When Generative Replay Meets Evolving Deepfakes: Dual Confusion-Aware Regularization for Incremental Deepfake Detection

Anonymous authors
Paper under double-blind review

Abstract

The rapid advancement of face generation techniques has introduced an increasing variety of forgery methods, making incremental deepfake detection essential for maintaining robust detectors. While generative replay provides a promising solution to alleviate catastrophic forgetting without storing historical data, its effectiveness is hindered by **domain confusion** between generated samples and real data. We observe that replay samples fall into two categories: when the replay generator closely resembles the newly introduced forgery model, generated real samples overlap with the fake domain and become **domain-risky**; when the generator differs significantly, generated samples maintain clearer domain separation and can be treated as **domain-safe**. To address this challenge, we propose a Dual **C**onfusion-**A**ware **RE**gularization strategy, termed **Dual-CARE**. A Domain-aware Confusion Score (DC Score) is introduced to quantify domain confusion and **dual-modulate** the optimization of both replay generators and the incremental detector. Guided by DC Score, replay generators are updated to approximate previous-task distributions, while the detector adopts different supervision strategies: domain-safe samples are directly supervised, whereas domain-risky samples are regulated using a Relative Separation Loss to balance supervision and confusion. Extensive experiments demonstrate that Dual-CARE effectively exploits generative replay and improves detection performance under evolving forgery scenarios.

1 Introduction

The misuse of deepfakes severely threatens online trust. While current detection methods (Sun et al., 2022; Cao et al., 2022; Huang et al., 2023; Yan et al., 2024a; Cheng et al., 2024; 2025b) focus on leveraging existing samples to improve generalization, they struggle against rapidly evolving forgery techniques. Moreover, retraining models on combined datasets introduces high computational costs and privacy risks. Therefore, an incremental learning strategy is essential to enable continuous model adaptation.

To preserve historical task information, current incremental learning-based forgery detection methods typically utilize various sample replay strategies, such as central/hard sample replay (Pan et al., 2023), adversarial perturbations (Sun et al., 2025), mixed prototypes (Tian et al., 2024), and sparse uniform replay (Cheng et al., 2025a). However, as shown in Fig. 1, replaying original samples faces two major bottlenecks: insufficient data diversity due to restricted storage, and potential privacy and security risks. We provide a detailed ***Related Work*** in Appendix. Although generative replay avoids these issues by synthesizing samples instead of storing originals, it introduces a fundamental challenge: since the goal of forgery detection is to distinguish real data from synthetic, *can these generated "real" samples truly serve as authentic data for model training?* Motivated by this question, we make the first attempt to explore the feasibility of generative replay in forgery detection.

As shown in Fig. 2, we study generative replay using LDM-generated fake faces together with real faces to train a detection model. Four generators, LDM (Rombach et al., 2022a),

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

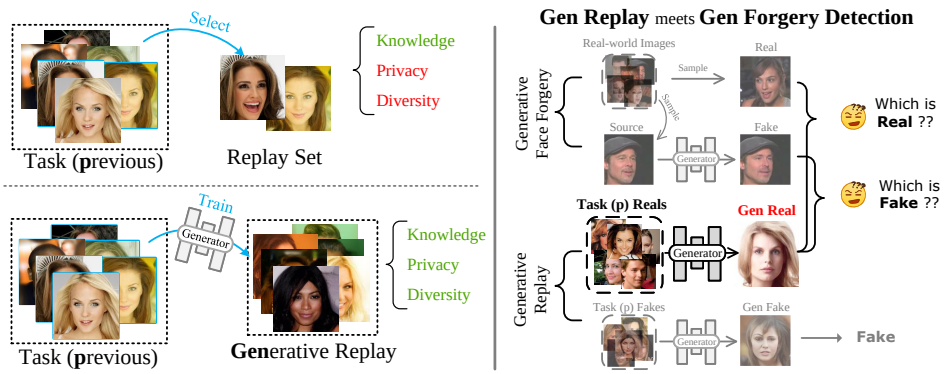


Figure 1: Left: Comparison between traditional sample replay and generative replay. Right: Generative replay makes forgery detection challenging because Gen-Real and Fake share similar generation processes, reducing their distinguishability.

