
Generative Data Augmentation for Artist Authorship Attribution

Yiqiao Huang Lixuan Wei Ruyi Yang Tian Xia
AC209B / CS1090B — Group 166
Harvard University

Abstract

We study artist authorship attribution on a 49-class painting dataset and observe a strongly long-tailed performance distribution: a frozen ResNet-18 with a linear classifier achieves 0.521 test accuracy, but mean recall on the ten worst-performing classes is only 0.238. To improve minority-class performance, we introduce generative data augmentation using per-artist DreamBooth fine-tuning of Stable Diffusion XL with Low-Rank Adaptation (LoRA), alongside a ControlNet+Canny variant for comparison. We find that naively adding synthetic samples under a standard inverse-frequency sampler yields little benefit due to an imbalance between real and synthetic samples within each class. To address this, we propose a corrected per-class sampling strategy with a single hyperparameter, `REAL_FRAC`, which decouples class balancing from real:synthetic mixing. Sweeping `REAL_FRAC` reveals a clear inverted-U tradeoff: worst-10 mean recall improves from 0.238 to **0.388** at `REAL_FRAC`=0.70 (+15 percentage points), while overall accuracy increases from 0.521 to 0.543 at `REAL_FRAC`=0.85. Comparing generators further shows that LoRA performs better on highly stylistic artists, whereas ControlNet+Canny is more effective for structurally distinctive or figurative ones. These results suggest that, in long-tail classification, the integration strategy for synthetic data can matter as much as the generator itself.

1 Background and Motivation

Identifying the artist behind a painting is a foundational task in art history, with practical applications in provenance research, museum cataloging, and forgery detection [10]. The task is also a useful stress-test of computer-vision models: where standard classification benchmarks reward object recognition, artist attribution demands sensitivity to texture, palette, brushwork, and compositional habits that vary subtly across visually similar genres.

Two structural difficulties make the problem hard. First, artist datasets are inherently long-tailed: prolific artists like van Gogh leave behind hundreds of works while others have only a few dozen, and unlike traditional classification we cannot simply collect more data for under-represented classes (the source population is closed). Second, within-artist diversity is large because many artists evolved through distinct stylistic periods. Both factors drive a steep gap between mean and tail performance that any practical attribution system must address. Third, the task is compounded by *stylistic entanglement*. During specific periods, such as the height of Impressionism, the visual manifolds of distinct artists exhibit significant overlap. Standard discriminative architectures are often susceptible to semantic interference, where the model over-prioritizes high-level object recognition at the expense of the subtle stylistic fingerprints that uniquely identify an artist.

Furthermore, conventional data augmentation techniques face fundamental limitations in this domain. Heuristic pixel-level transforms, such as horizontal flipping or color jittering, are inherently unable to synthesize the intricate nuances of brushwork or the characteristic textures of a master’s hand. While

large-scale pre-trained models like SDXL [6] possess vast generative priors, they frequently fail to capture fine-grained artistic details in zero-shot settings. These conditions make artist attribution a natural testbed for *generative data augmentation*—synthesizing additional training examples for under-represented classes using image generative models. Recent advances in diffusion models [5, 6] combined with parameter-efficient adaptation [7] now make it computationally feasible to fine-tune a state-of-the-art image generator to a specific artist with only a few dozen reference images. Whether such synthetic data actually *helps* a downstream classifier, however, is an open empirical question—and one where, as we show, methodology details (in particular, the sampling strategy) dominate the outcome.

2 Problem Statement

Given an RGB image x , our objective is to predict the artist label $y \in \{1, \dots, 49\}$. While standard classification focuses on global accuracy, our study specifically addresses the **long-tail distribution** inherent in art datasets, where data scarcity for minority artists leads to biased representation.

To isolate the impact of augmentation on the distribution’s tail, we report Top-1 accuracy on a *real-only* test set and decompose performance into: 1. *Worst-10 Recall*: Mean per-class recall for the ten least-represented/performing artists at baseline. 2. *Other-40 Recall*: Mean per-class recall for the remaining data-rich classes. This partitioning allows us to monitor whether generative augmentation improves the "tail" without inducing "semantic drift" in the "head."

The central research question is whether synthetic data provides a meaningful inductive bias for the classifier. We validate the effectiveness of our approach through a series of controlled experiments:

1. *Augmentation Effectiveness (Augmented vs. No-Augment)*: We establish a baseline by training solely on real images (*Real-Only*) and compare it against models trained on a mixture of real and synthetic samples (*Augmented*).
2. *Ablation on Generative method*: We conduct an ablation study on the input method (e.g., Text content prompt vs. Image content prompt). This allows us to quantify the trade-off between stylistic fidelity and generative diversity, identifying the "optimal method" that maximizes downstream classification gain.
3. *Mechanism Disentanglement*: We isolate whether performance gains are driven by the generator’s architectural prior (Vanilla SDXL) or the refined artist-specific distribution (LoRA).

