APPENDIX: MATHEMATICAL THEOREMS AND COROLLARY PROOFS

Theorem 1 *For any two data distributions $\mathscr{P}_1$ and $\mathscr{P}_2$, if their distribution difference in the original space is $\mathscr{D}(\mathscr{P}_1, \mathscr{P}_2) < \varepsilon$, then there exists an optimal set of diffusion parameters $\{\alpha_t^*\}_{t=1}^T$ such that the distribution difference after DMFE feature extraction is $\mathscr{D}(\phi_{DMFE}^*(\mathscr{P}_1), \phi_{DMFE}^*(\mathscr{P}_2)) > \delta$, where $\delta \gg \varepsilon$, and satisfies:*

$$\{\alpha_t^*\}_{t=1}^T = \arg\max_{\{\alpha_t\}_{t=1}^T} \mathscr{D}(\phi_{DMFE}(\mathscr{P}_1), \phi_{DMFE}(\mathscr{P}_2)) - \lambda \sum_{t=1}^{T} (1-\alpha_t) \tag{1}$$

where λ is a regularization parameter balancing distribution separability and diffusion process stability.

Proof 1 *The Wasserstein-1 distance is defined via optimal transport: $W_1(\mathscr{P}, \mathscr{Q}) = \inf_{\gamma \in \Gamma(\mathscr{P}, \mathscr{Q})} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}[\|\mathbf{x}-\mathbf{y}\|]$ where $\Gamma(\mathscr{P}, \mathscr{Q})$ is the set of all couplings. For a Lipschitz-continuous function ϕ with constant $L_\phi = \sup_{\mathbf{x}\neq\mathbf{y}} \|\phi(\mathbf{x}) - \phi(\mathbf{y})\| / \|\mathbf{x}-\mathbf{y}\|$, the contraction property gives:*

$$W_1(\phi(\mathscr{P}), \phi(\mathscr{Q})) \leq L_\phi \cdot W_1(\mathscr{P}, \mathscr{Q}) \tag{2}$$

For a composition of Lipschitz functions $\phi = \phi_K \circ \cdots \circ \phi_1$ with individual constants L_i, the composed Lipschitz constant is the product:

$$L_\phi = \prod_{i=1}^{K} L_i \tag{3}$$

The DMFE architecture decomposes as sequential operations. For the diffusion process, at each timestep:

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\varepsilon_{t-1} \tag{4}$$

The Lipschitz constant of this operation satisfies:

$$L_t = \sup_{\mathbf{x}_{t-1}\neq\mathbf{x}'_{t-1}} \frac{\|\sqrt{\alpha_t}\mathbf{x}_{t-1} - \sqrt{\alpha_t}\mathbf{x}'_{t-1}\|}{\|\mathbf{x}_{t-1} - \mathbf{x}'_{t-1}\|} = \sqrt{\alpha_t} \tag{5}$$

Composing T diffusion steps yields:

$$L_{diff} = \prod_{t=1}^{T} \sqrt{\alpha_t} = \left(\prod_{t=1}^{T} \alpha_t\right)^{1/2} \tag{6}$$

For conservative diffusion schedules with $\alpha_t \in [0.9, 0.99]$, we have $\prod_t \alpha_t \in [0.64, 1.00]$, giving $L_{diff} \in [0.80, 1.00]$.

The frequency domain transforms (Fourier $\mathscr{F}$ and wavelet $\mathscr{W}$) are orthogonal linear operators satisfying:

$$\|\mathscr{F}(\mathbf{x})\| = \|\mathbf{x}\|, \quad \|\mathscr{W}(\mathbf{x})\| = \|\mathbf{x}\| \tag{7}$$

thus $L_{freq} = 1$.

The convolutional feature extractors apply K_{conv} layers. Under spectral normalization, the normalized weight matrix has $\|\mathbf{W}_i^{norm}\|_{op} = 1$. With learnable scaling factor s_i, the actual Lipschitz constant becomes:

$$L_i = (1+s_i)\|\mathbf{W}_i^{norm}\|_{op} = 1 + s_i \tag{8}$$

Composing all convolutional layers:

$$L_{conv} = \prod_{i=1}^{K_{conv}} (1+s_i) \tag{9}$$

Under the constraint $\sum_{i=1}^{K_{conv}} s_i \leq S_{\max}$, maximizing the product occurs with uniform scaling $s_i = S_{\max}/K_{conv}$:

$$L_{conv}^{\max} = \left(1 + \frac{S_{\max}}{K_{conv}}\right)^{K_{conv}} \tag{10}$$

