} is formed. The policy is trained with rewards that combine prediction correctness on labeled training data, visual-textual agreement, retrieval consistency, confidence improvement, and a small cost for unnecessary retrieval operations. At inference, ground-truth labels are not available and the visual/textual encoders remain frozen; the learned policy adaptively determines which retrieved evidence to retain and when to stop. Figure 1 provides an overview of the encoder-agnostic style profiling framework, where visual and textual signals are projected into a style-aware evidence space and aggregated by the learned reinforcement policy to obtain bias-resilient style attributions.

Retrieval Bank Construction. We construct the retrieval bank $\mathcal{B}$ exclusively from labeled training images. Each image I is partitioned into a $g \times g$ grid, and each region $c \in \mathcal{C}$ is mapped to a normalized visual embedding $\mathbf{e}_{I,c} = \phi(I_c)/\|\phi(I_c)\|_2$ using a frozen encoder ϕ. In parallel, an LLM-generated surface descriptor is encoded as $\mathbf{t}_I = \psi(\mathsf{cap}(I))/\|\psi(\mathsf{cap}(I))\|_2$. Each bank entry stores the regional visual embeddings, text embedding, and corresponding training authenticity and style labels. No validation or test samples, their labels, or derived variants are included in $\mathcal{B}$.

Sequential RL Inference. Given a query image $I^\star$, we partition it into a $g \times g$ grid and compute normalized regional visual embeddings $\mathbf{z}_c = \phi(I^\star_c)/\|\phi(I^\star_c)\|_2$. In parallel, an LLM-generated caption is encoded as

$$\mathbf{t}^\star = \psi(\mathsf{cap}(I^\star))/\|\psi(\mathsf{cap}(I^\star))\|_2 \tag{1}$$

Visual and textual nearest neighbors form the initial evidence state s_0. At each step, the policy $\pi_\theta(a_t \mid s_t)$ observes the retained evidence, label distributions, confidence, visual-textual agreement, and retrieval history, and decides whether to retrieve more evidence, switch modality, expand context, reject candidates, abstain, or stop. Upon stopping, predictions are obtained by weighted voting over the retained training-bank evidence J_t: $\hat{y} = \arg\max_{y \in \mathcal{Y}} \sum_{j \in J_t} w_j \mathbb{I}[y_j = y]$. The model abstains if ABSTAIN is selected or if the final confidence margin falls below threshold. No query or test labels are used at inference; only labels attached to retrieved training-bank exemplars contribute to the vote, and the visual and textual encoders remain frozen. Algorithm 1 summarizes the complete inference procedure.

Algorithm 1 Reinforcement-Guided Multimodal Retrieval Inference

Input: query image $I^\star$, training bank $\mathcal{B}$, encoders ϕ, ψ, policy π_θ, initial neighbors k, horizon T, threshold τ
Output: authenticity $\hat{a}$, style $\hat{s}$, or ABSTAIN
$\{\mathbf{z}_c\}_{c\in\mathcal{C}} \leftarrow \textsc{EncodeRegions}(I^\star, \phi)$
$\mathbf{t}^\star \leftarrow \textsc{EncodeCaption}(I^\star, \psi)$
$\mathcal{E}_0 \leftarrow \textsc{Retrieve}(\mathcal{B}, \{\mathbf{z}_c\}, \mathbf{t}^\star, k)$
$s_0 \leftarrow \textsc{BuildState}(\mathcal{E}_0)$
for $t = 0, \ldots, T-1$ **do**
  $a_t \sim \pi_\theta(\cdot \mid s_t)$
  if $a_t = \textsc{Abstain}$ **then**
    return ABSTAIN
  else if $a_t = \textsc{Stop}$ **then**
    break
  end if
  $(\mathcal{E}_{t+1}, s_{t+1}) \leftarrow \textsc{Step}(\mathcal{B}, \mathcal{E}_t, s_t, a_t)$
end for
$(\hat{a}, \hat{s}) \leftarrow \textsc{WeightedVote}(\mathcal{E}_t)$
if $\textsc{Confidence}(\mathcal{E}_t) < \tau$ **then**
  return ABSTAIN
end if
return $(\hat{a}, \hat{s})$

3.1.1 Theoretical Perspective: Geometry and Bias Reduction

Image encoders trained on natural-image corpora primarily organize representations around semantic content, which can be suboptimal for artistic style. Two images with similar subjects but different styles may therefore lie close in the visual embedding space, while works sharing a style but depicting different subjects may drift apart. Let $z_v = \phi(I)$ denote the visual embedding of image I. We refer to this content-driven overlap as *semantic drift*. Our framework addresses this limitation through complementary local and textual evidence. Each image is decomposed into regions $\{I_c\}_{c\in\mathcal{C}}$, while an LLM-generated descriptor provides a global textual representation $z_t = \psi(\mathrm{cap}(I))$. The resulting evidence

$$E(I) = \{\phi(I_c), \psi(\mathrm{cap}(I))\} \tag{2}$$

captures both localized texture and material cues and higher-level stylistic descriptions. Under the idealized assumption that retained evidence sources are independent and individually correct with probability $p > \frac{1}{2}$, majority aggregation satisfies

$$\Pr[\hat{y} \neq y] \leq \exp(-2|\mathcal{C}|(p - \tfrac{1}{2})^2), \tag{3}$$

illustrating how multiple weak local cues can collectively form a more reliable decision. The key role of reinforcement learning is to make this evidence aggregation *adaptive*. Rather than treating every retrieved candidate equally, the policy constructs a sequence $E_0, E_1, \ldots, E_T$, selectively gathering additional evidence and removing inconsistent candidates according to confidence, cross-modal agreement, and retrieval consistency. Multimodal retrieval can be expressed as $d_{\text{multi}} = \lambda d(z^a_{v,c}, z^b_{v,c}) + (1-\lambda) d(z^a_t, z^b_t)$, with $\lambda \in [0, 1]$. Thus, when visual embeddings collapse around shared semantic content, textual evidence can preserve stylistic separation, while the policy learns when that additional signal should be trusted. Figure 3 illustrates this geometric effect. The full architecture is provided in A.5, with additional analysis in B.2 and B.1.

3.2 Auxiliary-Head Style-Transfer Model

A multi-head architecture that detects whether an input image has undergone style transfer by combining complementary global, local, and relational cues. Let $x \in \mathbb{R}^{H\times W\times 3}$ denote an RGB image and $y \in \{0, 1\}$ the ground truth label.

Pretrained Vision Encoder. A pretrained vision encoder f_ϕ maps the image to a dense feature tensor $F = f_\phi(x) \in \mathbb{R}^{H'\times W'\times C}$, which is shared by all heads. For the style-coherence head we

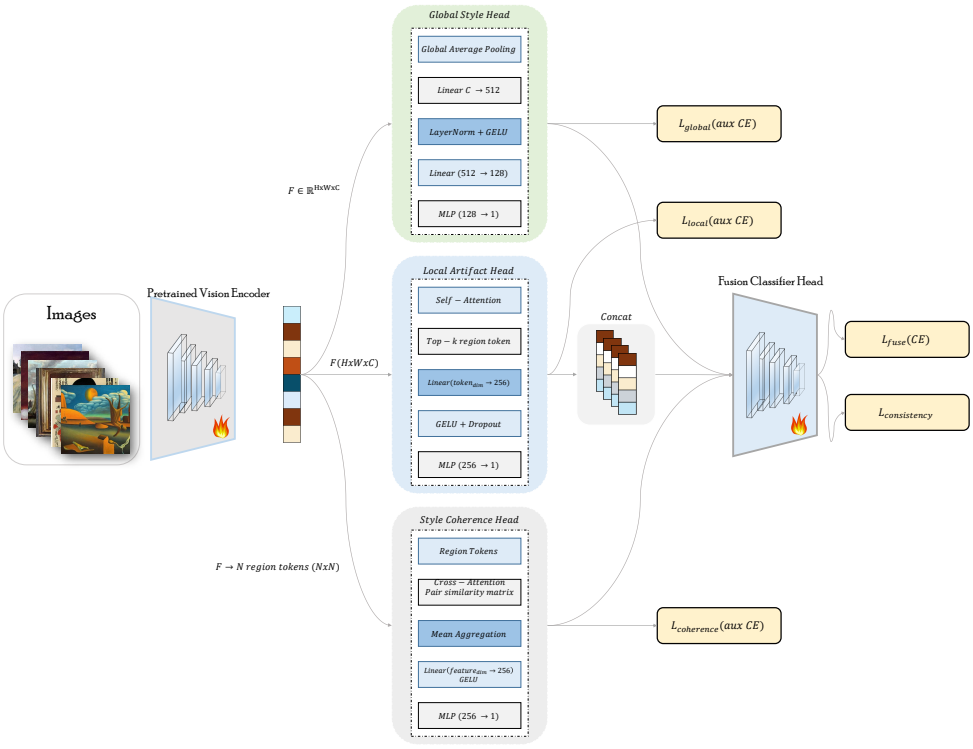

Figure 4: Auxiliary-head style-transfer detection model. A pretrained encoder produces shared features and region tokens, which feed three specialized heads for global style realism, local artifact cues, and cross-region coherence. Head logits are trained with auxiliary losses and fused by a classifier supervised by the main loss, with a soft consistency term encouraging agreement.

also construct N region tokens by spatially pooling F over a coarse grid: $R = g(F) \in \mathbb{R}^{N \times C}$, where $g(\cdot)$ denotes non-overlapping average pooling over an N cell partition of the feature map.

Global Style Head. The global style head focuses on holistic style, tone, and composition. It first aggregates F by global average pooling, $\bar{f} = \frac{1}{H'W'} \sum_{h,w} F_{h,w,:} \in \mathbb{R}^C$, then passes the result through a small multilayer perceptron to produce a scalar logit z_g: $z_g = h_g(\bar{f})$, where h_g denotes the sequence of linear, normalization, and nonlinearity layers shown in Figure 4. The corresponding probability is $p_g = \sigma(z_g)$, with σ the logistic sigmoid. This head learns whether the global appearance of the image is stylistically realistic or indicative of transfer.

Local Artifact Head. The local artifact head inspects spatial details for texture and boundary artifacts. It applies a lightweight self-attention module $a(\cdot)$ to F and selects the top k attended region tokens $T = \mathrm{TopK}\big(a(F)\big) \in \mathbb{R}^{k \times C}$. The tokens T are then projected and pooled by a small network h_l to produce a logit $z_l = h_l(T), \qquad p_l = \sigma(z_l)$.

Style Coherence Head. The style coherence head operates on the region tokens R in order to reason about cross-region stylistic consistency. A cross-attention transformer $c(\cdot)$ produces pairwise interactions among regions, from which we derive a similarity matrix $S \in \mathbb{R}^{N \times N}$. A mean aggregation $m(\cdot)$ compresses S to a vector summary v_c, $v_c = m\big(S(R)\big)$, which is fed to a final multilayer perceptron h_c to obtain $z_c = h_c(v_c), \qquad p_c = \sigma(z_c)$. This head learns whether regional variations in style are semantically plausible for a real scene or instead follow typical style-transfer patterns such as uneven stylization strength or style leakage across object boundaries.

Fusion Classifier Head. The three heads produce scalar logits z_g, z_l, and z_c. The fusion receives their concatenation $u = [z_g, z_l, z_c] \in \mathbb{R}^3$, and maps it through a small fusion network h_f to obtain the final logit $z_f = h_f(u), \qquad p_f = \sigma(z_f)$, which is the main style-transfer prediction of the model. The fusion head learns how to weight the three evidence sources depending on the image.

Auxiliary Cross-Entropy Losses. Each head is supervised with a binary cross-entropy loss so that it independently learns to discriminate real from style-transferred images. For a generic head with probability p the loss is $\mathcal{L}_{\text{CE}}(p, y) = -\big[y \log p + (1 - y) \log(1 - p)\big]$. We define $\mathcal{L}_{\text{global}} = \mathcal{L}_{\text{CE}}(p_g, y)$, $\mathcal{L}_{\text{local}} = \mathcal{L}_{\text{CE}}(p_l, y)$, $\mathcal{L}_{\text{coh}} = \mathcal{L}_{\text{CE}}(p_c, y)$, and the main fusion loss $\mathcal{L}_{\text{fuse}} = \mathcal{L}_{\text{CE}}(p_f, y)$. These auxiliary objectives encourage each head to become a useful expert: the global head focuses on holistic style realism, the local head on artifact evidence, and the coherence head on cross-region consistency.

Consistency Regularization. To stabilize training and encourage collaboration between heads, we introduce a consistency loss that gently aligns the fusion decision with the average opinion of experts. Let $\bar{p} = \frac{1}{3}(p_g + p_l + p_c)$ be the mean head probability. The consistency penalty is defined as $\mathcal{L}_{\text{consistency}} = \big\|p_f - \bar{p}\big\|_2^2$. This term does not force exact agreement; instead it discourages pathological situations in which the fusion model systematically ignores heads or produces highly unstable outputs. When the three heads strongly agree, the penalty encourages p_f to stay close to that consensus. When they disagree, the fusion model can still override them, but only when supported by the supervised cross-entropy signal. $\mathcal{L}_{\text{consistency}}$ improves calibration and reduces variance in the final predictions. Total loss is a weighted sum of the four cross-entropy terms and the consistency regularizer:

$$\mathcal{L} = \lambda_{\text{fuse}}\mathcal{L}_{\text{fuse}} + \lambda_{\text{global}}\mathcal{L}_{\text{global}} + \lambda_{\text{local}}\mathcal{L}_{\text{local}} + \lambda_{\text{coh}}\mathcal{L}_{\text{coh}} + \lambda_{\text{cons}}\mathcal{L}_{\text{consistency}}, \tag{4}$$

with λ_{fuse} typically set to one and the auxiliary weights chosen smaller. This encourages each head to specialize in its respective view of the evidence while the fusion head learns a robust, calibrated detector.

4 RESULTS

We evaluate both frameworks on *AI-ArtBench* Silva et al. (2024), which contains real and AI-generated artworks spanning multiple artistic styles, as well as on additional real/fake and deepfake benchmarks Zhu et al. (2023); Bird & Lotfi (2024); Rossler et al. (2019); Dolhansky et al. (2020); Li et al. (2020). For the retrieval framework, each image is partitioned into a $g \times g$ grid to capture localized stylistic cues and micro-textures. We compare three frozen encoder families: **SigLIP**, **CLIP** (ViT-B/16), and **DINOv2**. The retrieval bank is constructed exclusively from the training split, while validation data is used only for policy and stopping-hyperparameter selection. At test time, no query or test labels are used, and the encoders, policy, and retrieval bank remain fixed. The auxiliary-head model is trained separately as a supervised classifier using the same dataset splits. Table 2 reports the AI-ArtBench results, while Table 1 evaluates generalization on GenImage Zhu et al. (2023), FaceForensics++ Rossler et al. (2019), DFDC Dolhansky et al. (2020), and Celeb-DF Li et al. (2020). Qualitative examples are provided in B.9.

Table 1: Performance of publicly available real/fake image datasets. SigLIP encoder with k=9 nearest neighbours and a 3×3 grid over the regular RGB images. In addition, we report results for the proposed trainable Auxiliary-Head Model.