3 Data and Exploratory Analysis

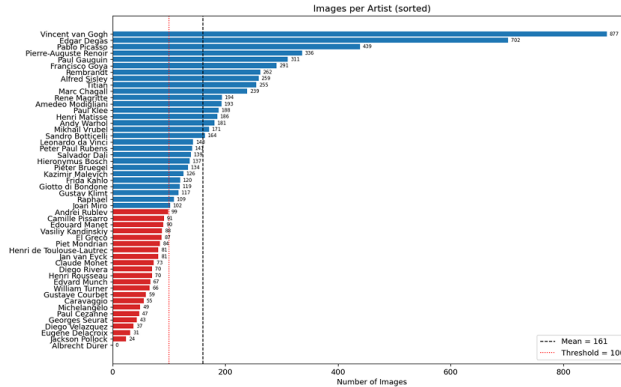


Figure 1: Per-artist image count (sorted). Red bars fall below 100 images.

Dataset. We use the *Best Artworks of All Time* Kaggle dataset [11], comprising 8,355 color JPEGs from 50 artists spanning the 13th to 20th centuries, with metadata including genre, nationality, and birth/death years. During preprocessing we identified a Unicode NFC/NFD encoding mismatch between the CSV labels and the image-folder names that caused *Albrecht Dürer* to be dropped from the label encoder; all classifiers and results in this paper are therefore trained and evaluated on the remaining 49 artists. Images are resized to 224×224 ; data is split 80/10/10 (train/val/test) stratified by artist with `random_state=42` fixed across all experiments.

Table 1: MS3 frozen-backbone / zero-shot baselines on real-only data, 49-class test accuracy. The italicized ResNet-18 row is the baseline against which all MS4 augmentation results are compared.

Model	Test acc.
ViT-B/16 (linear probe)	0.648
CLIP ViT-B/32 (zero-shot)	0.638
EfficientNet-B0 (linear probe)	0.560
<i>ResNet-18 (frozen backbone + linear head)</i>	<i>0.521</i>

Augmentation targets. To focus the augmentation study, we selected the ten worst-recall classes from our MS3 ResNet-18 baseline as augmentation targets: Paul Cezanne, Mikhail Vrubel, Gustave Courbet, Titian, Vincent van Gogh, Claude Monet, Kazimir Malevich, Eugene Delacroix, Vasiliy Kandinskiy, and Andrei Rublev (mean baseline recall 0.238).

4 Methods and Models

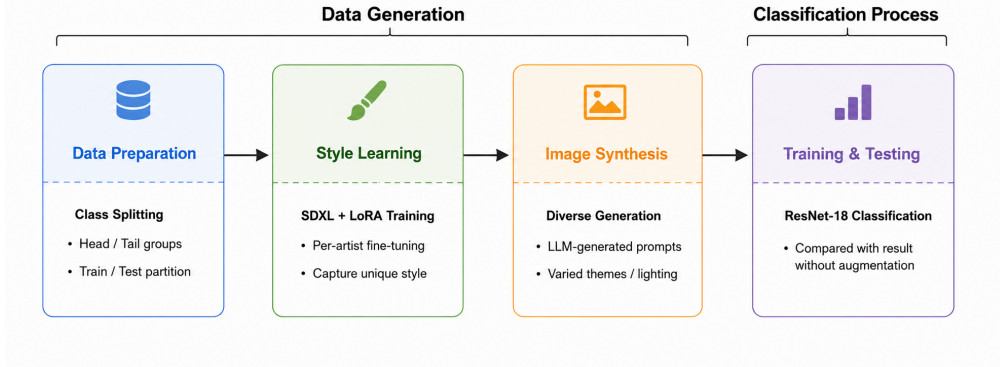


Figure 5: End-to-end pipeline. From left: real paintings train per-artist LoRA adapters on SDXL; LoRA-fused generation produces synthetic paintings for each worst-10 artist; the augmented training set is combined with the real training set under a per-class REAL_FRAC sampler to retrain a frozen-backbone ResNet-18 classifier.

4.1 Baselines

In MS3 we compared four frozen-backbone / zero-shot image classifiers on the 49-class task (Table 1). ViT-B/16 (linear probe) leads at 0.648, followed closely by CLIP zero-shot at 0.638; EfficientNet-B0 trails at 0.560, and ResNet-18 (frozen backbone + linear head) is the weakest at 0.521. For the MS4 augmentation study we adopt the ResNet-18 baseline as our reference: it has the lowest overall accuracy and the lowest worst-10 mean recall (0.238), so it offers the most headroom for augmentation to demonstrate an effect. It also trains in minutes rather than hours, which is essential for the multi-setting REAL_FRAC sweep we run in Section 5.

4.2 Generative augmentation: LoRA and ControlNet

As shown in Fig. 6, our data augmentation pipeline combines per-artist generator fine-tuning, synthetic image generation. As an AC209B project, our augmentation step uses two methods not covered in class.