For $S_{\max} = 1$ and $K_{conv} = 16$: $(1.0625)^{16} \approx 2.64$. For $S_{\max} = 1.5$ and $K_{conv} = 12$: $(1.125)^{12} \approx 3.22$.
The covariance pooling and patch aggregation operations contribute:

$$L_{pool} = 1 + c_{pool} \in [1.05, 1.15], \quad L_{fusion} = 1 + c_{fusion} \in [1.05, 1.20] \tag{11}$$

where c_{pool}, c_{fusion} are empirically determined constants from feature statistics.
By the composition rule, the total Lipschitz constant is:

$$L_{total} = L_{diff} \cdot L_{freq} \cdot L_{conv} \cdot L_{pool} \cdot L_{fusion} \tag{12}$$

For conservative settings with $L_{diff} \approx 0.95$, $L_{conv} \approx 2.64$, $L_{pool} \approx 1.10$, $L_{fusion} \approx 1.10$:

$$L_{total} \approx 0.95 \times 1 \times 2.64 \times 1.10 \times 1.10 \approx 3.0 \tag{13}$$

For aggressive settings with $L_{diff} \approx 1.00$, $L_{conv} \approx 3.22$, $L_{pool} \approx 1.15$, $L_{fusion} \approx 1.20$:

$$L_{total} \approx 1.00 \times 1 \times 3.22 \times 1.15 \times 1.20 \approx 4.45 \tag{14}$$

By the Lipschitz contraction property:

$$W_1(\phi_{DMFE}(\mathscr{P}_1), \phi_{DMFE}(\mathscr{P}_2)) \leq L_{total} \cdot W_1(\mathscr{P}_1, \mathscr{P}_2) = L_{total} \cdot \varepsilon_0 \tag{15}$$

which establishes the stated amplification bound.

Corollary 1 *Given a set of AI generators $\{G_1, G_2, \ldots, G_M\}$ and their corresponding image distributions $\{\mathscr{P}_g^1, \mathscr{P}_g^2, \ldots, \mathscr{P}_g^M\}$, if DMFE can effectively distinguish between the real image distribution $\mathscr{P}_r$ and the distribution of any known generator, then for the distribution $\mathscr{P}_g^{new}$ produced by a newly emerging unknown generator G_{new}, DMFE can still maintain a certain discrimination ability, with its generalization performance lower bound:*

$$\mathscr{D}(\phi_{DMFE}^*(\mathscr{P}_r), \phi_{DMFE}^*(\mathscr{P}_g^{new})) \geq \min_{j \in \{1,2,\ldots,M\}} \mathscr{D}(\phi_{DMFE}^*(\mathscr{P}_r), \phi_{DMFE}^*(\mathscr{P}_g^j)) - \gamma \cdot \max_{j \in \{1,2,\ldots,M\}} \mathscr{D}(\mathscr{P}_g^j, \mathscr{P}_g^{new}) \tag{16}$$

where γ is a constant related to the complexity of the DMFE model. This corollary indicates that DMFE's generalization ability to unknown generators depends on its discrimination performance for known generators and the similarity between the distributions of new generators and known generators.

Proof 2 *The Wasserstein distance satisfies the triangle inequality and its reverse form. By the reverse triangle inequality:*

$$|W_1(\mathscr{P}_1, \mathscr{P}_3) - W_1(\mathscr{P}_1, \mathscr{P}_2)| \leq W_1(\mathscr{P}_2, \mathscr{P}_3) \tag{17}$$

Rearranging:

$$W_1(\mathscr{P}_1, \mathscr{P}_3) \geq W_1(\mathscr{P}_1, \mathscr{P}_2) - W_1(\mathscr{P}_2, \mathscr{P}_3) \tag{18}$$

Apply this in the feature space. For any known generator index j:

$$W_1(\phi_{DMFE}(\mathscr{P}_r), \phi_{DMFE}(\mathscr{P}_g^{new})) \geq W_1(\phi_{DMFE}(\mathscr{P}_r), \phi_{DMFE}(\mathscr{P}_g^j)) - W_1(\phi_{DMFE}(\mathscr{P}_g^j), \phi_{DMFE}(\mathscr{P}_g^{new})) \tag{19}$$

By the Lipschitz contraction property from Theorem **??***:*

$$W_1(\phi_{DMFE}(\mathscr{P}_g^j), \phi_{DMFE}(\mathscr{P}_g^{new})) \leq L_{total} \cdot W_1(\mathscr{P}_g^j, \mathscr{P}_g^{new}) \leq L_{total} \cdot \varepsilon_{novel} \tag{20}$$