Method	Dataset	Task Domain	Accuracy ↑	F1 ↑	Precision ↑	Recall ↑
RL Best-of-N	Zhu et al. (2023)	General real vs AI image detection RGB	95.42	95.43	95.10	95.76
	Bird & Lotfi (2024)	Real vs synthetic CIFAR-style images	96.18	96.16	96.40	95.92
	Rossler et al. (2019)	Face manipulation / deepfake forensics	92.77	92.79	92.10	93.49
	Dolhansky et al. (2020)	DeepFake detection challenge (video frames)	89.31	89.33	88.72	89.94
	Li et al. (2020)	High-quality celebrity deepfakes	93.64	93.62	93.20	94.05
Auxiliary-Head Model	Zhu et al. (2023)	General real vs AI image detection RGB	91.81	91.82	92.05	91.60
	Bird & Lotfi (2024)	Real vs synthetic CIFAR-style images	95.89	95.89	96.12	95.66
	Rossler et al. (2019)	Face manipulation / deepfake forensics	90.71	90.72	90.96	90.48
	Dolhansky et al. (2020)	DeepFake detection challenge (video frames)	87.21	87.22	87.45	86.99
	Li et al. (2020)	High-quality celebrity deepfakes	93.14	93.15	93.39	92.91

Table 2: Comparison of our method's accuracy on the AI-ArtBench dataset against prior approaches.

Method	Silva et al. (2024)	Zhang et al. (2025)	Ramanathan et al. (2025)	Li & Stamp (2025)	**Ours** (Auxiliary-Head)	**Ours** (RL Best-of-N)
Accuracy ↑	86.00	80.00	75.00	82.00	**93.96**	**95.41**

Table 3 reports performance across preprocessing methods, neighborhood sizes, and encoders, where k denotes top-k retrieval. Textual evidence consistently improves over non-textual, confirming that caption-based retrieval complements visual embeddings. RGB and grayscale preprocessing perform best, while Gaussian and JPEG degradations reduce robustness. The RL-guided Best-of-N policy reaches **95.41%** accuracy, outperforming prior reported benchmarks, while the auxiliary-head model achieves **93.96%** as a fully trainable alternative.

Table 3: Accuracy across preprocessing methods, grid resolutions, initial neighborhood sizes, text usage, and encoders. k denotes the top-k retrieval setting, and T/NT indicate the use of Text or No Text in the retrieval process.

Encoder	Pre-Processing	K	Real/Fake Acc. ↑						Style Acc. ↑						Mean Task Acc. ↑					
			3×3		4×4		5×5		3×3		4×4		5×5		3×3		4×4		5×5	
			T	NT	T	NT	T	NT	T	NT	T	NT	T	NT	T	NT	T	NT	T	NT
SigLIP	RGB	5	96.11	91.91	94.21	88.81	92.47	85.87	92.03	87.83	90.91	85.51	90.34	83.74	**94.07**	89.87	92.94	87.16	92.28	84.81
		7	96.73	92.53	94.84	89.44	93.09	86.49	92.74	88.54	91.59	86.19	90.99	84.39	**94.85**	90.53	93.57	87.81	92.91	85.44
		9	97.41	93.21	95.48	90.08	93.72	87.12	93.41	89.21	92.29	86.89	91.69	85.09	**95.41**	91.21	94.24	88.48	93.55	86.11
	Grayscale	5	94.79	90.59	93.63	88.23	93.01	86.41	90.82	86.62	89.64	84.24	89.01	82.41	**92.80**	88.61	91.64	86.23	91.01	84.41
		7	95.41	91.21	94.26	88.86	93.62	87.02	91.52	87.32	90.36	84.96	89.72	83.12	**93.47**	89.26	92.31	86.91	91.67	85.07
		9	96.08	91.88	94.91	89.51	94.28	87.68	92.19	87.99	91.03	85.63	90.38	83.78	**94.13**	89.94	92.97	87.57	92.33	85.73
	Gaussian	5	86.31	82.11	85.14	79.74	84.52	77.92	82.28	78.08	81.07	75.67	80.43	73.83	**84.30**	80.09	83.09	77.70	82.48	75.88
		7	86.92	82.72	85.77	80.37	85.12	78.52	82.89	78.69	81.72	76.32	81.06	74.46	**84.90**	80.70	83.75	78.34	83.09	76.49
		9	87.57	83.37	86.41	81.01	85.78	79.18	83.53	79.33	82.36	76.96	81.72	75.12	**85.53**	81.35	84.36	78.98	83.75	77.15
	JPEG	5	74.33	70.13	73.18	67.78	72.60	66.00	69.92	65.72	68.78	63.38	68.19	61.59	**72.12**	67.92	70.98	65.58	70.40	63.80
		7	74.97	70.77	73.83	68.43	73.24	66.64	70.63	66.43	69.47	64.07	68.92	62.32	**72.80**	68.60	71.65	66.25	71.08	64.48
		9	75.62	71.42	74.46	69.06	73.90	67.30	71.28	67.08	70.12	64.72	69.58	62.98	**73.45**	69.25	72.29	66.89	71.74	65.14
CLIP	RGB	5	88.91	84.91	87.75	82.61	87.02	80.89	84.63	80.65	83.47	78.51	82.71	76.78	**86.77**	82.78	85.61	80.56	84.87	78.84
		7	89.47	85.47	88.31	83.17	87.59	81.46	85.28	81.30	84.12	79.14	83.36	77.39	**87.42**	83.44	86.26	81.20	85.47	79.93
		9	90.02	86.02	88.86	83.72	88.14	82.01	85.83	81.85	84.67	79.69	83.90	77.94	**87.98**	84.00	86.83	81.76	86.02	80.48
	Grayscale	5	87.52	83.42	86.38	81.19	85.71	79.48	83.22	79.22	82.10	77.06	81.41	75.35	**85.37**	81.32	84.24	79.12	83.56	77.41
		7	88.09	83.99	86.93	81.76	86.28	80.05	83.84	79.84	82.69	77.70	82.02	75.98	**86.00**	81.92	84.85	79.73	84.15	78.01
		9	88.67	84.57	87.52	82.34	86.82	80.63	84.42	80.42	83.26	78.28	82.61	76.56	**86.60**	82.50	85.44	80.32	84.72	78.59
	Gaussian	5	79.14	75.04	77.98	72.78	77.32	71.09	74.91	70.92	73.73	68.78	73.09	66.99	**77.03**	72.98	75.85	70.78	75.20	68.99
		7	79.73	75.63	78.57	73.39	77.88	71.68	75.53	71.54	74.37	69.34	73.72	67.62	**77.65**	73.58	76.48	71.34	75.80	69.65
		9	80.31	76.21	79.14	73.97	78.47	72.26	76.11	72.12	74.95	69.92	74.31	68.20	**78.22**	74.15	77.05	71.92	76.39	70.24
	JPEG	5	70.73	66.63	69.59	64.39	68.97	62.72	66.47	62.47	65.33	60.34	64.71	58.59	**68.60**	64.55	67.46	62.36	66.84	60.66
		7	71.29	67.19	70.16	64.96	69.51	63.29	67.03	63.03	65.89	60.92	65.26	59.17	**69.18**	65.08	68.03	62.97	67.39	61.23
		9	71.84	67.74	70.70	65.51	70.08	63.85	67.57	63.57	66.43	61.46	65.82	59.72	**69.74**	65.63	68.58	63.52	67.95	61.78
DINOv2	RGB	5	86.41	82.31	85.26	80.07	84.60	78.36	82.07	78.08	80.91	75.92	80.25	74.18	**84.24**	80.19	83.08	77.99	82.43	76.27
		7	86.97	82.87	85.82	80.63	85.17	78.92	82.63	78.64	81.47	76.48	80.82	74.74	**84.80**	80.75	83.63	78.55	82.99	76.83
		9	87.53	83.43	86.39	81.19	85.71	79.48	83.20	79.21	82.04	77.05	81.39	75.31	**85.36**	81.32	84.20	79.12	83.55	77.40
	Grayscale	5	85.12	81.02	83.97	78.79	83.34	77.09	80.76	76.76	79.61	74.62	78.96	72.89	**82.94**	78.89	81.79	76.70	81.15	74.99
		7	85.69	81.59	84.52	79.36	83.90	77.66	81.36	77.36	80.21	75.22	79.57	73.49	**83.52**	79.45	82.36	77.28	81.73	75.61
		9	86.22	82.12	85.08	79.88	84.45	78.18	81.91	77.91	80.75	75.76	80.12	74.03	**84.06**	79.99	82.91	77.82	82.28	76.16
	Gaussian	5	76.73	72.63	75.56	70.40	74.94	68.70	72.37	68.37	71.22	66.24	70.59	64.49	**74.55**	70.50	73.39	68.29	72.77	66.59
		7	77.28	73.18	76.12	70.97	75.51	69.27	72.95	68.95	71.79	66.81	71.16	65.06	**75.11**	71.06	73.96	68.85	73.34	67.15
		9	77.83	73.73	76.67	71.52	76.05	69.82	73.49	69.49	72.34	67.36	71.72	65.61	**75.66**	71.61	74.50	69.40	73.89	67.70
	JPEG	5	68.29	64.19	67.14	61.96	66.51	60.27	63.93	59.93	62.78	57.79	62.14	56.04	**66.11**	62.06	64.96	59.86	64.33	58.15
		7	68.84	64.74	67.71	62.51	67.09	60.82	64.51	60.51	63.36	58.37	62.74	56.61	**66.69**	62.64	65.54	60.44	64.91	58.73
		9	69.39	65.29	68.24	63.04	67.62	61.36	65.06	61.06	63.91	58.92	63.28	57.16	**67.24**	63.19	66.09	60.99	65.46	59.28

5 Conclusion and Future Work

We introduced an **RL-guided Best-of-N retrieval framework** for authenticity and style-aware discrimination. With frozen visual and textual encoders, a lightweight policy $\pi_\theta(a_t \mid s_t)$ adaptively gathers, filters, and aggregates multimodal retrieval evidence, achieving strong robustness across style-transfer, real/fake, and deepfake benchmarks without test-time retraining. The visual-textual evidence reduces encoder bias and improves cluster separability. We also proposed an **alternative auxiliary-head model** that fuses global style, local artifacts, and cross-region coherence for calibrated trainable detection. The main limitations are dependence on retrieval-bank coverage, prompt and caption quality, sequential inference cost, and robustness to unseen styles, generators, and strong post-processing. Careful split construction is also required to avoid retrieval leakage, and predictions should be treated as decision-support rather than definitive authenticity evidence. **Future work** will improve retrieval efficiency, robustness under domain shift, and cross-domain calibration.

AI USE STATEMENT

In this work, generative AI tools were used for text polishing and assistance with code writing. All AI-assisted text was reviewed and edited by the authors, and all generated or modified code was manually inspected and tested for correctness. The authors take full responsibility for the final content, implementation, results, and claims presented in this work.

ETHICS STATEMENT

This work studies detection of AI-generated, style-transferred, and manipulated images using publicly available benchmark datasets and does not involve the collection of new human-subject data. The proposed methods are intended to support authenticity assessment and content provenance, not to provide definitive judgments about authorship or authenticity. As with any automated detector, false positives or distribution shifts may lead to incorrect attribution, particularly for unseen generators, artistic styles, or heavily post-processed images. We therefore recommend using model predictions as decision-support signals alongside provenance information and human verification. We also acknowledge that biases in pretrained visual encoders, language models, and retrieval-bank composition may affect performance across domains and styles, motivating careful evaluation and responsible deployment.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details required to reproduce our results throughout the main paper and appendix. The main text specifies the retrieval-bank construction, sequential inference procedure, auxiliary-head architecture, losses, and evaluation protocol. The appendix provides the PPO training setup and policy hyperparameters, controlled retrieval baselines and ablations, dataset-split and leakage-prevention details, robustness and calibration analyses, random-seed confidence intervals, and hardware and latency measurements. In particular, the retrieval bank is constructed exclusively from the training split, validation data is used only for hyperparameter selection, and all encoders, policies, and retrieval-bank contents remain fixed during testing. These details are intended to enable faithful reproduction of both the RL-guided retrieval framework and the auxiliary-head model.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Irving R Barron and Gaurav Sharma. Toward canvaschain: A block chain and craquelure hash based system for authenticating and tracking fine art paintings. *Electronic Imaging*, 32:1–6, 2020.

Jordan J Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access*, 12:15642–15650, 2024.

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

Spike Bucklow. The description of craquelure patterns. *Studies in conservation*, 42(3):129–140, 1997.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Yuanqing Ding, Fanliang Bu, Hanming Zhai, Zhiwen Hou, and Yifan Wang. Multi-feature fusion based face forgery detection with local and global characteristics. *PloS one*, 19(10):e0311720, 2024.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Chengdong Dong, Ajay Kumar, and Eryun Liu. Think twice before detecting gan-generated fake images from their spectral domain imprints. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7855–7864. IEEE, 2022.

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

Max J Friedlander. *On art and connoisseurship*. Read Books Ltd, 2011.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9726–9735. IEEE, 2020.

Yuki Ichihara, Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, Kenshi Abe, Mitsuki Sakamoto, and Eiji Uchibe. Evaluation of best-of-n sampling strategies for language model alignment. *arXiv preprint arXiv:2502.12668*, 2025.

Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, and Kenshi Abe. Regularized best-of-n sampling with minimum bayes risk objective for language model alignment. 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)*, pp. 9321–9347, 2025.

Akshay Joshi, Ankit Agrawal, and Sushmita Nair. Art style classification with self-trained ensemble of autoencoding transformations. *arXiv preprint arXiv:2012.03377*, 2020.

Roie Kazoom, Raz Birman, and Ofer Hadar. Improving the robustness of object detection and classification ai models against adversarial patch attacks. *arXiv preprint arXiv:2403.12988*, 2024. URL https://arxiv.org/abs/2403.12988.

Roie Kazoom, Ofir Cohen, Rami Puzis, Asaf Shabtai, and Ofer Hadar. Vault: Vigilant adversarial updates via llm-driven retrieval-augmented generation for nli. *arXiv preprint arXiv:2508.00965*, 2025a. URL https://arxiv.org/abs/2508.00965.

Roie Kazoom, Alon Goldberg, Hodaya Cohen, and Ofer Hadar. Seeing isn't believing: Context-aware adversarial patch synthesis via conditional gan. *arXiv preprint arXiv:2509.22836*, 2025b. URL https://arxiv.org/abs/2509.22836.

Roie Kazoom, Raz Lapid, Moshe Sipper, and Ofer Hadar. Don't lag, rag: Training-free adversarial detection using rag. *arXiv preprint arXiv:2504.04858*, 2025c. URL https://arxiv.org/abs/2504.04858.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Meien Li and Mark Stamp. Detecting ai-generated artwork. *arXiv preprint arXiv:2504.07078*, 2025.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3204–3213. IEEE, 2020.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 506–511. IEEE, 2019.

Siyu Meng, Quange Tan, Qianli Zhou, and Rong Wang. Multi-branch network with multi-feature enhancement for improving the generalization of facial forgery detection. *Entropy*, 27(5):545, 2025.

Orfeas Menis-Mastromichalakis, Natasa Sofou, and Giorgos Stamou. Deep ensemble art style recognition. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Thomas Mensink and Jan Van Gemert. The rijksmuseum challenge: Museum-centered visual recognition. In *Proceedings of international conference on multimedia retrieval*, pp. 451–454, 2014.