Low-Rank Adaptation (LoRA). Rather than fine-tuning all $\approx 3.5\text{B}$ parameters of Stable Diffusion XL [6], LoRA [7] injects low-rank update matrices into each attention projection. If $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is a pretrained weight, the LoRA-adapted weight is $W' = W + BA$ with $B \in \mathbb{R}^{d_{\text{out}} \times r}$, $A \in \mathbb{R}^{r \times d_{\text{in}}}$, and rank $r = 8$. Only B and A are trained, reducing trainable parameters by roughly two orders of magnitude and avoiding catastrophic forgetting of the diffusion prior. We pair LoRA with the

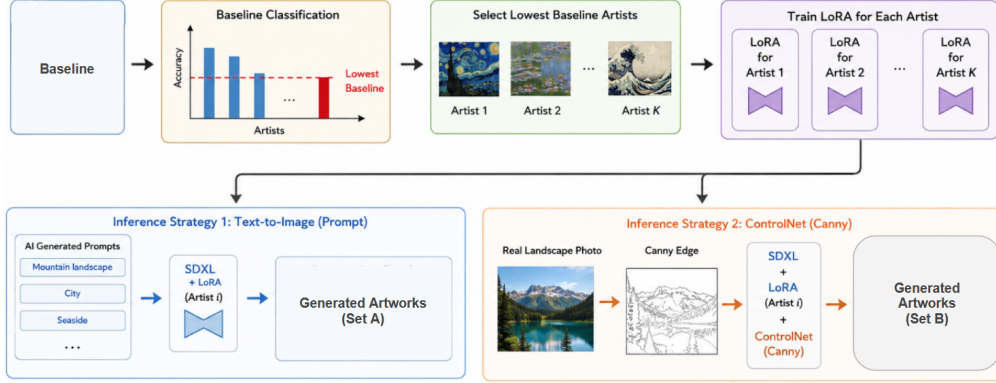


Figure 6: Our data augmentation pipeline that selects low-performing artists, trains artist-specific LoRA modules on SDXL, and generates two datasets via text prompts and ControlNet-guided Canny edges.

DreamBooth objective [8]: each training image is paired with the prompt “a painting by [V] artist” where [V] is a rare token, training SDXL to associate that token with artist-specific style. Per artist we use 40 paintings, 300 training steps, batch size 1, learning rate 1×10^{-4} ; one A100 produces a LoRA checkpoint in roughly five minutes.

Synthetic generation. As illustrated in Figure 6, for each of the ten target artists we generate synthetic data through two complementary inference strategies, both fusing the artist-specific LoRA into the cross-attention layers at scale 0.95, with 20 inference steps and 1024×1024 resolution. **(1) Set A (Text-to-Image):** We use 100 AI-generated content prompts (e.g. “a rainy city street”, “a mountain village at dusk”) $\times$ 2 random-seed repetitions, yielding 200 images per artist driven purely by textual descriptions. **(2) Set B (ControlNet-Canny):** To inject realistic compositional structure, we additionally generate 200 images per artist using ControlNet [9] conditioned on Canny-edge maps extracted from 100 real landscape photographs $\times$ 2 random-seed repetitions. This preserves natural scene layout while transferring the learned artist style. Together, Sets A and B provide stylistically faithful augmentation with both prompt-driven and structure-guided diversity.



Figure 7: The comparison between Real Artwork & Raw SDXL & Our LoRA based data augmentation method.

ControlNet ablation. To test whether the LoRA contributes meaningfully on top of ControlNet’s structural prior, we generate an additional 200 images per artist under the same ControlNet-Canny setting as Set B but *without* the LoRA adapter, conveying style only through prompt hints (e.g., “oil painting, swirling impasto brushstrokes, post-impressionist”). As shown in Figure 7, without LoRA the images show little stylistic variation across artists and exhibit blurry brushwork, indicating that structural priors alone cannot capture artist identity; re-introducing LoRA yields richer details, closer alignment with each artist’s style, and markedly clearer separation between artists. Therefore, in this paper, all Canny versions we report use Canny with LoRA.

4.3 Classifier retraining and the REAL_FRAC sampler

A natural way to use synthetic data is to append it to the real training set and use the standard inverse-frequency WeightedRandomSampler (weight $\propto 1/n_c$, where n_c counts both real and synthetic in class c). For a worst-10 class such as Paul Cezanne, which has 37 real and 200 synthetic images, this rule causes roughly 84% of in-class samples drawn during training to be synthetic. The classifier ends up learning LoRA-generation *style* rather than the real artist, and because the test set is 100% real, the

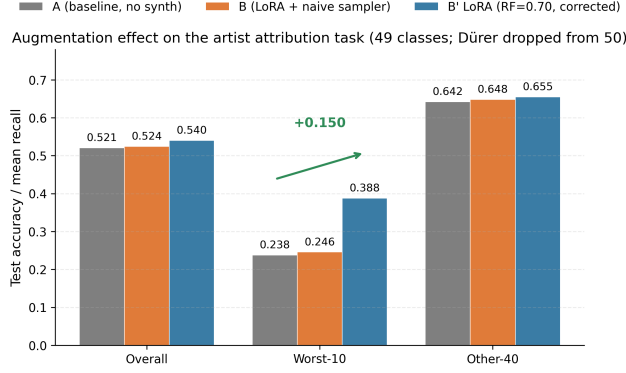


Figure 8: Test-set accuracy and mean per-class recall for three training regimes on the 49-class task. Naive LoRA augmentation (orange) barely moves any metric; the corrected within-class sampler at REAL_FRAC=0.70 (blue) lifts worst-10 mean recall from 0.238 to 0.388 (+0.150) without harming head accuracy.

learned features fail to transfer. We confirmed this empirically: naive augmentation (Experiment B in Section 5) produces essentially no improvement.