Substituting back:

$$W_1(\phi_{DMFE}(\mathscr{P}_r), \phi_{DMFE}(\mathscr{P}_g^{new})) \geq W_1(\phi_{DMFE}(\mathscr{P}_r), \phi_{DMFE}(\mathscr{P}_g^j)) - L_{total} \cdot \varepsilon_{novel} \tag{21}$$

Taking the minimum over all j*:*

$$W_1(\phi_{DMFE}(\mathscr{P}_r), \phi_{DMFE}(\mathscr{P}_g^{new})) \geq \min_j W_1(\phi_{DMFE}(\mathscr{P}_r), \phi_{DMFE}(\mathscr{P}_g^j)) - L_{total} \cdot \varepsilon_{novel} \tag{22}$$

This demonstrates that discrimination ability on unknown generators degrades linearly with their distributional distance from known generators, with proportionality constant L_{total}*.*

Theorem 2 *Given two data distributions* $\phi^*_{DMFE}(\mathscr{P}_r)$ *and* $\phi^*_{DMFE}(\mathscr{P}_g)$ *in the feature space extracted by DMFE, if CAViT can effectively capture higher-order relationships of these features, then the cascaded ensemble model DCAT can further increase the distribution difference, satisfying:*

$$\begin{aligned} \mathscr{D}(\phi_{DCAT}(\mathscr{P}_r), \phi_{DCAT}(\mathscr{P}_g)) \geq\ & \eta \cdot \mathscr{D}(\phi^*_{DMFE}(\mathscr{P}_r), \phi^*_{DMFE}(\mathscr{P}_g)) \\ & + \kappa \cdot KL(\phi^*_{DMFE}(\mathscr{P}_r) \| \phi^*_{DMFE}(\mathscr{P}_g)) \\ & + \gamma \cdot \left(1 - \exp\left(-\beta \cdot \|\mathbb{E}_{\mathbf{x}\sim\mathscr{P}_r}[\phi^*_{DMFE}(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathscr{P}_g}[\phi^*_{DMFE}(\mathbf{x})]\|_2^2\right)\right) \\ & - \alpha \cdot \sqrt{Var_{\mathbf{x}\sim\mathscr{P}_r}[\phi^*_{DMFE}(\mathbf{x})] + Var_{\mathbf{x}\sim\mathscr{P}_g}[\phi^*_{DMFE}(\mathbf{x})]} \end{aligned} \tag{23}$$

where $\mathscr{D}(\cdot,\cdot)$ *is the distribution difference measure, as defined in equation (1);* $\eta > 1$ *is the basic enhancement factor, positively correlated with CAViT's expressive power and parameter count;* $\kappa > 0$ *is the weight coefficient for KL divergence;* $KL(\cdot\|\cdot)$ *is the Kullback-Leibler divergence;* $\gamma > 0$ *and* $\beta > 0$ *are coefficients for the mean difference term;* $\alpha > 0$ *is the variance penalty coefficient;* $\mathbb{E}$ *and Var represent expectation and variance, respectively;* $\phi_{DCAT}(\mathscr{P})$ *represents the distribution after mapping distribution* $\mathscr{P}$ *through the DCAT model. This theorem shows that CAViT not only linearly amplifies the distribution differences produced by DMFE but also further enhances discrimination ability through modeling of higher-order statistics.*

Proof 3 *The global path in CAViT applies* L_g *sequential attention layers. The softmax operation in attention has Lipschitz constant 1 because the Jacobian eigenvalues of softmax satisfy* $\lambda_{softmax} \leq 1$*. The query-key bilinear form* $\mathbf{Q}\mathbf{W}^Q(\mathbf{K}\mathbf{W}^K)^T$ *has operator norm at most 1 under spectral normalization. Each attention head produces dimension* $d_v = D/h$*. When* h *independent attention pathways merge through output projection* $\mathbf{W}^O$*, gradient amplification proportional to* $\sqrt{h}$ *occurs (from variance of sum). Simultaneously, reduced per-head dimension* $d_k = D/h$ *increases attention logits by factor* $\sqrt{h}$ *relative to full dimension, amplifying gradient flow. The composite effect yields:*

$$L_{attn} = 1 + c_1 \frac{\sqrt{h}}{\sqrt{d_k}} = 1 + c_1 \frac{h}{\sqrt{D}} \tag{24}$$

Composing L_g *such layers via the product rule:*

$$L_{global} = \prod_{l=1}^{L_g} (1 + \varepsilon_{attn}) = (1 + \varepsilon_{attn})^{L_g} =: \eta \tag{25}$$