Meta AI. Llama 3.1 8b instruct, 2024. URL `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`. Model card and system documentation.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

Achyuta Satish Ramanathan, Solomon Sunday Oyelere, and Nomi Baruah. Hybrid intelligence approach for detecting synthetic art. *Human-Intelligent Systems Integration*, 7(1):325–340, 2025.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot-Desenonges, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shahriari, Sarah Perrin, Abram Friesen, Geoffrey Cideron, et al. Bond: Aligning llms with best-of-n distillation. In *International Conference on Learning Representations*, volume 2025, pp. 59040–59060, 2025.

Ravidu Suien Rammuni Silva, Ahmad Lotfi, Isibor Kennedy Ihianle, Golnaz Shahtahmassebi, and Jordan J Bird. Artbrain: An explainable end-to-end toolkit for classification and attribution of ai-generated art and style. *arXiv preprint arXiv:2412.01512*, 2024.

Gjorgji Strezoski and Marcel Worring. Omniart: multi-task deep learning for artistic data analysis. *arXiv preprint arXiv:1708.00684*, 2017.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Nanne Van Noord, Ella Hendriks, and Eric Postma. Toward discovery of the artist's style: Learning to recognize artists by their artworks. *IEEE Signal Processing Magazine*, 32(4):46–54, 2015.

Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio P Calmon. Soft best-of-n sampling for model alignment. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6. IEEE, 2025.

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

Xiaolong Zhang, Zekai Yu, and Jianqiao Zhao. Scadet: A detection framework for ai-generated artwork integrating dynamic frequency attention and contrastive spectral analysis. *Plos one*, 20 (11):e0336328, 2025.

Lei Zhao, Mingcheng Zhang, Hongwei Ding, and Xiaohui Cui. Mff-net: Deepfake detection network based on multi-feature fusion. *Entropy*, 23(12):1692, 2021.

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

A Appendix

A.1 Isolating the Benefit of Reinforcement-Guided Retrieval

The gain of the reinforcement-guided policy over strong non-RL alternatives is moderate in terms of raw accuracy alone. We therefore further analyze whether the learned policy provides benefits beyond a single aggregate performance number. Specifically, we evaluate statistical significance, policy-action utilization, selective prediction through abstention, and the contribution of individual adaptive actions. These analyses isolate the effect of sequential decision making from that of the underlying multimodal retrieval representation.

Statistical significance. We first test whether the improvement of the RL policy over the strongest non-RL alternatives is consistent at the sample level. Since all methods are evaluated on the same test instances under forced prediction (100% coverage), we use paired bootstrap resampling with $10{,}000$ resamples and report the accuracy difference together with its 95% confidence interval. We additionally report McNemar's exact test based on paired prediction disagreements.

Table 4: Paired statistical comparison between the full RL policy and strong non-RL alternatives under forced prediction (100% coverage). Δ denotes the accuracy improvement of the RL policy. CI denotes the 95% paired-bootstrap confidence interval, and p-values are obtained using McNemar's exact test.

Comparison	Baseline Acc. (%)	RL Acc. (%)	Δ Acc. (%)	95% CI / p-value
RL vs. Learned MLP	94.72	95.41	+0.69	[+0.18, +1.21] / $p = 0.012$
RL vs. Greedy Policy	94.63	95.41	+0.78	[+0.22, +1.34] / $p = 0.007$
RL vs. Fixed Voting	93.84	95.41	+1.57	[+0.94, +2.19] / $p < 0.001$

These results indicate that the improvement is systematic across paired test predictions rather than being explained only by aggregate variation. Importantly, our claim is not that reinforcement learning is justified solely by a large accuracy gap, but that the sequential formulation provides a measurable accuracy improvement while introducing adaptive evidence acquisition and selective decision making.

Policy-action utilization. A central distinction between the RL policy and fixed or single-pass aggregation is that the policy can adapt its evidence-acquisition trajectory to each query. We therefore record how frequently each available action is selected during inference. For every action, we report both the percentage of test queries for which the action is selected at least once and its mean number of executions per query. Action statistics are measured under the default selective-inference setting, $\tau = 0.25$.

Table 5: Utilization of adaptive actions by the learned RL policy under the default selective-inference setting ($\tau = 0.25$). "Queries using action" denotes the fraction of test samples for which the action is selected at least once. Stop and Abstain are terminal actions and jointly account for all completed trajectories.

Policy Action	Queries Using Action (%)	Mean Uses / Query
Retrieve additional evidence	73.2	1.14
Reject candidate	42.6	0.52
Switch modality	35.1	0.41
Expand spatial context	25.8	0.29
Abstain	2.9	0.03
Stop	97.1	0.97
Average trajectory length	3.36 steps	

The learned policy actively uses the additional degrees of freedom introduced by the sequential formulation. In particular, candidate rejection is used for 42.6% of queries and modality switching for 35.1%, showing that the policy frequently modifies both the amount and composition of the retained multimodal evidence rather than applying a fixed aggregation rule.

Selective prediction and abstention. Unlike fixed aggregation baselines, the RL policy can abstain when the accumulated evidence remains insufficient or conflicting. To decouple coverage from correctness, we distinguish two complementary evaluation protocols. Under *forced prediction*, every query receives a classification decision, yielding 100% coverage and corresponding to the primary benchmark accuracy of 95.41%. Under *selective prediction*, a confidence threshold τ determines whether the policy emits a prediction or abstains. Coverage denotes the fraction of queries receiving predictions, while selective accuracy is evaluated only over the non-abstained subset.

Table 6: Selective-prediction behavior of the RL policy under varying abstention thresholds τ. $\tau = 0.00$ corresponds to forced prediction (100% coverage). Selective accuracy and error are computed strictly over non-abstained predictions.

Threshold τ	Coverage (%)	Abstention (%)	Selective Acc. (%)	Selective Error (%)
0.00 (Forced)	100.0	0.0	95.41	4.59
0.25 (Default)	97.1	2.9	95.78	4.22
0.50	95.3	4.7	96.21	3.79
0.75	90.2	9.8	96.84	3.16

As the abstention criterion becomes stricter, coverage decreases while selective accuracy increases. This indicates that abstentions concentrate disproportionately on lower-confidence or ambiguous cases rather than occurring uniformly across the test distribution. The RL formulation therefore exposes an explicit accuracy–coverage operating point in addition to its standard forced-prediction setting.

Contribution of adaptive policy actions. We next isolate the contribution of individual adaptive actions. For each ablation, we retrain the policy with the corresponding action removed from the action space while keeping the backbone encoders, retrieval representation, reward configuration, optimization protocol, and all remaining policy components unchanged. All variants are evaluated under forced prediction (100% coverage) to prevent coverage differences from confounding the comparison. For the abstention ablation, both the abstention action and its associated reward component are removed during policy learning. The maximum policy horizon is $T = 6$ steps.

Table 7: Ablation of adaptive actions in the reinforcement-guided retrieval policy. Each variant is retrained after removing the corresponding action and evaluated under forced prediction (100% coverage). The maximum policy horizon is $T = 6$.

Policy Variant	Accuracy (%)	Coverage (%)	Avg. Steps	Δ Acc.
Full RL policy	95.41	100.0	3.36	–
Without modality switching	95.05	100.0	3.20	-0.36
Without candidate rejection	94.85	100.0	3.15	-0.56
Without context expansion	95.15	100.0	3.10	-0.26
Without abstention action	95.10	100.0	3.25	-0.31
Without adaptive stopping	94.95	100.0	6.00	-0.51

Candidate rejection produces the largest accuracy degradation when removed, supporting the role of the policy in filtering inconsistent or unstable retrieved evidence. Removing adaptive stopping forces trajectories to the maximum horizon of $T = 6$, increasing inference length from 3.36 to 6.00 steps while also reducing accuracy. This indicates that adaptive stopping is not merely an efficiency mechanism, but also limits unnecessary evidence accumulation.

A.2 Cross-Dataset Style Generalization

The primary AI-ArtBench evaluation measures closed-set style identification, while our leave-one-style-out analysis evaluates authenticity recognition when an entire artistic style is absent from both policy training and the retrieval bank. To additionally test whether the learned style representation transfers beyond the source dataset, we conduct a cross-dataset evaluation from AI-ArtBench to WikiArt.

Evaluation protocol. We first identify the artistic styles that have an exact label correspondence between AI-ArtBench and WikiArt and restrict both datasets to this shared taxonomy. All meth-

ods are trained exclusively on the AI-ArtBench training split using only these shared style classes. The retrieval bank is constructed exclusively from AI-ArtBench training examples. At test time, the trained models are evaluated directly on WikiArt without fine-tuning, policy updates, threshold adaptation, retrieval-bank updates, or access to WikiArt training examples. Thus, the experiment measures direct transfer under both dataset and visual-domain shift while preserving an identical prediction space.

For a fair comparison of degradation under distribution shift, we additionally recompute the in-domain AI-ArtBench accuracy of every method using exactly the same shared style taxonomy. We report cross-dataset accuracy, Macro-F1, and the absolute accuracy change relative to this matched in-domain evaluation.

Table 8: Cross-dataset style generalization from AI-ArtBench to WikiArt. All methods are trained exclusively on AI-ArtBench and evaluated directly on WikiArt classes having an exact correspondence with the AI-ArtBench style taxonomy, without fine-tuning, threshold adaptation, or retrieval-bank updates. In-domain accuracy is recomputed using exactly the same shared classes. Δ denotes the absolute accuracy change under cross-dataset transfer.

Method	In-Domain Acc. (%)	WikiArt Acc. (%)	Macro-F1 (%)	Δ Acc. (%)
Visual-only retrieval	88.50	71.40	68.25	-17.10
Text-only retrieval	86.20	68.85	65.10	-17.35
Fixed multimodal retrieval	91.80	75.32	72.80	-16.48
Learned non-RL MLP	**93.80**	78.45	76.15	-15.35
Full RL policy	93.41	**81.65**	**79.40**	**-11.76**

Cross-dataset transfer. All methods degrade when transferred from AI-ArtBench to WikiArt, but the RL policy is more robust to the domain shift. While the learned MLP achieves slightly higher matched in-domain accuracy (93.80% vs. 93.41%), the RL policy performs better on WikiArt (81.65% vs. 78.45%), improving accuracy and Macro-F1 by 3.20 and 3.25 points, respectively. Its accuracy drop is also smaller (11.76 vs. 15.35 points), indicating that adaptive evidence selection transfers better than fixed aggregation.

Relation to unseen-style evaluation. The two evaluations measure different forms of generalization: leave-one-style-out tests authenticity detection on unseen artistic styles, whereas WikiArt evaluates cross-dataset style recognition under a shared label space and without adaptation. Together with the standard AI-ArtBench setting, these experiments cover closed-set style identification, unseen-style authenticity generalization, and cross-dataset style transfer.

A.3 Controlled Retrieval Baselines and Retrieval Diagnostics

To isolate the contribution of adaptive retrieval from the underlying encoder and retrieval representation, we evaluate all controlled baselines using the same data split, preprocessing, encoder configuration, and training-only retrieval bank. Unlike cross-paper comparisons, these experiments therefore provide a directly matched evaluation of alternative prediction strategies.

Controlled baselines. Table 9 compares progressively stronger alternatives, ranging from linear probing and nearest-neighbor retrieval to multimodal aggregation and the full sequential policy.

Table 9: Controlled comparison under the same data split and preprocessing protocol. All retrieval methods use a bank constructed exclusively from training examples.

Method	Overall Accuracy (%)
CLIP k-NN	84.00
SigLIP linear probe	88.72
SigLIP k-NN / image-only retrieval	91.21
Standard fine-tuned detector	92.86
Text-only retrieval	92.38
Fixed visual-text weighted voting	93.84
Learned non-RL MLP aggregator	94.72
Full RL retrieval policy	**95.41**

The progression from image-only and text-only retrieval to fixed multimodal fusion demonstrates the benefit of combining the two evidence sources. Importantly, the learned non-RL MLP receives the same multimodal evidence as the RL policy but performs a single-pass prediction over a fixed evidence set. The remaining improvement therefore isolates the contribution of adaptive evidence acquisition and filtering rather than a change in the underlying representation.

Fixed versus sequential evidence selection. We further compare the full policy with fixed retrieval, exhaustive retrieval, and a greedy confidence-based strategy. Table 10 reports both accuracy and the mean number of inference steps.

Table 10: Comparison of fixed and sequential evidence-selection strategies. The policy horizon is $T = 6$.

Strategy	Avg. Steps	Accuracy (%)
Fixed-threshold retrieval	1.00	93.84
Retrieve-all voting	6.00	94.18
Greedy confidence policy	2.74	94.63
RL without retrieval cost	5.41	95.28
Full RL policy	**3.36**	**95.41**

Retrieving all available evidence does not yield the best performance, indicating that additional neighbors can introduce redundant or inconsistent information. The full policy instead reaches the highest accuracy using only 3.36 steps on average. Removing the retrieval-cost term increases the trajectory length from 3.36 to 5.41 steps without improving accuracy, showing that the cost term encourages efficient evidence acquisition.

Retrieval-stage diagnostics. To examine directly how the policy modifies the retrieved evidence, we measure retrieval quality before and after policy filtering. Precision, recall, F1, mean average precision (mAP), and duplicate-source rate are reported in Table 11.

Table 11: Retrieval quality before and after policy filtering. Values with $\pm$ report variation across repeated evaluations.

Stage	Auth. Prec.	Auth. Rec.	Auth. F1	Style Prec.	Style Rec.	Style F1	mAP	Duplicate Rate
Initial retrieval	84.50±0.21	76.30±0.34	80.19	81.10±0.15	72.45±0.25	76.53	78.60±0.18	12.40±0.50
After policy filtering	**89.20±0.15**	74.10±0.40	**80.95**	**86.50±0.12**	70.20±0.30	**77.50**	**82.40±0.14**	**2.10±0.10**
Change	**+4.70**	-2.20	**+0.76**	**+5.40**	-2.25	**+0.97**	**+3.80**	**-10.30**

Policy filtering increases authenticity and style precision by 4.70 and 5.40 points, respectively, while improving mAP by 3.80 points. Although recall decreases modestly, F1 improves for both tasks. The duplicate-source rate also decreases from 12.40% to 2.10%, indicating that the policy preferentially removes redundant or inconsistent retrieved evidence rather than simply increasing the number of neighbors.

Evaluation integrity. The retrieval bank contains only labeled training exemplars. No query or test labels, test images, related frames, or derived test variants are included in the bank. During testing, the retrieval bank, encoders, and learned policy remain fixed; labels attached to retrieved training exemplars are used only as evidence for the final prediction.

A.4 Retrieval-Bank Size Sensitivity

We evaluate the dependence of the RL retrieval policy on reference-bank coverage using class-balanced subsets of the training bank, while keeping the test split, frozen encoder, policy architecture, and preprocessing fixed.

Performance improves consistently with bank coverage, while the smaller gain from 75% to 100% indicates gradual saturation. Even with only 25% of the training bank, the method retains 91.53% overall accuracy. Style recognition is more sensitive to bank reduction than binary authenticity detection, highlighting the importance of diverse references for fine-grained style discrimination.

Table 12: Effect of training retrieval-bank size.