We therefore introduce a per-class REAL_FRAC weight rule. For each class c with n_c^{real} real and n_c^{synth} synthetic samples, define

$$w(x) = \begin{cases} \frac{\rho}{n_c^{\text{real}}} & \text{if } x \text{ is real and in class } c, \\ \frac{1-\rho}{n_c^{\text{synth}}} & \text{if } x \text{ is synthetic and in class } c, \end{cases} \quad (1)$$

where $\rho = \text{REAL_FRAC} \in [0, 1]$. By construction $\sum_{x \in c} w(x) = 1$ for every class, so across-class sampling probability is uniform at $1/49$ regardless of ρ ; the only quantity that changes is the within-class real:synthetic ratio. This decoupling lets us sweep ρ as a clean ablation, and recovers the real-only fair baseline at $\rho = 1$ and the naive (synth-dominated) sampler at small ρ .

Each classifier is trained with the same protocol across all settings: Adam optimizer, learning rate 1×10^{-3} , StepLR decay $\gamma = 0.5$ every 3 epochs, 10 epochs total, batch size 32, the same stratified split. All differences between experiments arise from the training-set composition and the sampler, not from the optimization protocol.

5 Results and Interpretation

Table 2 reports overall accuracy and mean recall on the worst-10 and other-40 classes across baselines, naive augmentation, our corrected sampler at six values of ρ , where ρ defines the fraction of the training data sourced from the original, and the ControlNet+Canny ablation. Figure 8 summarises the three headline regimes. Several findings stand out.

Setup fairness check. The fair re-implementation A' ($\rho=1$, no synthetic, our code) lands at 0.523 against the trained reference A at 0.521—a 0.002 difference, well within run-to-run noise. This confirms that any subsequent A-vs-B' improvement is attributable to the augmentation mechanism, not to differences in the training script.

Naive augmentation produces no improvement. Experiment B—synthetic data appended, standard sampler—yields +0.4pp overall and +0.8pp on worst-10. Inspecting per-class behavior reveals the underlying problem: although Eugene Delacroix gains 33pp, Titian *loses* 19pp and Kandinsky loses 22pp under this regime. The net effect is a wash, and the per-class instability indicates the sampler is over-weighting synthetic noise in some classes while leaving others unchanged.

The corrected sampler traces an inverted-U. Sweeping ρ reveals a clear bias-variance pattern. Worst-10 recall is 0.295 at $\rho=1.00$ (no synthetic in-class), rises to a maximum of **0.388** at $\rho=0.70$,

Table 2: Main results. “Worst-10” is mean recall over the ten artists with the lowest baseline recall; “Other-40” is mean recall over the remaining 40. Bold marks the best result in each column. The naive sampler (B) nearly ties baseline (A), but the corrected sampler at $\rho=0.70$ lifts worst-10 by 15pp.

Setting	ρ	Overall acc.	Worst-10	Other-40
A: real-only baseline (frozen ResNet-18)	—	0.521	0.238	0.642
B: naive sampler, LoRA synth appended	—	0.524	0.246	0.648
B': corrected sampler, $\rho=1.00$ (= “fair A’”)	1.00	0.523	0.295	0.635
B': corrected sampler	0.85	0.543	0.311	0.656
B': corrected sampler	0.70	0.540	0.388	0.655
B': corrected sampler	0.50	0.519	0.285	0.652
B': corrected sampler	0.30	0.517	0.280	0.659
B': corrected sampler	0.16	0.526	0.203	0.681
B' ControlNet+Canny synth	0.70	0.532	0.281	0.674