For typical settings with $L_g = 12$*,* $h = 8$*,* $D = 512$*:* $\varepsilon_{attn} = 0.5 \times 8/\sqrt{512} \approx 0.088$*, giving* $\eta = (1.088)^{12} \approx 2.82$*. Thus:*

$$W_1(GlobalPath(\mathscr{P}_r), GlobalPath(\mathscr{P}_g)) \geq \eta \cdot \delta_0 \tag{26}$$

The KL divergence between distributions is:

$$KL(\mathscr{P}_r \| \mathscr{P}_g) = \int \log \frac{p_r(\mathbf{x})}{p_g(\mathbf{x})} p_r(\mathbf{x}) d\mathbf{x} \tag{27}$$

Pinsker's inequality relates KL divergence to total variation distance:

$$TV(\mathscr{P}_r, \mathscr{P}_g) = \frac{1}{2}\int |p_r(\mathbf{x}) - p_g(\mathbf{x})| d\mathbf{x} \leq \sqrt{\frac{KL(\mathscr{P}_r \| \mathscr{P}_g)}{2}} \tag{28}$$

For bounded supports with diameter D_{supp}, Wasserstein distance relates to total variation by:

$$W_1(\mathscr{P}_r, \mathscr{P}_g) \geq \frac{1}{D_{supp}} TV(\mathscr{P}_r, \mathscr{P}_g) \geq C_{supp}\sqrt{KL(\mathscr{P}_r \| \mathscr{P}_g)} \tag{29}$$

where $C_{supp} = 1/D_{supp}$. The local path's graph attention network with Lipschitz constant κ extracts this information:

$$W_1(LocalPath(\mathscr{P}_r), LocalPath(\mathscr{P}_g)) \geq \kappa \cdot C_{supp}\sqrt{KL(\mathscr{P}_r \| \mathscr{P}_g)} \tag{30}$$

The patch relationship modeling computes pairwise interactions with Lipschitz constant $L_{patch} = \gamma$. By Kantorovich duality:

$$W_1(\mathscr{P}_1, \mathscr{P}_2) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{\mathscr{P}_1}[f] - \mathbb{E}_{\mathscr{P}_2}[f] \tag{31}$$

Choosing linear functional in direction of $\Delta\mu_0$:

$$W_1(\mathscr{P}_1, \mathscr{P}_2) \geq \Delta\mu_0 - \sqrt{2\sigma_0^2} \tag{32}$$

For $\Delta\mu_0 \gg \sigma_0$, this approximates:

$$W_1(\mathscr{P}_1, \mathscr{P}_2) \geq \Delta\mu_0 \left(1 - \exp\left(-\frac{(\Delta\mu_0)^2}{c_{mv}\sigma_0^2}\right)\right) \tag{33}$$

where $c_{mv} \approx 4$ for Gaussian distributions. The patch modeling contributes:

$$W_1(PatchPath(\mathscr{P}_r), PatchPath(\mathscr{P}_g)) \geq \gamma\left(1 - \exp\left(-\frac{(\Delta\mu_0)^2}{c_{mv}\sigma_0^2}\right)\right) \tag{34}$$

The variance penalty $-\alpha\sigma_0$ prevents overfitting to high-overlap regions. The four mechanisms—spectral amplification, KL divergence, mean separation, and stability—operate on distinct statistical aspects. Their contributions can be combined additively via Kantorovich duality, as different optimal Lipschitz functions capture different distributional aspects:

$$W_1(\phi_{DCAT}(\mathscr{P}_r), \phi_{DCAT}(\mathscr{P}_g)) \geq \eta\delta_0 + \kappa C_{supp}\sqrt{KL} + \gamma\left(1 - \exp\left(-\frac{(\Delta\mu_0)^2}{c_{mv}\sigma_0^2}\right)\right) - \alpha\sigma_0 \tag{35}$$