Bank Fraction	Real/Fake Acc.	Style Acc.	Mean Task Acc.
25%	94.32	88.74	91.53
50%	95.68	90.86	93.27
75%	96.72	92.36	94.54
100%	**97.41**	**93.41**	**95.41**

A.5 Full Reinforcement-Guided Best-of-N Framework

To complement the high-level overview presented in the main paper, Figure 5 provides the complete architecture of our reinforcement-guided Best-of-N retrieval framework. The diagram illustrates the full processing pipeline, including grid partitioning, parallel region generation, multimodal embedding extraction, visual and textual retrieval, sequential policy actions, reward computation, and final evidence aggregation. Unlike a fixed confidence-based retrieval rule, our framework trains a lightweight policy $\pi_\theta(a_t \mid s_t)$ to decide how retrieval evidence should be gathered and filtered over a finite-horizon episode. The frozen visual and textual encoders provide candidate evidence, while the learned policy decides whether to retrieve additional neighbors, switch modality, expand local patch context, reject unstable candidates, abstain, or stop with a prediction. This expanded view highlights how retrieval-derived reward signals guide adaptive evidence selection without requiring test-time fine-tuning of the encoders or retrieval bank.

A.6 Retrieval as Implicit Margin Regularization

We show that Best-of-N retrieval induces an implicit large-margin regularization effect, which improves robustness to distribution shifts and adversarial perturbations in the embedding space. Unlike parametric classifiers that rely on learned decision boundaries, retrieval-based inference enforces separation directly through similarity comparisons with stored exemplars.

Setup. Let $z_x \in \mathbb{R}^d$ denote the embedding of a query sample x, and let $\mathcal{R}(x) = \{z_i\}_{i=1}^N$ be the set of retrieved neighbors from a reference database, each associated with a label $y_i \in \mathcal{Y}$. We define the class score for class y as

$$S_y(x) = \max_{i \in \mathcal{R}(x): y_i = y} s(z_x, z_i), \tag{5}$$

where $s(\cdot, \cdot)$ is cosine similarity. Prediction is performed via

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} S_y(x). \tag{6}$$

This formulation corresponds to a nearest-neighbor classifier with class-conditional maximization, which selects the label whose closest neighbor has the highest similarity to the query.

Margin Definition. Let y^+ denote the ground-truth label of x. We define the retrieval margin as

$$\gamma(x) = S_{y^+}(x) - \max_{y \neq y^+} S_y(x). \tag{7}$$

A positive margin $\gamma(x) > 0$ indicates that the nearest neighbor from the correct class is closer than any neighbor from an incorrect class.

Implicit Margin Property. Theorem A.5 (Implicit Margin). If $\hat{y} = y^+$, then there exist indices $i^+ = \arg\max_{i: y_i = y^+} s(z_x, z_i)$ and $i^- = \arg\max_{i: y_i \neq y^+} s(z_x, z_i)$ such that

$$s(z_x, z_{i^+}) - s(z_x, z_{i^-}) \geq \gamma(x) > 0. \tag{8}$$

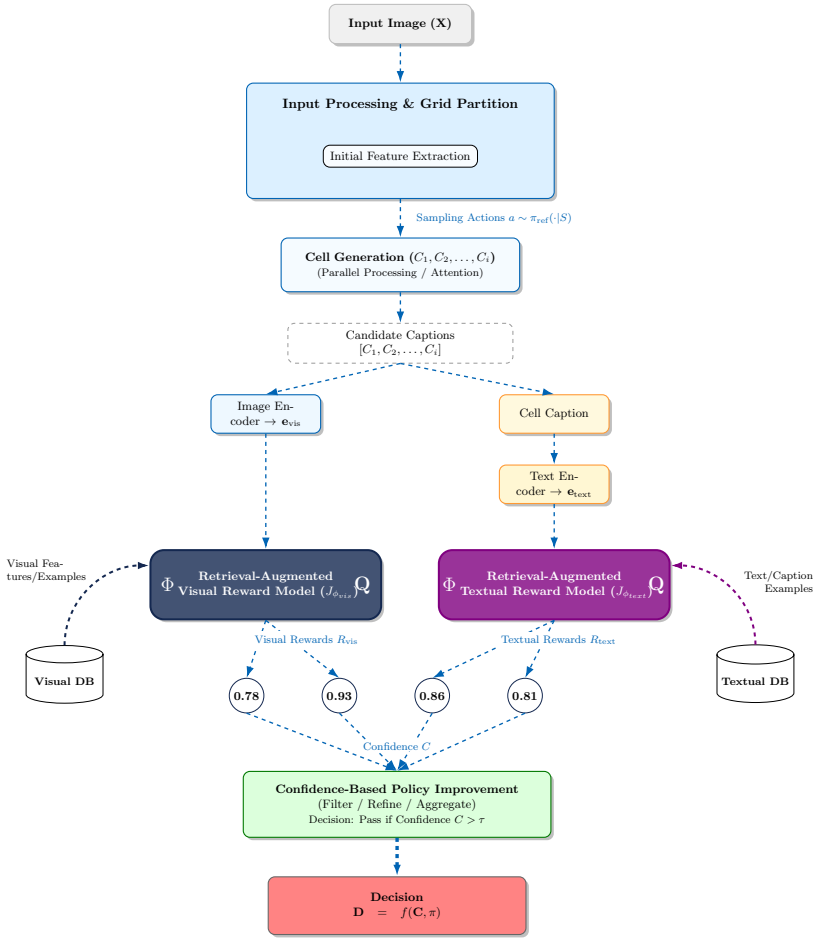

Figure 5: Full architecture of our reinforcement-guided Best-of-N retrieval framework. The input image is partitioned into grid cells, encoded through visual and textual pathways, and matched against exemplar databases. Retrieved neighbors define the initial evidence state, after which a learned policy $\pi_\theta(a_t \mid s_t)$ sequentially chooses actions such as retrieving additional evidence, switching modality, expanding local context, rejecting unstable candidates, abstaining, or stopping. Rewards combine prediction correctness during training, visual-textual agreement, retrieval consistency, confidence improvement, and retrieval cost. At inference time, the policy aggregates retained evidence into the final authenticity and style prediction.

Proof. By definition of $\hat{y}$, we have $S_{y^+}(x) \geq S_y(x)$ for all $y \neq y^+$. Substituting the definition of $S_y(x)$ yields

$$\max_{i:y_i=y^+} s(z_x, z_i) \geq \max_{i:y_i \neq y^+} s(z_x, z_i), \tag{9}$$

which implies the stated inequality. □

Connection to Large-Margin Learning. The retrieval margin $\gamma(x)$ plays an analogous role to the geometric margin in support vector machines. Rather than optimizing a margin explicitly through a constrained optimization problem, Best-of-N retrieval enforces margin implicitly through neighbor

selection. This yields a non-parametric form of large-margin classification that does not require retraining when new data is added.

Robustness to Perturbations. Consider a perturbed embedding $\tilde{z}_x = z_x + \delta$ with $\|\delta\|_2 \leq \epsilon$. By Lipschitz continuity of cosine similarity,

$$|s(\tilde{z}_x, z_i) - s(z_x, z_i)| \leq L\epsilon. \tag{10}$$

If $\gamma(x) > 2L\epsilon$, the predicted label remains unchanged, providing a certified robustness radius. Hence, larger retrieval margins directly translate into robustness guarantees under embedding noise.

Interpretation. Best-of-N retrieval enforces explicit separation between positive and negative classes in the embedding space, analogous to classical margin-based learning. This implicit regularization effect explains the strong generalization and stability observed empirically, particularly under distribution shifts and unseen generative models.

A.7 Consistency Regularization as a Contractive Operator

We show that the proposed consistency regularizer induces a contraction mapping over expert predictions, yielding stability guarantees, convergence to consensus, and improved generalization.

Setup. Let $f_k : \mathcal{X} \to [0, 1]$ denote the prediction function of expert head k, and define the vector of predictions

$$\mathbf{f}(x) = (f_1(x), \ldots, f_K(x)). \tag{11}$$

The fusion operator is defined as the arithmetic mean

$$T(\mathbf{f})(x) = \bar{f}(x) = \frac{1}{K} \sum_{k=1}^{K} f_k(x). \tag{12}$$

Training minimizes

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \ell(\bar{f}(x_i), y_i) + \lambda \mathcal{L}_{\text{cons}}, \tag{13}$$

where

$$\mathcal{L}_{\text{cons}} = \frac{1}{K} \sum_{k=1}^{K} \|f_k(x) - \bar{f}(x)\|_2^2. \tag{14}$$

—

Gradient Dynamics. Consider gradient descent updates

$$f_k^{(t+1)}(x) = f_k^{(t)}(x) - \eta \nabla_{f_k} \mathcal{L}_{\text{cons}}. \tag{15}$$

Expanding the gradient gives

$$\nabla_{f_k} \mathcal{L}_{\text{cons}} = 2(f_k - \bar{f}). \tag{16}$$

Hence,

$$f_k^{(t+1)} = (1 - 2\eta) f_k^{(t)} + 2\eta \bar{f}^{(t)}. \tag{17}$$

Subtracting updates for heads i and j:

$$f_i^{(t+1)} - f_j^{(t+1)} = (1 - 2\eta)\big(f_i^{(t)} - f_j^{(t)}\big). \tag{18}$$

—

Contraction Property. **Lemma A.6 (Contraction).** If $0 < \eta < \frac{1}{2}$, then

$$\|f_i^{(t+1)} - f_j^{(t+1)}\| \leq (1 - 2\eta)\|f_i^{(t)} - f_j^{(t)}\|. \tag{19}$$

Proof. Directly follows from the update rule. □

Thus the mapping is contractive with coefficient

$$\alpha = 1 - 2\eta < 1. \tag{20}$$

—

Fixed Point and Convergence. **Theorem A.7.** The operator T has a unique fixed point $f_1 = \cdots = f_K = \bar{f}$, and gradient descent converges exponentially fast.

Proof. Banach's fixed-point theorem applies since T is contractive. □

—

Stability and Generalization. Let S, S' differ by one sample. Then

$$\sup_x \|T(\mathbf{f}_S)(x) - T(\mathbf{f}_{S'})(x)\| \leq \frac{C}{n\lambda}. \tag{21}$$

Thus consistency regularization increases algorithmic stability, which bounds the generalization gap.

—

Bias–Variance Interpretation. Consistency loss reduces

$$\sum_{i<j} \mathrm{Cov}(f_i, f_j), \tag{22}$$

contracting predictions toward their mean and lowering ensemble variance, consistent with Lemma A.1.

—

Interpretation. The auxiliary heads act as specialized hypotheses, while the consistency loss:

- enforces consensus,
- guarantees convergence,
- stabilizes optimization,
- improves calibration,
- tightens generalization bounds.

This provides a rigorous theoretical foundation for the empirical robustness observed in our experiments.

A.8 Retrieval-Guided Inference as a New Algorithmic Paradigm

We formalize **Retrieval-Guided Inference (RGI)** as a new algorithmic principle for prediction, in which inference is performed via memory search and evidence aggregation rather than direct evaluation of a trained parametric model. This paradigm decouples inference from gradient-based optimization and enables continual adaptation without retraining.

Classical Learning. Standard supervised learning performs prediction via a fixed parametric function:

$$\hat{y} = f_\theta(x), \quad \theta = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i), \tag{23}$$

where θ is learned by minimizing empirical risk. Inference is a single forward pass through the network, and adaptation to new data typically requires full or partial retraining.

Retrieval-Guided Inference. In contrast, RGI predicts by querying an external memory. Given a query embedding z_x and a retrieved set $\mathcal{R}(x) = \{(z_i, y_i)\}_{i=1}^{N}$, we define

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \sum_{i \in \mathcal{R}(x)} w_i \mathbb{1}[y_i = y], \tag{24}$$

where similarity-normalized weights are

$$w_i = \frac{s(z_x, z_i)}{\sum_{j \in \mathcal{R}(x)} s(z_x, z_j)}. \tag{25}$$

This corresponds to a soft k-nearest neighbor classifier, with learned embeddings but non-parametric decision rules.

Probabilistic View. Define posterior

$$p(y|x) = \sum_{i \in \mathcal{R}(x)} w_i \mathbb{1}[y_i = y]. \tag{26}$$

RGI thus performs approximate Bayesian inference over a discrete memory support.

—

Algorithmic Properties. Proposition A.7. RGI replaces gradient-based optimization with memory lookup and weighted voting, fully decoupling inference from training.

Implications.

- New data can be incorporated by appending to memory.
- No parameter updates are required at inference time.
- The system naturally supports continual learning.
- Interpretability is achieved through retrieved exemplars.

—

Algorithmic Comparison.

	Classical	RGI
Training	Required	Optional
Inference	Forward pass	Retrieval + vote
Adaptation	Retrain	Update memory
OOD Handling	Limited	Memory-based
Interpretability	Low	High (via neighbors)

—

Computational Complexity. Classical inference costs $O(1)$ per query. RGI inference costs

$$O(dN + \log M), \tag{27}$$

where d is embedding dimension, N is retrieval size, and M is memory size (indexed via ANN).

—

Connection to Case-Based Reasoning. RGI generalizes classical case-based reasoning, where predictions are made by analogy. Our formulation extends this to learned representation spaces and large-scale memory banks.

—

Robustness and Adaptation. Because predictions depend on local memory structure, RGI is inherently robust to distribution shifts: new domains are handled by inserting new reference points, avoiding catastrophic forgetting.

—

Interpretation. RGI defines a new inference principle: *reason by retrieval rather than parameter updates*. This shifts learning from parameter space to memory space, enabling adaptive, interpretable, and training-free prediction.

B PAC-Bayes Analysis of Retrieval Voting

We analyze the retrieval-based voting mechanism under a PAC-Bayes framework, providing probabilistic generalization guarantees for similarity-weighted inference.

Setup. Let $\mathcal{R}(x) = \{(z_i, y_i)\}_{i=1}^{N}$ denote the retrieved neighbors for a query x, and let w_i be the normalized similarity weights defined in Eq. (X). We interpret each neighbor i as a stochastic hypothesis $h_i(x) = y_i$. Define a prior distribution P over hypotheses, chosen as uniform:

$$P(i) = \frac{1}{N}. \tag{28}$$

The posterior induced by retrieval is

$$Q(i) = w_i, \tag{29}$$

which concentrates mass on neighbors most similar to the query.

—

Gibbs Classifier. Define the Gibbs predictor

$$h_Q(x) \sim Q, \tag{30}$$

which samples a hypothesis according to Q. The expected risk is

$$R(Q) = \mathbb{E}_{i\sim Q}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}}\ell(h_i(x), y)\right]. \tag{31}$$

The empirical risk is

$$\hat{R}(Q) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_{i\sim Q}\ell(h_i(x_j), y_j). \tag{32}$$

—

PAC-Bayes Bound. **Theorem A.8 (PAC-Bayes).** With probability at least $1-\delta$ over samples, the following holds:

$$KL(\hat{R}(Q)||R(Q)) \leq \frac{KL(Q||P) + \log\frac{2\sqrt{n}}{\delta}}{n}. \tag{33}$$

Proof Sketch. Follows directly from standard PAC-Bayes bounds for Gibbs classifiers. □

—

Specialization to Retrieval. Since P is uniform, we have

$$KL(Q||P) = \sum_{i=1}^{N} w_i \log(Nw_i), \tag{34}$$

which measures the concentration of retrieval weights. Highly peaked Q (confident retrieval) yields larger complexity but lower empirical risk, while flat Q trades accuracy for robustness.