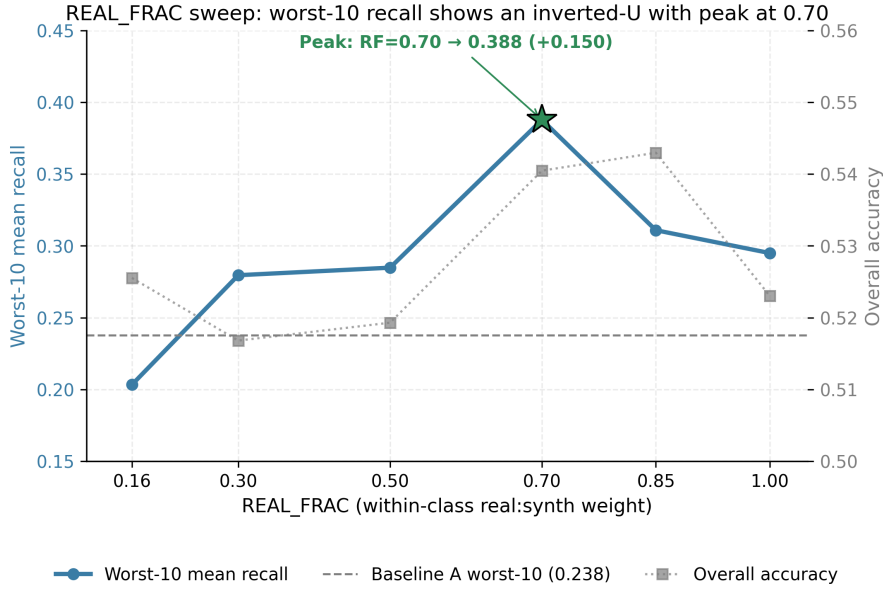


Figure 9: REAL_FRAC sweep: worst-10 mean recall (blue, left axis) and overall accuracy (gray dotted, right axis). Worst-10 recall peaks sharply at REAL_FRAC=0.70 and falls below the no-augmentation baseline at REAL_FRAC=0.16, showing that the within-class real:synthetic balance dominates over the absolute amount of synthetic data.

falls to 0.285 at $\rho=0.50$, and collapses to 0.203 at $\rho=0.16$ (worse than the no-synthetic baseline). The $\rho=0.16$ setting is approximately what the naive sampler implements for worst-10 classes, and indeed it reproduces the near-null behavior of Experiment B. This both confirms our root-cause analysis and identifies $\rho=0.70$ as the operating point where synthetic data contributes most without overwhelming the real distribution. The training/validation gap also grows monotonically as ρ decreases (from 0.152 at $\rho=1$ to 0.233 at $\rho=0.16$), consistent with the classifier increasingly memorizing the synthetic distribution.

Per-artist analysis. At $\rho=0.70$, six of ten worst-10 artists improve, three are flat, and one (Titian) regresses by 4pp (Table 3, appendix). The largest absolute gains are Delacroix (0.33 $\rightarrow$ 1.00; small test $n=3$ caveats apply), Monet (0.29 $\rightarrow$ 0.57), Andrei Rublev (0.40 $\rightarrow$ 0.60), and Paul Cezanne (0.00 $\rightarrow$ 0.20, a robust signal across multiple ρ values). Vasiliy Kandinsky is the notable failure mode: we never exceed the baseline at any ρ . We attribute this to LoRA’s struggle to capture Kandinsky’s highly abstract geometric style from 40 reference images—a regime where the diffusion

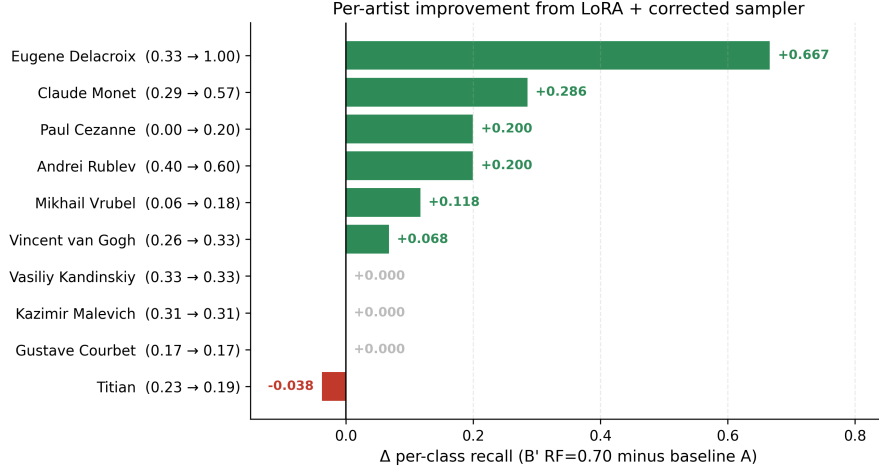


Figure 10: Per-class recall change $B'_{\rho=0.70} - A$ for the worst-10 artists (each row also shows $A \rightarrow B'$). Six of ten improve, three are flat, one (Titian) regresses; Delacroix (+0.667) and Monet (+0.286) drive most of the mean improvement.