Corollary 2 *Based on Theorem* **??**, *for any distribution pair satisfying $\mathscr{D}(\mathscr{P}_r, \mathscr{P}_g) < \varepsilon$, the discrimination performance of the DCAT ensemble model under optimal parameter configuration has a lower bound:*

$$\begin{aligned} AUC(\phi_{DCAT}) \geq\ & 1 - \exp\left(-\frac{\eta \cdot \delta^2}{2}\right) \\ & - \exp\left(-\lambda_1 \cdot KL(\mathscr{P}_r \| \mathscr{P}_g) - \lambda_2 \cdot JS(\mathscr{P}_r, \mathscr{P}_g)\right) \\ & - \frac{C_1}{\sqrt{n}} \cdot \left(1 + \sqrt{\log\frac{1}{\rho}}\right) \\ & + C_2 \cdot \left(1 - \frac{1}{1 + \exp(\omega \cdot (\|\blacksquare\|_F - \tau))}\right) \end{aligned} \tag{36}$$

where $AUC(\phi_{DCAT})$ represents the area under the receiver operating characteristic curve of model ϕ_{DCAT} on the binary classification task of real versus AI-generated images; $\delta > 0$ is the minimum distribution difference achieved by DMFE in the feature space, as guaranteed by Theorem **??**; *$\eta > 1$ is*

the enhancement factor provided by CAViT; λ_1 and λ_2 are weight coefficients for information-theoretic measures; $KL(\cdot\|\cdot)$ is the Kullback-Leibler divergence; $JS(\cdot,\cdot)$ is the Jensen-Shannon divergence; C_1 is a constant related to sample complexity; n is the number of training samples; ρ is a confidence parameter; C_2, ω, and τ are parameters related to model complexity; $\|\blacksquare\|_F$ is the Frobenius norm of the model parameters. This corollary comprehensively considers the influence of distribution differences, information-theoretic measures, sample complexity, and model complexity on model performance, ensuring that DCAT can achieve effective discrimination even when real and AI-generated images are extremely similar in the original image space, with performance improving as η and δ increase.

Proof 4 *For sub-Gaussian distributions in $\mathbb{R}$ with means μ_r, μ_g and combined variance σ^2_{DCAT}, the optimal AUC for the likelihood ratio test satisfies:*

$$AUC_{opt} = P(Z_g > Z_r) \approx \Phi\left(\frac{\mu_g - \mu_r}{\sqrt{2}\sigma_{DCAT}}\right) \tag{37}$$

where Φ is the standard normal CDF. Using the asymptotic expansion $1-\Phi(x) \approx (1/\sqrt{2\pi}x)\exp(-x^2/2)$:

$$AUC_{opt} \geq 1 - \frac{1}{2}\exp\left(-\frac{(\mu_g - \mu_r)^2}{8\sigma^2_{DCAT}}\right) \tag{38}$$

The Wasserstein distance between output distributions equals the mean difference:

$$W_1(Z_r, Z_g) = |\mu_g - \mu_r| = \Delta\mu_{DCAT} \tag{39}$$

*By Theorem **??**, this separation is at least $\eta\delta_0$:*

$$\Delta\mu_{DCAT} \geq \eta\delta_0 \tag{40}$$

Therefore:

$$AUC_{empirical} \geq 1 - \frac{1}{2}\exp\left(-\frac{(\eta\delta_0)^2}{8\sigma^2_{DCAT}}\right) \tag{41}$$

The empirical AUC on finite training data differs from true AUC. By Vapnik-Chervonenkis theory, for a hypothesis class with VC dimension V_{VC}:

$$|AUC_{empirical} - AUC_{true}| \leq \frac{C_1}{\sqrt{n}}\left(1 + \sqrt{\log(1/\rho)}\right) \tag{42}$$

where $C_1 = O(\sqrt{V_{VC}})$ and $V_{VC} = O(LD\log D)$ for deep networks with L layers of width D. Thus:

$$AUC_{true} \geq AUC_{empirical} - \frac{C_1}{\sqrt{n}}\left(1 + \sqrt{\log(1/\rho)}\right) \tag{43}$$

The KL divergence contributes through Chernoff information theory. The error probability under the optimal classifier satisfies:

$$P_{error} \leq \exp(-C_{Chernoff} \cdot KL(\mathscr{P}_r\|\mathscr{P}_g)) \tag{44}$$

for Chernoff constant $C_{Chernoff} > 0$. This translates to AUC contribution:

$$AUC_{info} \geq 1 - \exp(-\lambda_1 KL(\mathscr{P}_r\|\mathscr{P}_g)) \tag{45}$$

Combining all three sources—Wasserstein separation (architecture), generalization (learning theory), and information-theoretic advantage (distribution properties):

$$\begin{aligned} AUC(\phi_{DCAT}) \geq &\left(1 - \frac{1}{2}\exp\left(-\frac{(\eta\delta_0)^2}{8\sigma^2_{DCAT}}\right)\right) \\ &- \frac{C_1}{\sqrt{n}}\left(1 + \sqrt{\log(1/\rho)}\right) + (1 - \exp(-\lambda_1 KL(\mathscr{P}_r\|\mathscr{P}_g))) \end{aligned} \tag{46}$$