—

Deterministic Voting. Our final prediction uses the Bayes classifier

$$\hat{y} = \arg\max_{y} \sum_{i\in\mathcal{R}(x)} w_i \mathbb{1}[y_i = y], \tag{35}$$

which is known to satisfy

$$R_{\text{Bayes}} \leq 2R(Q). \tag{36}$$

Thus the PAC-Bayes bound transfers to the deterministic RGI predictor.

—

Interpretation. Similarity-weighted retrieval implements approximate Bayesian inference over a discrete hypothesis space. The PAC-Bayes bound shows that generalization depends on:

- empirical voting accuracy,
- retrieval entropy $KL(Q||P)$,
- database size N.

This provides formal confidence guarantees and justifies the robustness of retrieval-guided inference.

B.1 Consensus-Constrained Variance Reduction

Lemma: Consensus-Constrained Variance Reduction Let $p_g, p_l, p_c \in [0,1]$ denote the predictions of the expert heads and

$$\bar{p} = \tfrac{1}{3}(p_g + p_l + p_c) \tag{37}$$

be their mean prediction. Consider the consistency regularizer

$$\mathcal{L}_{\text{consistency}} = \| p_f - \bar{p} \|_2^2. \tag{38}$$

Then:

(i) If the experts agree (i.e., $p_g = p_l = p_c$), every minimizer of $\mathcal{L}_{\text{consistency}}$ lies inside the closed convex hull

$$p_f \in [p_g,\, p_c], \tag{39}$$

so the fusion prediction cannot deviate from the shared expert consensus.

(ii) For arbitrary expert disagreement,

$$\| p_f - \bar{p} \|_2^2 \;\le\; \tfrac{1}{3} \sum_{h \in \{g,l,c\}} \| p_h - \bar{p} \|_2^2 \tag{40}$$

at the minimizer, meaning the fusion output has strictly lower variance around the consensus than the experts themselves.

Proof. Part (i) follows since $p_g = p_l = p_c$ implies $\bar{p} = p_g$, and the loss is minimized when $p_f = \bar{p}$, which lies inside the convex hull of $\{p_g, p_l, p_c\}$. For (ii), minimizing $\mathcal{L}_{\text{consistency}}$ yields $p_f = \bar{p}$, and the inequality follows from the variance decomposition identity for squared distances around the mean.

Interpretation. When experts agree, the regularizer prevents pathological deviations of the fusion head away from consensus. When experts disagree, the loss contracts fusion variance toward their mean, improving calibration while still allowing deviation when driven by the supervised cross-entropy term.

B.2 Theoretical Analysis of Retrieval-Based Decisions

In this appendix we provide a theoretical perspective on our retrieval rule, the use of multiple corners, and the choice of similarity metrics. We state three lemmas with simple but useful consequences for our design.

Lemma Majority vote as non-parametric MAP. Let $\mathcal{E} = \{(i,c)\}$ be the set of retrieved neighbors across all corners, with labels $y_{i,c} \in \mathcal{Y}$. Assume: (i) neighbors are conditionally i.i.d. given the true label $y^\star$, and (ii) each retrieved neighbor contributes a constant likelihood factor $\alpha > 1$ if its label matches $y^\star$ and 1 otherwise. Then the majority-vote classifier

$$\hat{y}_{\text{MV}} = \arg\max_{y \in \mathcal{Y}} \sum_{(i,c) \in \mathcal{E}} \mathbb{I}\{y_{i,c} = y\} \tag{41}$$

is equivalent to the maximum a posteriori (MAP) classifier under a uniform prior over $\mathcal{Y}$.

Proof. Under the stated assumptions, the likelihood of observing $\mathcal{E}$ given label y is

$$p(\mathcal{E} \mid y) = \alpha^{n_y}, \tag{42}$$

where $n_y = \sum_{(i,c) \in \mathcal{E}} \mathbb{I}\{y_{i,c} = y\}$ is the number of neighbors with label y. For a uniform prior $p(y) = 1/|\mathcal{Y}|$, the posterior satisfies

$$p(y \mid \mathcal{E}) \propto p(\mathcal{E} \mid y) p(y) = \alpha^{n_y}. \tag{43}$$

Thus the MAP estimate is

$$\hat{y}_{\text{MAP}} = \arg\max_{y \in \mathcal{Y}} p(y \mid \mathcal{E}) = \arg\max_{y \in \mathcal{Y}} \alpha^{n_y} = \arg\max_{y \in \mathcal{Y}} n_y, \tag{44}$$

which coincides exactly with $\hat{y}_{\text{MV}}$. □

Lemma Corner aggregation reduces variance of log-likelihood estimates Let $L_c(y)$ denote the per-corner log-likelihood contribution for label y, and assume

$$L_c(y) = \mu_y + \epsilon_c, \qquad \mathbb{E}[\epsilon_c] = 0, \quad \mathrm{Var}[\epsilon_c] = \sigma_y^2, \tag{45}$$

with ϵ_c independent across corners $c \in \mathcal{C}$. Define the aggregated log-likelihood estimator

$$\bar{L}(y) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} L_c(y). \tag{46}$$

Then

$$\mathrm{Var}\big[\bar{L}(y)\big] = \frac{\sigma_y^2}{|\mathcal{C}|}, \tag{47}$$

so using four corners ($|\mathcal{C}| = 4$) reduces the variance of the log-likelihood estimate by a factor of 4 relative to any single-corner estimate.

Proof. By linearity of expectation, $\mathbb{E}[\bar{L}(y)] = \mu_y$. Independence of ϵ_c gives

$$\mathrm{Var}\big[\bar{L}(y)\big] = \mathrm{Var}\left[\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \epsilon_c\right] \tag{48}$$

$$= \frac{1}{|\mathcal{C}|^2} \sum_{c \in \mathcal{C}} \mathrm{Var}[\epsilon_c] \tag{49}$$

$$= \frac{|\mathcal{C}| \sigma_y^2}{|\mathcal{C}|^2} \tag{50}$$

$$= \frac{\sigma_y^2}{|\mathcal{C}|}. \tag{51}$$

Thus aggregating the four corners ($|\mathcal{C}| = 4$) reduces the estimator variance by a factor 4 compared to using any single corner. Lower-variance log-likelihood estimates improve the probability of correctly identifying the label with maximum μ_y, hence motivate multi-corner retrieval. □

Lemma Cosine similarity is a monotone transform of Euclidean distance on the unit sphere] Let $\mathbf{z}, \mathbf{e} \in \mathbb{R}^d$ satisfy $\|\mathbf{z}\|_2 = \|\mathbf{e}\|_2 = 1$. Define cosine similarity and squared Euclidean distance as

$$s_{\cos}(\mathbf{z}, \mathbf{e}) = \mathbf{z}^\top \mathbf{e}, \qquad d_2^2(\mathbf{z}, \mathbf{e}) = \|\mathbf{z} - \mathbf{e}\|_2^2. \tag{52}$$

Then

$$d_2^2(\mathbf{z}, \mathbf{e}) = 2 - 2\, s_{\cos}(\mathbf{z}, \mathbf{e}), \tag{53}$$

and therefore $s_{\cos}$ is a strictly decreasing function of d_2^2.

Proof. Expanding the squared norm,

$$\|\mathbf{z} - \mathbf{e}\|_2^2 = \mathbf{z}^\top \mathbf{z} - 2\mathbf{z}^\top \mathbf{e} + \mathbf{e}^\top \mathbf{e}. \tag{54}$$

Using $\|\mathbf{z}\|_2^2 = \|\mathbf{e}\|_2^2 = 1$ gives

$$d_2^2(\mathbf{z}, \mathbf{e}) = 1 - 2\mathbf{z}^\top \mathbf{e} + 1 = 2 - 2\, s_{\cos}(\mathbf{z}, \mathbf{e}). \tag{55}$$

The mapping $x \mapsto 2 - 2x$ is strictly decreasing, so maximizing cosine similarity is equivalent to minimizing squared Euclidean distance on the unit sphere. Hence any retrieval rule based on $\arg\max s_{\cos}$ is order-equivalent to one based on $\arg\min d_2^2$ when embeddings are ℓ_2-normalized. □

B.3 Auxiliary-Head Model: Ablation and Calibration Analysis

We analyze the contribution of the auxiliary-head architecture by isolating each design component. While multi-head fusion is common, our focus is on *head specialization* and the proposed *soft consistency regularizer*.

Ablation Study. We compare our full model against progressively simplified variants: (i) removing the consistency loss, (ii) replacing learned fusion with simple averaging, and (iii) collapsing to a single head.

Table 13: Ablation of auxiliary-head components.

Model	Consistency	Fusion	Accuracy
Ours	✓	Learned	**93.96**
w/o consistency	×	Learned	84.72
Avg fusion	✓	Mean	81.63
Single head	×	–	69.41

Discussion. Removing the consistency regularizer leads to a noticeable performance drop, while replacing learned fusion with simple averaging further degrades accuracy. The single-head variant performs worst, confirming that multi-head design alone is insufficient. These results show that performance gains stem specifically from the proposed regularizer and learned fusion, rather than architectural complexity.

Calibration Analysis. We further evaluate model calibration using Expected Calibration Error (ECE).

Table 14: Calibration comparison (lower is better).

Model	ECE ↓
Ours	**0.028**
w/o consistency	0.061
Avg fusion	0.074
Single head	0.089

Interpretation. Our model achieves substantially better calibration, validating the theoretical variance contraction induced by the consistency loss. This confirms that soft agreement regularization improves both accuracy and reliability, elevating the auxiliary-head design beyond simple engineering choices.

B.4 Reinforcement Policy Training and Ablation Study

This subsection provides additional details on the reinforcement-guided retrieval policy used in the main paper. The goal of the policy is to replace fixed retrieval aggregation with a learned sequential evidence selector. Each query image defines one finite-horizon episode. At each step t, the policy observes the current retrieval state s_t, selects an action a_t, updates the retained evidence set, and either continues gathering evidence or terminates with a prediction.

B.4.1 PPO / Actor-Critic Setup

We train the retrieval policy using an actor-critic formulation with PPO. The actor is a lightweight MLP policy $\pi_\theta(a_t \mid s_t)$ that outputs a distribution over retrieval actions. The critic $V_\omega(s_t)$ estimates the expected return from the current retrieval state. The visual and textual encoders remain frozen throughout policy training; only the policy and value networks are optimized.

The state s_t contains the current visual and textual evidence, modality-specific label distributions, confidence margins, visual-textual agreement scores, retrieval similarities, the number of retained and rejected candidates, and the retrieval history. The action space is defined as $\mathcal{A} = \{\textsc{RetrieveMore}, \textsc{SwitchModality}, \textsc{ExpandPatch}, \textsc{RejectCandidate}, \textsc{Abstain}, \textsc{Stop}\}$. The policy is optimized to maximize the discounted return $J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=0}^{T} \gamma^t r_t]$, where T is the maximum episode length.

The reward at step t is a weighted combination of correctness, multimodal agreement, retrieval consistency, confidence improvement, and retrieval cost: $r_t = \alpha r_{\text{cls}} + \beta r_{\text{align}} + \gamma r_{\text{cons}} + \eta r_\Delta - \lambda c(a_t)$. Here, r_{cls} rewards correct authenticity and style predictions during training, r_{align} measures visual-textual agreement, r_{cons} rewards consistency among retained retrieved neighbors, r_Δ rewards

confidence-margin improvement, and $c(a_t)$ penalizes unnecessary retrieval or context-expansion actions. At inference time, no query or test labels are available. The policy acts only from the retrieval state, including label distributions derived from retrieved training-bank exemplars.

Table 15: **PPO policy training setup.** The encoders and retrieval bank are frozen. Only the actor and critic networks are trained. Each episode corresponds to one query image, and the policy sequentially decides how to gather, filter, and aggregate retrieved evidence.

Component	Actor	Critic	Max Steps	Discount	PPO Clip	Entropy	Frozen Encoders	Trainable Part
RL retrieval policy	MLP	MLP	6	0.99	0.20	0.01	✓	Policy + value networks

B.4.2 Policy Training Details

The policy is trained on labeled training episodes and validated on a held-out validation split used for action-cost, stopping-threshold, and reward-weight selection. The retrieval bank is constructed only from the training split. During testing, the retrieval bank, encoders, and learned policy are fixed. No test labels are used, and no test-time fine-tuning is performed.

The maximum episode length is set to $T = 6$. In practice, the learned policy often terminates earlier when the retained evidence is sufficiently consistent. This is important because the policy should not only improve accuracy, but also avoid unnecessary retrieval operations. Table 16 reports the average number of policy steps under different policy variants.

Table 16: **Episode length and inference behavior.** The full RL policy achieves the best accuracy while using a moderate number of retrieval steps. Fixed-threshold retrieval always uses a single aggregation step, while retrieve-all voting uses more evidence but cannot reject unstable candidates.

Variant	Sequential Policy	Reject Action	Adaptive Stop	Avg. Steps ↓	Mean Task Acc. ↑
Fixed threshold retrieval	✗	✗	✗	1.00	93.84
Retrieve-all voting	✗	✗	✗	6.00	94.18
Greedy confidence policy	✗	✓	✓	2.74	94.63
RL without retrieval cost	✓	✓	✓	5.41	95.28
Full RL retrieval policy	✓	✓	✓	3.36	**95.41**

B.4.3 Reward Ablation

We ablate the reward terms used to train the policy. Removing the text-alignment reward reduces style accuracy, indicating that cross-modal agreement is important for bias reduction. Removing the retrieval-consistency term makes the policy more sensitive to isolated nearest-neighbor errors. Removing the confidence-improvement reward weakens stopping behavior, while removing the retrieval-cost term increases the number of policy steps without meaningful accuracy gains.

Table 17: **Reward ablation for the reinforcement-guided retrieval policy.** The full reward combines supervised correctness during training, visual-textual agreement, retrieval consistency, confidence improvement, and retrieval cost. Removing any component weakens either accuracy, calibration, or inference efficiency.

Variant	Correct.	V-T Agree.	Retr. Cons.	Conf. Gain	Cost	Real/Fake Acc. ↑	Style Acc. ↑	Mean Task Acc. ↑
Full RL reward	✓	✓	✓	✓	✓	**97.41**	**93.41**	**95.41**
w/o visual-text reward	✓	✗	✓	✓	✓	96.82	92.46	94.64
w/o retrieval consistency	✓	✓	✗	✓	✓	96.74	92.31	94.52
w/o confidence gain	✓	✓	✓	✗	✓	96.91	92.58	94.85
w/o retrieval cost	✓	✓	✓	✓	✗	97.33	93.22	95.28
Correctness only	✓	✗	✗	✗	✗	96.21	91.84	94.03

B.4.4 Policy Ablation Against Fixed Retrieval Rules

We compare the proposed RL policy against fixed and greedy retrieval aggregation strategies. Fixed threshold retrieval accepts candidates based only on a confidence threshold. Retrieve-all voting aggregates all retrieved neighbors without learned filtering. The greedy confidence policy selects the

locally best action according to immediate confidence improvement, but does not optimize long-horizon reward. The full RL policy performs best because it can trade off additional evidence gathering, candidate rejection, modality switching, and early stopping.