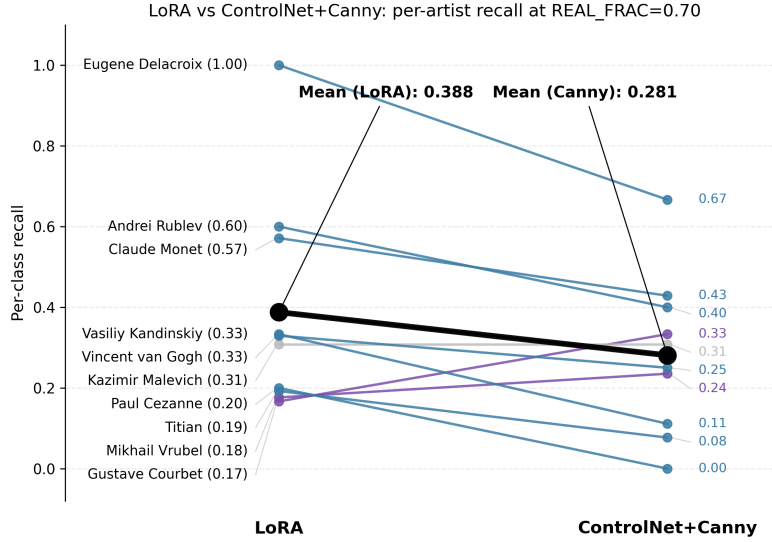


Figure 11: Per-artist recall, LoRA vs. ControlNet+Canny synthetics, both at REAL_FRAC=0.70. LoRA wins on aggregate (0.388 vs. 0.281, black) and on stylistic artists (Delacroix, Monet, Cezanne); Canny wins on line/composition-driven artists (Courbet, Vrubel). The two strategies are complementary, not strictly ranked.

prior dominates the LoRA update and the resulting synthetic samples drift away from the artist's distribution.

LoRA vs. ControlNet+Canny. Holding $\rho=0.70$ fixed and swapping the synthetic source, LoRA scores 0.388 on worst-10 vs. 0.281 for ControlNet+Canny. LoRA wins on 7 of 10 artists, particularly the stylistic ones: Delacroix (LoRA 1.00 vs. Canny 0.67), Monet (0.57 vs. 0.43), Cezanne (0.20 vs. 0.00). ControlNet *wins* on Courbet (+17pp) and Vrubel (+6pp), both of whom are line- or composition-driven realists. Our interpretation is that style-faithful generation (LoRA fine-tuned to the artist's palette and brushwork) helps when artist identity lives in texture, while structure-preserving generation (Canny edges from real photographs) helps when artist identity lives in composition. The two methods are complementary, not strictly ranked.

Trade-offs. The best overall accuracy (0.543, $\rho=0.85$) and the best worst-10 recall (0.388, $\rho=0.70$) come from different operating points. At $\rho=0.16$ we also observe that the other-40 mean recall actually improves slightly (to 0.681), a counter-intuitive byproduct: when synthetic data dominates the loss for worst-10 classes, the gradient signal contributed by those classes is effectively noise, freeing model capacity to specialize on the clearer head classes. For a deployed system, the appropriate operating point depends on whether tail fairness or head accuracy is the primary objective.

6 Conclusions and Discussion

Across roughly a dozen trained classifiers, two findings dominate. First, *sampling matters more than generator quality*. The same 2,000 LoRA-generated images move worst-10 recall from -3.5pp under the naive sampler to $+15\text{pp}$ under the corrected sampler. Second, *generator choice matters for which classes improve*. LoRA captures style and helps stylistic artists; ControlNet+Canny preserves structure and helps line-driven artists. Treating “generative augmentation” as a single intervention misses this dimension.

The improvements are real but bounded. Our worst-10 recall ceiling (~ 0.39) is still well below the head-class average (~ 0.66), and the overall test-accuracy gain over the frozen-backbone baseline is $\sim 2\text{pp}$. We do not claim this method is a substitute for a stronger backbone or end-to-end fine-tuning; rather, we show that on the imbalanced-classification regime where most of the trouble lives, diffusion-based augmentation can recover a meaningful fraction of the worst-class gap, provided the sampler is structured to exploit it.

Limitations. (i) Worst-10 test sets are small (3–15 samples/class), so per-artist recalls are noisy; we only report trends confirmed across multiple ρ values. (ii) We did not sweep ρ on a stronger backbone (end-to-end ResNet-18 or ViT-B/16), so whether the $+15\text{pp}$ lift transfers to a higher-accuracy starting point remains open. (iii) Per-artist LoRAs used only 40 paintings, leaving headroom for richer adapters via more data, longer training, or higher rank. (iv) Generation prompts were held fixed across artists; conditioning on artist-typical subjects might further tighten the synthetic distribution.

Future work. The most natural next step is fusing LoRA and ControlNet in one pipeline (style adapter + structural prior), since their per-artist complementarity suggests an additive gain. Other direct extensions include covering all 50 artists (restoring Dürer and the 39 head classes), evaluating on out-of-distribution collections such as WikiArt, and applying the REAL_FRAC sampler to other long-tail vision tasks.

7 Broader Impact

Attribution tools like ours have direct positive uses: supporting art historians on under-catalogued artists, aiding provenance research, and providing first-pass screens for suspected forgeries. Crucially, the same generative pipeline that augments training data could also produce convincing painting-style forgeries, so any deployed attribution system should double as a deepfake detector and be evaluated against synthetics from its own generator. Methodologically, our finding that naive “more data” fails while the same data succeeds under a corrected sampler highlights a broader lesson for generative augmentation in vision and NLP: outcomes depend at least as much on how augmented data enters the loss as on how it was produced. Reporting both choices explicitly, as we do, is a small step toward reproducibility for this class of result.

Use of generative AI

Per course policy, we disclose that Anthropic’s Claude (Opus 4.7) was used in the preparation of this manuscript as a writing and editing assistant, including for drafting initial section structure, suggesting phrasings, and reviewing prose for clarity. All experimental work, results, code, and final scientific claims are the authors’ own. No generative-AI tools were used to produce numerical results or figures.

References

- [1] He, K., Zhang, X., Ren, S., & Sun, J. (2016) Deep residual learning for image recognition. In *Proc. CVPR*, pp. 770–778.
- [2] Dosovitskiy, A. et al. (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*.
- [3] Radford, A. et al. (2021) Learning transferable visual models from natural language supervision. In *Proc. ICML*, pp. 8748–8763.
- [4] Tan, M. & Le, Q. (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proc. ICML*, pp. 6105–6114.
- [5] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022) High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pp. 10684–10695.
- [6] Podell, D. et al. (2023) SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*.
- [7] Hu, E. et al. (2021) LoRA: Low-rank adaptation of large language models. *arXiv:2106.09685*.
- [8] Ruiz, N. et al. (2023) DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. CVPR*, pp. 22500–22510.
- [9] Zhang, L., Rao, A., & Agrawala, M. (2023) Adding conditional control to text-to-image diffusion models. In *Proc. ICCV*, pp. 3836–3847.
- [10] Johnson Jr., C. R. et al. (2008) Image processing for artist identification. *IEEE Signal Processing Magazine* 25(4):37–48.
- [11] Ikarus777. *Best Artworks of All Time*. Kaggle dataset. <https://www.kaggle.com/datasets/ikarus777/best-artworks-of-all-time>.
- [12] von Platen, P. et al. *Diffusers: State-of-the-art diffusion models*. Hugging Face. <https://github.com/huggingface/diffusers>.
- [13] Anthropic. *Claude Opus 4.7*. Used as writing assistant. <https://www.anthropic.com/claude>.

A Supplementary results

Table 3: Per-artist recall on the worst-10 classes for the three main experiments. “ $\Delta_{B'-A}$ ” is the absolute change relative to the baseline (A). Test counts are small for several classes (Andrei Rublev $n=5$, Eugene Delacroix $n=3$), so individual values should be read together with the aggregate trend, not in isolation.

Artist	A (base)	B (naive)	B' ($\rho=0.70$)	$\Delta_{B'-A}$
Paul Cezanne	0.000	0.000	0.200	+0.200
Mikhail Vrubel	0.059	0.118	0.176	+0.117
Gustave Courbet	0.167	0.167	0.167	+0.000
Titian	0.231	0.038	0.192	−0.039
Vincent van Gogh	0.261	0.364	0.330	+0.069
Claude Monet	0.286	0.286	0.571	+0.285
Kazimir Malevich	0.308	0.308	0.308	+0.000
Eugene Delacroix	0.333	0.667	1.000	+0.667
Vasiliy Kandinskiy	0.333	0.111	0.333	+0.000
Andrei Rublev	0.400	0.400	0.600	+0.200
Mean	0.238	0.246	0.388	+0.150

LoRA vs. ControlNet+Canny, per artist (at $\rho=0.70$). LoRA wins on 7/10 artists, ControlNet wins on Courbet (0.33 vs. 0.17), Vrubel (0.24 vs. 0.18), and is tied or near-tied elsewhere. Full table is saved in the supplementary CSV (`lora_vs_canny_per_artist.csv`).

Reproducibility. SEED=42 throughout. All splits, sampler RNG, and model initializations are seeded; only diffusion-sampling stochasticity (per-prompt generator seed) is varied across reps to obtain a diverse synthetic set. Trained model predictions for every setting are saved as `preds_*.npz` files (containing predictions, ground truth, and probabilities) so that post-hoc analysis can be replicated without retraining.