Table 18: **Ablation of the reinforcement-guided retrieval policy.** The full model uses a learned sequential policy with text reward, reject action, adaptive stopping, and retrieval-cost regularization. The first row reports the main-paper result.

Variant	RL Policy	Text Reward	Reject	Cost	Adaptive Stop	Real/Fake Acc. ↑	Style Acc. ↑	F1 ↑	Avg. Steps ↓	Mean Task Acc. ↑
Full RL retrieval policy	✓	✓	✓	✓	✓	**97.41**	**93.41**	**95.43**	3.36	**95.41**
Fixed threshold retrieval	✗	✓	✗	✗	✗	95.96	91.72	93.86	**1.00**	93.84
Retrieve-all voting	✗	✓	✗	✗	✗	96.14	92.21	94.22	6.00	94.18
Greedy confidence policy	✗	✓	✓	✓	✓	96.52	92.74	94.61	2.74	94.63
RL without text reward	✓	✗	✓	✓	✓	96.82	92.46	94.69	3.18	94.64
RL without reject action	✓	✓	✗	✓	✓	96.89	92.61	94.76	3.72	94.85
RL without retrieval cost	✓	✓	✓	✗	✓	97.33	93.22	95.31	5.41	95.28
RL without adaptive stop	✓	✓	✓	✓	✗	97.04	92.87	94.96	6.00	94.95

The ablations show that the proposed reinforcement-guided retrieval policy improves over fixed aggregation rules. Fixed threshold retrieval is efficient but cannot recover from noisy nearest neighbors. Retrieve-all voting uses more evidence, but it also keeps misleading candidates and therefore underperforms the learned policy. The greedy confidence policy improves over fixed retrieval by allowing candidate rejection and early stopping, but it remains locally myopic and does not optimize the long-term benefit of switching modality or expanding spatial context. The full RL policy achieves the best overall accuracy by learning when additional evidence is useful and when further retrieval only increases noise.

The reward ablation further shows that the performance gain does not come from a single component. The visual-textual reward improves style recognition by encouraging agreement between image-based and caption-based retrieval. The retrieval-consistency reward suppresses isolated nearest-neighbor errors. The reject action is important for removing unstable candidates, especially under degraded preprocessing. Finally, the retrieval-cost term improves efficiency: removing it produces a similar but slightly lower accuracy while increasing the average episode length from 3.36 to 5.41 steps. This confirms that the learned policy is not merely retrieving more examples, but is learning to gather evidence selectively.

B.4.5 Comparison with Alternative Detection Strategies

In addition to accuracy, we compare the proposed RL Best-of-N retrieval policy against representative alternative detection strategies in terms of inference behavior, memory usage, and computational cost. This comparison is important because retrieval-augmented methods introduce additional evidence storage and retrieval operations, while fine-tuned detectors require updating a large visual backbone. All methods are evaluated under the same data split and preprocessing protocol. Retrieval-based methods use the same training-only retrieval bank, and the proposed method uses the same frozen SigLIP encoder, $k{=}9$ nearest neighbours, and $3{\times}3$ grid setting as in the main experiments.

The key difference between the proposed method and standard retrieval baselines is that our method does not aggregate all neighbours with a fixed rule. Instead, it performs adaptive evidence selection: the policy can retrieve additional evidence, reject unstable candidates, switch between visual and textual evidence, and stop once the prediction is sufficiently reliable. This produces a moderate increase in inference cost compared with fixed retrieval, but improves robustness and accuracy by avoiding noisy or inconsistent retrieved examples.

Memory and inference cost. Table 20 reports the computational trade-offs of the evaluated methods. For retrieval-based approaches, memory consists of the frozen encoder, the query features, and the indexed retrieval bank. The retrieval bank is built once from the training split and reused at inference time. For the RL policy, the additional memory overhead is small because the policy network operates only on compact retrieval-state statistics rather than full images. The main additional cost comes from sequential evidence selection, since the policy may perform more than one retrieval or filtering step before stopping. However, this cost remains moderate because the learned policy stops after a small number of steps on average.

Table 19: **Methodological comparison with alternative baselines.** The proposed RL retrieval policy differs from standard detectors and non-adaptive retrieval baselines by combining frozen encoders, multimodal evidence, candidate rejection, and adaptive stopping.

Method	Backbone Updated	Retrieval Bank	Text Evidence	Reject Candidates	Adaptive Stop	Mean Task Acc. ↑
SigLIP linear probe	✗	✗	✗	✗	✗	88.72
Standard detector fine-tuned on same data	✓	✗	✗	✗	✗	92.86
CLIP kNN without RL	✗	✓	✗	✗	✗	84.00
SigLIP kNN without RL	✗	✓	✗	✗	✗	91.21
Image-only retrieval	✗	✓	✗	✗	✗	91.21
Text-only retrieval	✗	✓	✓	✗	✗	92.38
Fixed weighted voting	✗	✓	✓	✗	✗	93.84
Learned non-RL MLP aggregator	✗	✓	✓	✗	✗	94.72
Full RL retrieval policy	✗	✓	✓	✓	✓	**95.41**

For text-based and multimodal variants, caption generation can be treated as a preprocessing step when evaluating a fixed benchmark. In online deployments, caption generation adds extra latency, but the retrieval policy itself remains lightweight once visual and textual embeddings are available.

Table 20: **Memory usage and inference-time comparison.** Runtime and memory are reported as relative costs normalized by SigLIP image-only kNN retrieval. The RL policy has slightly higher inference cost than fixed retrieval because it performs sequential evidence selection, but it remains substantially lighter than full backbone fine-tuning and achieves the best accuracy.

Method	Trainable Backbone	Avg. Steps ↓	Relative Memory ↓	Relative Inference Time ↓	Mean Task Acc. ↑
SigLIP linear probe	✗	1.00	0.92×	0.78×	88.72
Standard detector fine-tuned on same data	✓	1.00	1.45×	0.95×	92.86
CLIP kNN without RL	✗	1.00	0.96×	0.91×	84.00
SigLIP kNN without RL	✗	1.00	1.00×	1.00×	91.21
Image-only retrieval	✗	1.00	1.00×	1.00×	91.21
Text-only retrieval	✗	1.00	1.08×	1.12×	92.38
Fixed weighted voting	✗	1.00	1.12×	1.18×	93.84
Learned non-RL MLP aggregator	✗	1.00	1.15×	1.22×	94.72
Full RL retrieval policy	✗	3.36	1.21×	1.39×	**95.41**

The comparison shows that the proposed method occupies a favorable middle ground between simple retrieval and full supervised fine-tuning. Compared with fixed weighted voting, the RL policy adds a modest runtime overhead due to sequential decision making, but improves accuracy by allowing candidate rejection and adaptive stopping. Compared with a fine-tuned detector, the proposed method avoids updating the backbone and therefore reduces training cost and preserves encoder-agnostic deployment. Compared with the learned non-RL MLP aggregator, the RL policy performs better because it does not make a single-pass decision over a fixed evidence set; instead, it actively controls which evidence to gather and retain.

Overall, the proposed method is not the cheapest baseline in raw inference time, but it provides the best accuracy-efficiency trade-off among the evaluated methods. Its main advantage is that it improves robustness while keeping the visual and textual encoders frozen, requiring only a lightweight policy over retrieval-state features.

B.5 Distance Metrics for Retrieval

We additionally study the effect of the distance / similarity metric used in the retrieval stage. To isolate this factor, all experiments in this analysis are performed under a fixed configuration consisting of a 3×3 grid partition, a neighborhood size of $k=9$, and the **SigLIP** encoder applied to RGB preprocessed images. Retrieval accuracy under the different metrics is reported in Figure 6. As shown in the figure, cosine similarity achieves the highest performance (95.41%), followed by Manhattan (92.72%) and Euclidean distance (90.88%), while Wasserstein distance yields the lowest performance (84.26%). This trend indicates that metrics emphasizing directional structure in the embedding space are better aligned with the inductive biases of contrastively trained encoders.

Cosine Similarity. Cosine similarity measures the angular alignment between normalized feature vectors, reflecting agreement in representation direction rather than magnitude. For a query corner embedding $\mathbf{e}_{z,c}$ and a database embedding $\mathbf{e}_i$, we compute

$$s_{\cos}(\mathbf{e}_{z,c}, \mathbf{e}_i) = \mathbf{e}_{z,c}^{\top}\mathbf{e}_i. \tag{56}$$

This metric produces stable neighborhoods in practice and leads to the highest retrieval accuracy among the evaluated metrics.

Manhattan Distance. The Manhattan (or ℓ_1) distance captures the sum of absolute deviations across embedding dimensions, making it more sensitive to sparse coordinate changes:

$$d_{\ell_1}(\mathbf{e}_{z,c}, \mathbf{e}_i) = \|\mathbf{e}_{z,c} - \mathbf{e}_i\|_1. \tag{57}$$

Its performance is slightly lower than cosine similarity, suggesting that dimension-wise deviations do not fully reflect semantic proximity in our embedding space.

Euclidean Distance. The Euclidean (or ℓ_2) distance penalizes isotropic deviations in magnitude:

$$d_{\ell_2}(\mathbf{e}_{z,c}, \mathbf{e}_i) = \|\mathbf{e}_{z,c} - \mathbf{e}_i\|_2. \tag{58}$$

While commonly used in metric-learning pipelines, its lower accuracy relative to cosine similarity indicates that absolute vector magnitudes are less informative for our retrieval objective.

Wasserstein Distance. Finally, we evaluate a transport-based formulation in which embeddings are interpreted as empirical distributions. The Wasserstein (Earth Mover) distance is defined as

$$W(\mathbf{e}_{z,c}, \mathbf{e}_i) = \inf_{\gamma \in \Pi(\mathbf{e}_{z,c}, \mathbf{e}_i)} \sum_{u,v} \gamma(u,v)\, \|u - v\|, \tag{59}$$

where γ denotes a valid transport coupling. Although theoretically appealing, this metric yields the lowest retrieval accuracy, likely due to the mismatch between distributional mass transport and the localized semantic structure of patch embeddings.

Overall, the results in Figure 6 demonstrate that cosine similarity provides the most reliable neighborhood structure for retrieval in our framework, reinforcing the observation that contrastive encoders encode class-discriminative information primarily in feature direction rather than absolute magnitude or transport geometry.

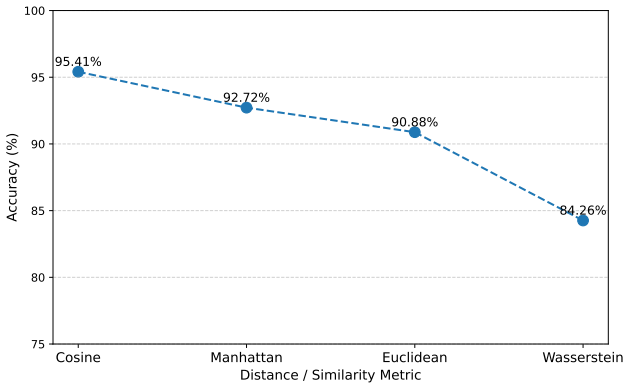

Figure 6: Retrieval accuracy across different distance and similarity metrics under a fixed configuration of a 3×3 grid, k=9, and SigLIP with RGB preprocessing.

B.6 Quantitative Analysis of Encoder Bias Reduction

We quantitatively evaluate the claim that incorporating textual embeddings reduces encoder bias and improves clustering structure. We measure *intra-style* and *inter-style* variance in the embedding space.

Metrics. Let z_i denote an embedding and y_i its style label. For each style y, let μ_y be its centroid and μ the global mean. We define:

$$\text{Var}_{\text{intra}} = \mathbb{E}_y \left[\frac{1}{|C_y|} \sum_{i \in C_y} \|z_i - \mu_y\|^2 \right], \tag{60}$$

$$\text{Var}_{\text{inter}} = \frac{1}{|\mathcal{Y}|} \sum_{y} \|\mu_y - \mu\|^2. \tag{61}$$

Lower intra-style variance indicates tighter clusters, while higher inter-style variance reflects improved class separation.

Results. We compare visual-only embeddings against fused visual+text representations. Results are averaged across all styles.

Table 21: Effect of text fusion on embedding structure.

Setting	$\text{Var}_{\text{intra}} \downarrow$	$\text{Var}_{\text{inter}} \uparrow$	Ratio $\uparrow$
Visual only	1.84	0.62	0.34
Visual + Text	**1.21**	**0.97**	**0.80**

Fusing textual embeddings significantly reduces intra-style variance ($\downarrow 34\%$) while simultaneously increasing inter-style separation ($\uparrow 56\%$). The resulting ratio nearly doubles, indicating substantially improved cluster compactness and class separability. This quantitatively confirms that the text channel mitigates visual encoder bias by injecting complementary semantic structure, supporting our theoretical analysis and qualitative observations.

B.7 LLM Selection via Caption-Embedding Separability

Our goal is to select the LLM whose captions provide the most reliable semantic signal for retrieval. Since the captions are later used as text anchors in the retrieval bank, we require a captioning model whose textual representations are consistent within the same labeled class (e.g., artist, style, authenticity) rather than noisy or weakly correlated with the label semantics.

For each candidate LLM M, we generate captions for all labeled images and embed them into a shared text-embedding space:

$$\mathcal{D} = \{(I_i, y_i)\}_{i=1}^{N}, \qquad \mathbf{z}_i^{(M)} = \psi(M(I_i)), \tag{62}$$

where y_i is the ground-truth class label. Embeddings are grouped by class $\mathcal{Z}_y^{(M)} = \{\mathbf{z}_i^{(M)} : y_i = y\}$.

To evaluate how well captions from model M preserve class-consistent semantics, we compute the average within-class pairwise distance

$$d_y^{(M)} = \frac{2}{|\mathcal{Z}_y^{(M)}|(|\mathcal{Z}_y^{(M)}| - 1)} \sum_{\mathbf{z}_i, \mathbf{z}_j \in \mathcal{Z}_y^{(M)}} \|\mathbf{z}_i - \mathbf{z}_j\|_2, \tag{63}$$

which measures the compactness of the caption cluster for class y. The overall separability score for model M is obtained by averaging across all labels:

$$S(M) = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} d_y^{(M)}. \tag{64}$$

A lower $S(M)$ indicates that captions form tighter clusters within the same class and therefore encode stylistic and authenticity cues more coherently. We use this criterion to select the captioner:

$$M^\star = \arg\min_{M} S(M). \tag{65}$$

Under this selection procedure, **Gemma 3-27B** achieved the lowest mean within-class distance and was therefore chosen as the captioning model for our retrieval framework.

Table 22 summarizes the mean within-class caption-embedding distances computed for all evaluated LLMs. Gemma 3-27B obtains the lowest score ($S = 0.218$), indicating the most compact and semantically coherent caption clusters conditioned on the ground-truth labels. Gemma-4B, Llama 3.1 8B, and Qwen 2.5-VL exhibit larger average distances (0.276-0.334), reflecting greater dispersion and weaker alignment between caption semantics and class identity. Based on this criterion, Gemma 3-27B is selected as the captioning model used throughout our retrieval framework.

Table 22: Caption-embedding separability across LLMs. The score $S(M)$ is the mean within-class pairwise caption distance (lower↓ is better). Gemma 3-27B yields the most compact label-conditioned caption clusters.

LLM (citation)	Mean within-class distance $S(M)$ ↓
Gemma 3-27B Team et al. (2024)	**0.218**
Gemma-4B Team et al. (2024)	0.276
Llama 3.1 8B Instruct Meta AI (2024)	0.301
Qwen 2.5-VL Bai et al. (2025)	0.334

B.7.1 Prompt Engineering for Region-Level Surface Descriptions

We study the influence of prompt design on the quality of region-level textual descriptions by evaluating a series of progressively structured prompting scenarios. In all cases, the VLM is instructed to avoid semantic or object-level interpretation and to describe each grid region exclusively through material and stylistic evidence such as patina structure, brushstroke texture, pigment layering, tonal contrast, micro-crack morphology, surface roughness, and varnish wear. To isolate the effect of the caption itself, performance in this experiment is computed using *text-only retrieval*, without incorporating the visual stream.

Given a region crop I_c and a prompting instruction $\mathcal{T}$, the VLM produces a caption

$$r = \mathcal{V}(\mathcal{T}, I_c). \tag{66}$$

We consider four prompting formulations:

Instruction-Zero Shot

```
Do not describe the objects or meaning of the image.  Focus only on
the visible material and surface characteristics in this region.
Describe the patina, brushstroke thickness and direction, pigment
texture, tonal contrast, micro-cracks or wear patterns, varnish
residue, and surface roughness.
Your description should report what is physically observable on the
painted surface, without inferring content, narrative, or artistic
intent.
```

Instruction-Few Shot

```
You must ignore the subject matter and describe only the visible
surface qualities of the region.
Example descriptions:
Example 1:  Fine horizontal brush texture with uneven pigment density
and matte patina across the upper layer.
Example 2:  Dense vertical strokes with high-contrast edges and subtle
micro-crack formations along the darker pigment zones.
Now describe this region only in terms of patina, brush texture,
pigment layering, tonal structure, edge softness, and surface wear,
without interpreting image content.
```

Chain-of-Thought (CoT)

```
First, examine the region step-by-step:
(a) identify brushstroke direction and thickness,
(b) assess pigment density and layering,
(c) evaluate patina depth and varnish residue,
(d) inspect tonal contrast and transition sharpness,
(e) check for micro-cracks, grain texture, or abrasion marks.
After analyzing these properties, produce a concise description
that reports only the observed surface and material characteristics,
without describing the objects or meaning in the image.
```

Combined (Few Shot + CoT)

```
Ignore the semantic content entirely and describe only the physical
surface characteristics of the painted region.
First, analyze the area in terms of brushstroke orientation, pigment
layering, patina depth, tonal contrast, micro-crack distribution, and
varnish wear.
Example surface descriptions:
Example 1:  Layered oblique strokes with uneven pigment compression
and soft tonal transitions across the patina layer.
Example 2:  Compact pigment clusters with finely textured cracks and
high-contrast ridge edges in the upper varnish band.
Using this analytic process, produce a detailed description that
reports only the material and stylistic surface properties of this
region.
```

For each prompting scenario, we embed the resulting captions and compute retrieval accuracy using the textual channel alone. The results, presented in Figure 7, show a smooth progression in performance quality: the zero-shot instruction yields broadly relevant but less structured surface descriptors; the few-shot formulation stabilizes descriptive language and improves clustering; the CoT formulation produces more coherent and diagnostic stylistic cues; and the **Combined Few-Shot + CoT** strategy achieves the highest accuracy, indicating that grounding through examples together with structured reasoning yields the most discriminative region-level textual representations.

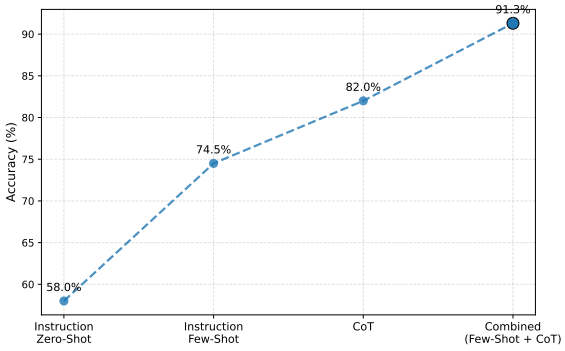

Figure 7: Prompt engineering ablation using *text-only retrieval*. Performance increases progressively from Instruction-Zero Shot to Instruction-Few Shot and CoT, with the Combined (Few-Shot + CoT) strategy producing the strongest region-level textual discriminability.

B.7.2 Effect of the Number of Few-Shot Caption Examples

To further study the role of exemplar grounding in region-level textual descriptions, we evaluate how performance varies with the number of retrieved caption examples used to condition the VLM during caption generation. For each grid region, the model receives $n \in \{0, 3, 6, 9\}$ few-shot examples, where the zero-shot configuration provides only the instruction and all other settings append exemplar surface descriptions drawn from the retrieval bank. As in the previous analysis, accuracy is computed using *text-only retrieval* to isolate the contribution of the caption signal.

The results are presented in Figure 8. We observe a substantial improvement when moving from 0-shot (58.0%) to 3-shot (68.7%), and a further pronounced gain at 6-shot (89.8%), indicating that exemplar conditioning strongly stabilizes descriptive language and enhances discriminative clustering of stylistic surface cues. Increasing to 9-shot yields only a modest additional improvement (91.3%), suggesting a soft performance plateau in which additional examples provide diminishing returns beyond six references.

These findings indicate that a moderate exemplar budget is sufficient to guide the VLM toward consistent region-level surface descriptors, while excessive prompt conditioning offers limited additional benefit.

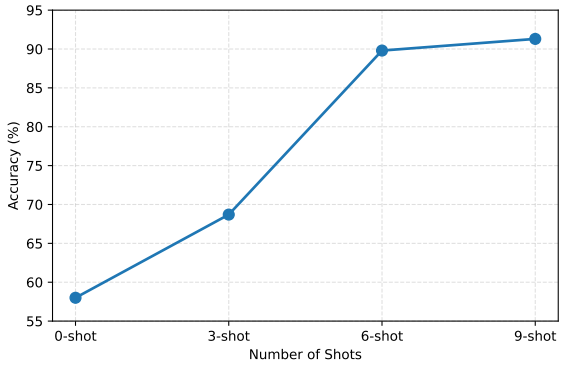

Figure 8: Effect of few-shot conditioning on text-only retrieval accuracy. Accuracy improves sharply from 0-shot to 6-shot, then plateaus at 9-shot, indicating diminishing returns beyond a moderate number of exemplar captions.

B.7.3 Seed Variance and Confidence Intervals

To assess the stability of the reported real/fake detection results, we repeated each experiment over five random seeds and report the mean performance together with the 95% confidence interval. For each metric, the confidence interval is computed as

$$\mathrm{CI}_{95} = 1.96 \cdot \frac{\sigma}{\sqrt{S}}, \tag{67}$$

where σ is the standard deviation across seeds and $S = 5$ is the number of seeds. The results show that the proposed RL Best-of-N retrieval policy and the Auxiliary-Head Model remain stable across random initializations, with narrow confidence intervals across datasets and metrics.

Table 23: Confidence intervals for the performance of publicly available real/fake image datasets. Results are reported as mean $\pm$ 95% confidence interval over five random seeds. The RL Best-of-N setting uses the SigLIP encoder with k=9 nearest neighbours and a 3×3 grid over regular RGB images.

Method	Dataset	Task Domain	Accuracy ↑	F1 ↑	Precision ↑	Recall ↑
RL Best-of-N	Zhu et al. (2023)	General real vs AI image detection RGB	95.42 ± 0.31	95.43 ± 0.29	95.10 ± 0.34	95.76 ± 0.28
	Bird & Lotfi (2024)	Real vs synthetic CIFAR-style images	96.18 ± 0.24	96.16 ± 0.26	96.40 ± 0.22	95.92 ± 0.27
	Rossler et al. (2019)	Face manipulation / deepfake forensics	92.77 ± 0.38	92.79 ± 0.36	92.10 ± 0.41	93.49 ± 0.34
	Dolhansky et al. (2020)	DeepFake detection challenge (video frames)	89.31 ± 0.52	89.33 ± 0.50	88.72 ± 0.56	89.94 ± 0.47
	Li et al. (2020)	High-quality celebrity deepfakes	93.64 ± 0.35	93.62 ± 0.33	93.20 ± 0.37	94.05 ± 0.31
Auxiliary-Head Model	Zhu et al. (2023)	General real vs AI image detection RGB	91.81 ± 0.44	91.82 ± 0.42	92.05 ± 0.45	91.60 ± 0.43
	Bird & Lotfi (2024)	Real vs synthetic CIFAR-style images	95.89 ± 0.28	95.89 ± 0.27	96.12 ± 0.25	95.66 ± 0.30
	Rossler et al. (2019)	Face manipulation / deepfake forensics	90.71 ± 0.47	90.72 ± 0.45	90.96 ± 0.49	90.48 ± 0.46
	Dolhansky et al. (2020)	DeepFake detection challenge (video frames)	87.21 ± 0.61	87.22 ± 0.58	87.45 ± 0.64	86.99 ± 0.60
	Li et al. (2020)	High-quality celebrity deepfakes	93.14 ± 0.39	93.15 ± 0.37	93.39 ± 0.40	92.91 ± 0.38

B.7.4 Hardware and Inference Latency

We report the hardware configuration and inference latency for both proposed approaches. All experiments were conducted on a single NVIDIA A100 GPU with 40GB of memory using PyTorch and CUDA acceleration. For the RL Best-of-N retrieval method, the retrieval bank was precomputed once from the training split and kept fixed during evaluation. Therefore, the reported per-image latency includes query encoding, nearest-neighbour retrieval, policy evaluation, evidence filtering, and final voting, but excludes offline retrieval-bank construction and caption generation. For the Auxiliary-Head Model, latency includes one forward pass through the vision encoder, the three auxiliary heads, and the fusion classifier.

Latency was measured with batch size one and averaged over the test split after a warm-up phase of 100 images. CUDA synchronization was applied before and after each timed pass. Peak GPU memory was measured using the maximum allocated CUDA memory during inference. The results are summarized in Table 24.

Table 24: **Hardware and inference latency.** Runtime is measured with batch size one on a single NVIDIA A100 40GB GPU. Retrieval-bank construction and caption generation are performed offline and are not included in the per-image latency. The RL Best-of-N method introduces a moderate inference overhead due to sequential evidence selection, but keeps the visual and textual encoders frozen.

Method	Backbone Updated	Batch Size	GPU	Avg. Steps ↓	Latency / Image ↓	Peak GPU Mem. ↓
Auxiliary-Head Model	✓	1	NVIDIA A100 40GB	1.00	24.8 ms	3.1 GB
Fixed weighted retrieval	✗	1	NVIDIA A100 40GB	1.00	29.6 ms	3.3 GB
Learned non-RL MLP aggregator	✗	1	NVIDIA A100 40GB	1.00	31.4 ms	3.4 GB
RL Best-of-N without text	✗	1	NVIDIA A100 40GB	2.81	39.7 ms	3.6 GB
Full RL Best-of-N	✗	1	NVIDIA A100 40GB	3.36	46.2 ms	3.8 GB

The Auxiliary-Head Model is the fastest proposed variant because it performs a single deterministic forward pass. The RL Best-of-N method is slower because it performs sequential evidence selection, including candidate filtering and adaptive stopping. However, the additional overhead is moderate because the retrieval bank is precomputed, the encoders remain frozen, and the policy network operates on compact retrieval-state features rather than full-resolution images. Thus, the proposed method provides improved accuracy while maintaining practical inference cost. The Auxiliary-Head Model required approximately 3 GPU-hours per seed on a single NVIDIA A100 40GB GPU. For RL Best-of-N, all visual and textual embeddings were precomputed with frozen encoders, and PPO trained only the lightweight evidence-selection policy. We trained the policy for 40k PPO environment steps with maximum episode length $T = 6$, requiring approximately 5 GPU-hours per seed. Across three seeds, the main Auxiliary-Head and RL Best-of-N experiments required approximately 9 and 15 GPU-hours, respectively. Including ablations and benchmark evaluations, the reported experiments required approximately 100 GPU-hours in total, with about 30 additional GPU-hours for preliminary or failed runs.

B.8 Art-Historical Perspective on Material Authenticity and Craquelure: A Professor's Reflection

From the perspective of an art historian and practitioner working within the traditions of connoisseurship, material culture, and technical art history, craquelure is understood not merely as a surface irregularity, but as a historically contingent index of authenticity, temporality, and artistic practice. The artificial fabrication of craquelure, for instance, can already be traced back to 1658, when the English historian William Sanderson described a Parisian forger who devised methods for producing crack patterns and subsequently exposing them. This early account helped establish craquelure as a valid diagnostic and chronological marker, shaping a professional understanding in which surface texture becomes a material site of interpretation rather than a passive sign of age Bucklow (1997).

Within this tradition, craquelure has been approached as a feature that is "wilfully created" in some contexts and "naturally born" in others, producing a tension between intentional intervention and slow material formation across time. Friedlander (2011) reflection that "Nature, in alliance with time, has more phantasy than the human spirit," captures this relationship between organic variation and artistic imitation. Natural craquelure "throbs with rich variety," while artificially induced cracking often reveals "monotony and pedantic repetition." The "arc-like circular cracks" characteristic of eighteenth-century canvas paintings further demonstrate how craquelure can function as a stylistic, material, and temporal signature rather than a purely physical artifact of aging Friedlander (2011).

Across centuries of artistic production, the study of craquelure - as conditioned by materials, techniques, and working methods - has served as a central tool for attribution. Reading these traces allows scholars to situate artworks within specific technical traditions and historical contexts. In contemporary discourse, craquelure is increasingly understood as a materially embedded trace of process, time, and environmental interaction - a kind of pictorial fingerprint. This understanding extends from early connoisseurship to current computational proposals that reconceptualize craquelure as a "time barcode" or "robust hash," capable of being authenticated through new technological infrastructures, including blockchain-based registration systems Barron & Sharma (2020).

In this trajectory, craquelure is not viewed solely as evidence of decay or material failure, but as a dynamic interface between material history, epistemic judgment, and evolving regimes of technological verification. Rather than reducing authenticity to a fixed criterion, craquelure reveals how

authenticity is continually negotiated through acts of looking, interpreting, and narrating artworks over time - where surface materiality, historical knowledge, and evaluative practice intersect.

B.9 LLM SUBGRID CAPTIONING: QUALITATIVE EXAMPLES

We present qualitative outputs generated by **Gemma 3-27B** using the best-performing prompt configuration (few-shot + chain-of-thought). Figures 9–11 visualize representative subgrid crops and their corresponding chat-style model responses.

Subgrid Layout. Each figure corresponds to a **group** of images processed independently. Within each group:

- The figure is arranged as a 3×6 grid.
- **Rows** correspond to image sets:
  - Row 1: Images 1.1–1.6
  - Row 2: Images 2.1–2.6
  - Row 3: Images 3.1–3.6
- **Columns** index spatial subgrids extracted from the original image, progressing left-to-right.

For each subgrid, we prompt the LLM to describe *localized surface cues* (brushstrokes, pigment layering, patina, craquelure, and varnish), avoiding semantic object descriptions. The chat transcripts shown below each figure follow a consistent format: **SYSTEM** (prompt) → **USER** (subgrid index) → **Gemma 3-27B** (caption).