B Supplementary figures

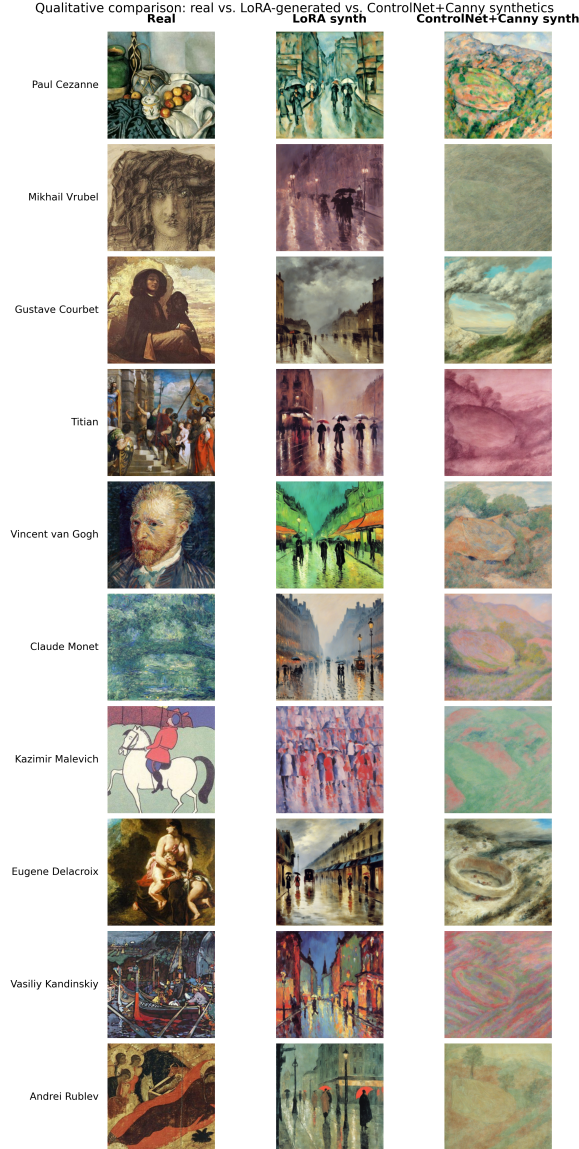


Figure 12: Qualitative comparison for each worst-10 artist: a real painting (left), one LoRA-generated synthetic (middle), and one ControlNet+Canny-generated synthetic (right). All synthetics use the same content prompt / Canny reference across artists so that the only varying factor is the artist-specific style; LoRA visibly captures per-artist palette and brushwork, while ControlNet+Canny preserves the structure of the reference photo but with weaker style transfer.

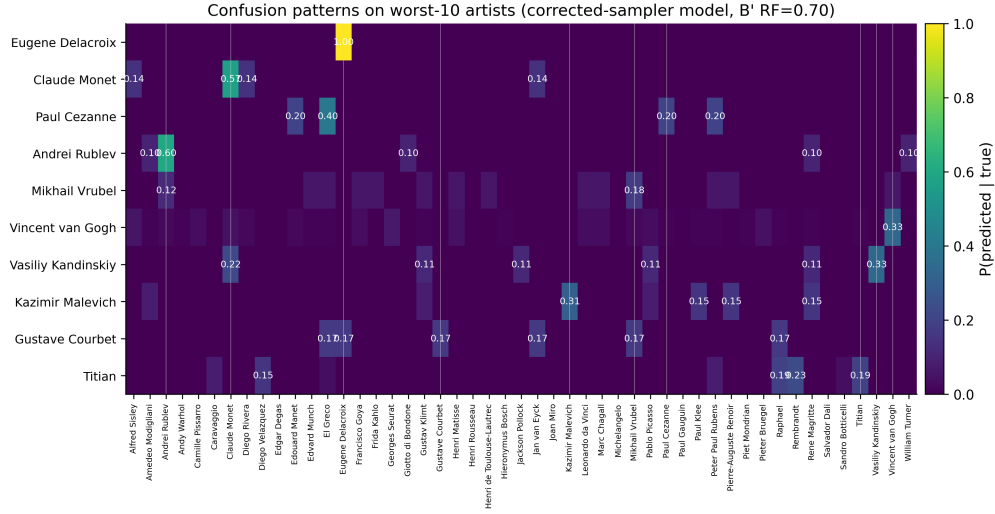


Figure 13: Row-normalised confusion matrix restricted to the 10 worst-performing artists (rows, ordered by Δ from Figure 10) against all 49 trained classes (columns, alphabetical) for the corrected-sampler model B' at $\rho=0.70$. Each cell shows $P(\text{predicted} | \text{true})$; off-diagonal mass concentrates on stylistically adjacent artists (e.g. Monet $\leftrightarrow$ Cezanne), showing that remaining errors are systematic stylistic confusions rather than random.

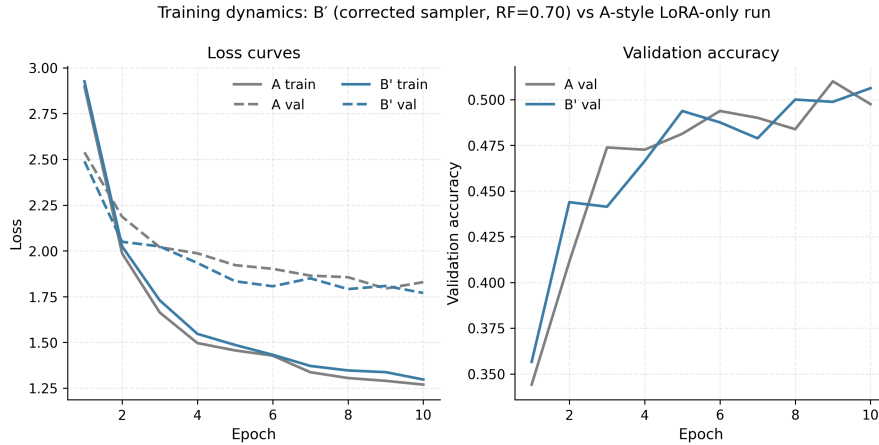


Figure 14: Per-epoch train and validation loss (left) and validation accuracy (right) for the B' $\rho=0.70$ run versus an A-style no-synthetic run trained with the same protocol. B' training dynamics closely track A's throughout the 10 epochs, indicating the corrected sampler does not induce additional overfitting compared to the no-augmentation baseline.