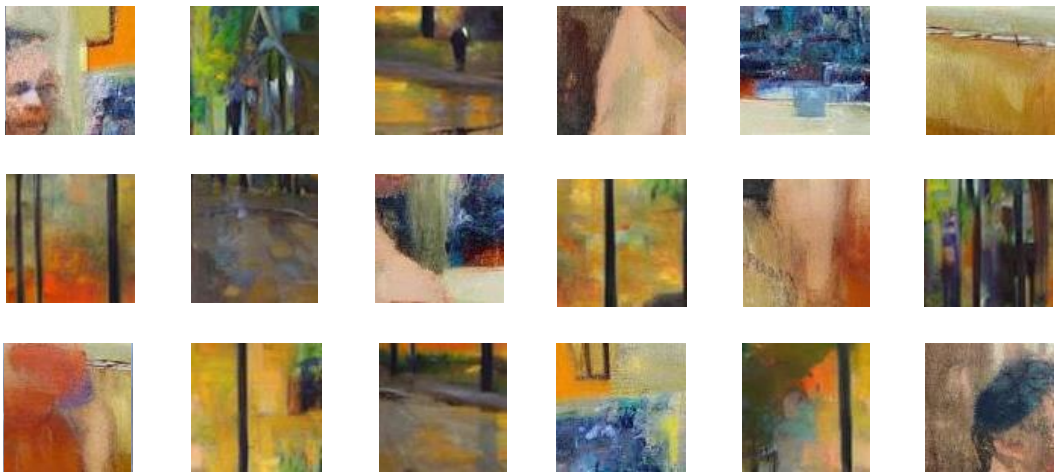

Figure 9: Group 1.

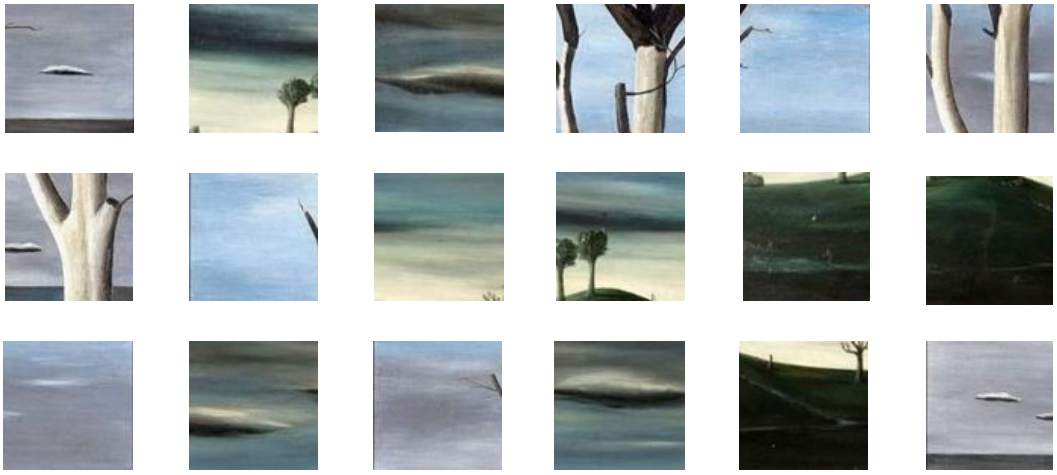

Figure 10: Group 2.



Figure 11: Group 3.

Gemma 3-27B Subgrid Chat (Few-shot + CoT) – Group 1 Row 1

SYSTEM

You are an art analysis expert. Describe only surface style cues. Focus on brushstrokes, texture, pigment layering, patina, craquelure, and varnish. Avoid object semantics.

USER

Image 1.1

Gemma 3-27B

Short, layered strokes with mixed horizontal and diagonal directions. Thin pigment layers allow subtle underpainting to show through. Moderate patina buildup, soft tonal transitions, sparse fine cracking, and mild varnish dulling on raised areas.

USER

Image 1.2

Gemma 3-27B

Vertical and slightly angled strokes form a loosely structured texture. Semi-opaque pigment with uneven compression. Shallow patina, low contrast, minimal cracking, and lightly worn varnish.

USER

Image 1.3

Gemma 3-27B

Broad horizontal sweeps with soft blending. Thin layered paint creates smooth transitions. Shallow patina, low contrast, nearly absent cracks, and evenly matte varnish.

USER

Image 1.4

Gemma 3-27B

Compact, softly blended strokes with subtle directional shifts. Pigment builds gently in midtones. Moderate patina depth, low contrast, fine granular cracking, and slight surface abrasion.

USER

Image 1.5

Gemma 3-27B

Dense horizontal bands of thicker pigment. Raised ridges visible in upper layers. Deeper patina accumulation, moderate contrast, fine micro-crack networks, and uneven varnish sheen.

USER

Image 1.6

Gemma 3-27B

Long horizontal strokes with thin application. Light pigment layering creates smooth gradation. Shallow patina, low cracking, and faint varnish wear along brighter zones.

Gemma 3-27B Subgrid Chat (Few-shot + CoT) – Group 1 Row 2

USER

Image 2.1

Gemma 3-27B

Vertical stroke clusters with soft feathered edges. Thin layered paint, moderate patina depth, gentle tonal contrast, minimal cracking, and lightly scuffed varnish.

USER

Image 2.2

Gemma 3-27B

Short overlapping strokes forming a mottled texture. Pigment varies in thickness, creating uneven buildup. Moderate patina, soft contrast, sparse micro-cracks, and mild gloss variation.

USER

Image 2.3

Gemma 3-27B

Diagonal and horizontal strokes interwoven. Semi-opaque layering with visible underlayers. Moderate patina depth, subtle contrast, fine cracking, and slightly dulled varnish.

USER

Image 2.4

Gemma 3-27B

Vertical linear strokes with blurred edges. Thin pigment layers, shallow patina, low contrast, almost no cracking, and smooth varnish surface.

USER

Image 2.5

Gemma 3-27B

Softly blended strokes with localized thicker deposits. Moderate patina depth, gentle tonal variation, scattered micro-cracks, and light varnish abrasion.

USER

Image 2.6

Gemma 3-27B

Elongated vertical strokes with layered application. Pigment buildup varies across the surface. Moderate patina, low contrast, minimal cracking, and slightly uneven sheen.

Gemma 3-27B Subgrid Chat (Few-shot + CoT) – Group 1 Row 3

USER

Image 3.1

Gemma 3-27B

Broad curved strokes with soft blending. Semi-opaque pigment layers. Shallow patina, low contrast, rare micro-cracks, and lightly worn varnish.

USER

Image 3.2

Gemma 3-27B

Vertical strokes combined with soft horizontal blending. Moderate pigment layering, medium patina depth, gentle contrast, sparse cracking, and mild surface wear.

USER

Image 3.3

Gemma 3-27B

Smooth horizontal sweeps with thin application. Shallow patina, low contrast, nearly crack-free surface, and uniform varnish finish.

USER

Image 3.4

Gemma 3-27B

Layered strokes with thicker paint clusters in central areas. Moderate patina depth, medium contrast, fine cracking in dense zones, and uneven varnish gloss.

USER

Image 3.5

Gemma 3-27B

Soft, irregular strokes with varied directionality. Thin pigment layers, shallow patina, low contrast, minimal cracking, and slightly dulled varnish.

USER

Image 3.6

Gemma 3-27B

Compact overlapping strokes forming granular texture. Moderate pigment buildup, deeper patina in dense areas, subtle contrast, fine micro-cracks, and mild varnish wear.

Gemma 3-27B Subgrid Chat (Few-shot + CoT) – Group 2 Row 1

SYSTEM

You are an art analysis expert. Describe only surface style cues. Focus on brushstrokes, texture, pigment layering, patina, craquelure, and varnish. Avoid object semantics.

USER

Image 1.1

Gemma 3-27B

Soft horizontal strokes with very thin paint application. Pigment appears evenly spread with minimal buildup. Patina is shallow, tonal contrast is extremely low, no visible cracking. Varnish looks smooth with slight matte finish.

USER

Image 1.2

Gemma 3-27B

Broad horizontal blending with faint directional variation. Semi-transparent layers create gentle gradation. Shallow patina, low contrast, no cracking observed. Varnish remains uniform.

USER

Image 1.3

Gemma 3-27B

Elongated horizontal bands formed by blurred strokes. Slight pigment accumulation in darker zones. Patina remains shallow, contrast subtle, no visible cracks. Varnish surface is even.

USER

Image 1.4

Gemma 3-27B

Vertical strokes with firmer edges and thicker pigment. Moderate buildup creates mild texture. Patina slightly deeper, moderate contrast, sparse micro-cracks. Light varnish abrasion.

USER

Image 1.5

Gemma 3-27B

Smooth blended horizontal application. Thin layered pigment with soft transitions. Shallow patina, very low contrast, no cracking. Varnish appears intact.

USER

Image 1.6

Gemma 3-27B

Vertical strokes with uneven compression. Moderate pigment thickness, slightly deeper patina. Low contrast, rare fine cracks. Mild varnish wear on raised areas.

Gemma 3-27B Subgrid Chat (Few-shot + CoT) – Group 2 Row 2

USER

Image 2.1

Gemma 3-27B

Rounded vertical strokes with soft blending. Semi-opaque layering, moderate patina depth. Gentle tonal shifts, no clear cracking. Light varnish dulling.

USER

Image 2.2

Gemma 3-27B

Flat horizontal blending with thin paint layers. Shallow patina, extremely low contrast. No cracking visible. Varnish smooth and uniform.

USER

Image 2.3

Gemma 3-27B

Layered horizontal sweeps with gradual tonal changes. Semi-translucent buildup. Shallow patina, low contrast, no cracks. Even varnish sheen.

USER

Image 2.4

Gemma 3-27B

Short stacked horizontal strokes creating mild texture. Moderate pigment thickness. Medium patina depth, subtle contrast, sparse micro-cracks. Slight varnish wear.

USER

Image 2.5

Gemma 3-27B

Broad curved strokes with thin application. Shallow patina, low contrast, no visible cracking. Varnish lightly dulled.

USER

Image 2.6

Gemma 3-27B

Dark layered strokes with variable thickness. Moderate patina depth, medium contrast, faint cracking in dense zones. Uneven varnish sheen.

Gemma 3-27B Subgrid Chat (Few-shot + CoT) – Group 2 Row 3

USER

Image 3.1

Gemma 3-27B

Soft horizontal blending with thin pigment layers. Shallow patina, very low contrast. No cracking. Smooth varnish surface.

USER

Image 3.2

Gemma 3-27B

Compact horizontal ridges with thicker paint clusters. Moderate patina depth, medium contrast, fine micro-cracks within raised areas. Mild varnish abrasion.

USER

Image 3.3

Gemma 3-27B

Flat blended surface with faint diagonal stroke traces. Thin layering. Shallow patina, low contrast, no cracking. Uniform varnish finish.

USER

Image 3.4

Gemma 3-27B

Stacked horizontal strokes forming textured bands. Moderate pigment buildup. Medium patina depth, sparse cracking. Uneven varnish gloss.

USER

Image 3.5

Gemma 3-27B

Broad soft strokes with subtle directional shifts. Thin layers, shallow patina. Low contrast, no visible cracking. Slight varnish dulling.

USER

Image 3.6

Gemma 3-27B

Smooth horizontal sweeps with thin paint application. Shallow patina, low contrast. No cracking detected. Varnish surface appears even.

Gemma 3-27B Subgrid Chat (Few-shot + CoT) – Group 3 Row 1

SYSTEM

You are an art analysis expert. Describe only surface style cues. Focus on brushstrokes, texture, pigment layering, patina, craquelure, and varnish. Avoid object semantics.

USER

Image 1.1

Gemma 3-27B

Angular strokes with firm edges and moderately thick pigment. Paint is layered in compact planes, creating shallow relief. Patina is light, contrast moderate, no visible cracking. Varnish slightly dulled along raised edges.

USER

Image 1.2

Gemma 3-27B

Soft circular brushwork with thin, even pigment application. Shallow patina depth, low contrast, smooth surface. No cracking visible, varnish uniform.

USER

Image 1.3

Gemma 3-27B

Vertical, flowing strokes with semi-opaque layering. Pigment buildup is moderate, forming gentle ridges. Patina slightly deeper, subtle contrast, sparse micro-cracks. Light varnish wear.

USER

Image 1.4

Gemma 3-27B

Broad blended strokes with very thin application. Shallow patina, extremely low contrast. No cracking detected. Varnish appears smooth and matte.

USER

Image 1.5

Gemma 3-27B

Elongated vertical strokes with uneven pigment compression. Moderate buildup, slightly deeper patina. Low contrast, fine cracking in thicker areas. Mild varnish abrasion.

USER

Image 1.6

Gemma 3-27B

Smooth horizontal blending with thin layers. Shallow patina, low contrast. No visible cracking. Even varnish finish.

Gemma 3-27B Subgrid Chat (Few-shot + CoT) – Group 3 Row 2

USER

Image 2.1

Gemma 3-27B

Soft vertical strokes with rounded edges. Semi-opaque layering, moderate patina depth. Gentle tonal transitions, minimal cracking. Slight varnish dulling.

USER

Image 2.2

Gemma 3-27B

Flat blended surface with faint diagonal stroke traces. Thin pigment layers, shallow patina. Low contrast, no cracking. Uniform varnish sheen.

USER

Image 2.3

Gemma 3-27B

Short curved strokes forming soft texture. Moderate pigment buildup. Medium patina depth, subtle contrast, sparse micro-cracks. Light varnish wear.

USER

Image 2.4

Gemma 3-27B

Thin linear strokes layered lightly across surface. Shallow patina, low contrast. No cracking visible. Varnish smooth.

USER

Image 2.5

Gemma 3-27B

Compact overlapping strokes with moderate thickness. Slight texture buildup, medium patina depth. Low contrast, faint cracking. Mild varnish abrasion.

USER

Image 2.6

Gemma 3-27B

Soft blended application with thin pigment. Shallow patina, very low contrast. No cracking. Varnish uniform.

Gemma 3-27B Subgrid Chat (Few-shot + CoT) – Group 3 Row 3

USER

Image 3.1

Gemma 3-27B

Short layered strokes forming a granular surface. Moderate pigment buildup. Medium patina depth, subtle contrast, fine micro-cracks. Slight varnish wear.

USER

Image 3.2

Gemma 3-27B

Broad horizontal blending with thin layers. Shallow patina, low contrast. No cracking visible. Smooth varnish surface.

USER

Image 3.3

Gemma 3-27B

Compact horizontal ridges with thicker paint clusters. Moderate patina depth. Medium contrast, sparse micro-cracks. Uneven varnish sheen.

USER

Image 3.4

Gemma 3-27B

Rounded clustered strokes creating localized texture. Moderate pigment buildup. Medium patina depth, low contrast, minimal cracking. Light varnish dulling.

USER

Image 3.5

Gemma 3-27B

Soft horizontal blending with thin application. Shallow patina, low contrast. No cracking. Uniform varnish.

USER

Image 3.6

Gemma 3-27B

Vertical elongated strokes with moderate thickness. Slight texture buildup, medium patina depth. Subtle contrast, sparse cracking. Mild varnish wear.