LDM-I (Rombach et al., 2022a), DDIM (Song et al., 2020), and DDPM (Ho et al., 2020), are used to replay both fake and real distributions, with comparable FID values maintained to ensure fairness. The results show that when the replay generator closely resembles the fake model, such as LDM or LDM-I, the generated samples are poorly distinguished as real or fake. In contrast, generators that differ more from the fake model, such as DDIM or DDPM, achieve accurate detection. These observations suggest that generative replay can be effective in some scenarios but unreliable in others. We refer to the reliable cases as **domain-safe** and the unreliable ones as **domain-risky**. Understanding the underlying factors that determine whether a scenario is domain-safe or domain-risky is crucial for designing robust replay strategies.

To fully exploit generative replay for Generated Deepfake Detection, we propose a Dual Confusion-Aware REgularization (Dual-CARE) strategy. The key challenge lies in effectively utilizing replayed samples while mitigating the domain confusion introduced by generative artifacts. To address this, we first introduce a Domain-aware Confusion Score (DC Score) to quantify the degree of domain confusion during training, which is then used to **dual-modulate** the loss tradeoff in both generator updating and detector learning. Based on this score, we train a pair of updating replay generators that progressively simulate the distributions of all previous-task domains. A domain-aware fast training scheme guided by the DC Score is adopted to enhance the stability and efficiency of generator updating. For detector training, replayed samples are further categorized into "domain-safe" and "domain-risky" groups. The domain-safe samples are directly used for supervised training, while for domain-risky samples we introduce a Relative Separation Loss (RS Loss) to balance the supervision between informative forgery cues and potential confusion introduced by generative artifacts. Our contributions can be summarized as follows:

- We are the first to explore the feasibility of applying generative replay to generated face forgery detection, analyzing how the overlap between generated "real" and fake samples introduces domain risks that challenge traditional replay assumptions.
- We propose Dual-CARE to utilize replayed samples while mitigating interference from generative artifacts, introducing RS Loss and DC Score to adaptively balance information preservation and confusion suppression.
- Extensive experiments demonstrate that Dual-CARE consistently enhances incremental detection accuracy across various replay generators and alleviates the negative effects of domain overlap.

2 Motivation

As illustrated in Fig. 2, when the replay generator differs from the forgery generator, the generation artifacts shared by the replayed real and fake samples remain comparable and therefore do not interfere with the detector's learning. In contrast, when the replay generator closely resembles the forgery generator, its own generative artifacts are likely

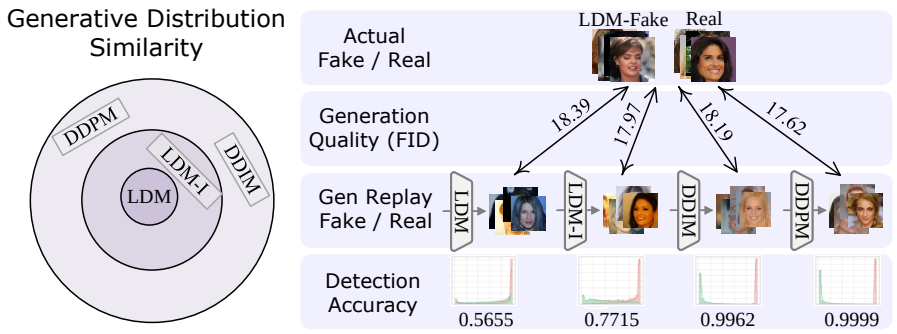


Figure 2: Influence of distribution similarity on generative replay. LDM-I (ODE version (Song et al., 2020) of LDM) generates images more similar to LDM than image-level DDPM (using SDE) and DDIM. All replay generators are controlled to have comparable FID scores, ensuring similar generation quality. However, under such a constrained condition, when the replay distribution is closer to the original fake distribution, detection accuracy drops, indicating that generative replay performs better when the replay generator differs more from the deepfake generator.

to be interpreted as forgery cues. This misalignment causes the replayed real samples to drift away from the true real distribution, thereby confusing the detector. We refer to this phenomenon as the **Domain Confusion Effect**.

Upon closer examination, this effect indicates that the replayed fake samples themselves do not substantially affect detection performance, regardless of the replay generator used. The core issue instead lies in the replayed real samples, which cannot always be safely used for training. Although their FID suggests that they approximate the original real distribution, these samples still contain subtle synthetic artifacts that can distort the classifier's decision boundary. Therefore, the key challenge is to leverage the informative content embedded in these replayed real samples while mitigating their generative bias. Moreover, when the replay generator is dissimilar to the original forgery generator, the replayed real samples tend to be less disruptive and can be directly used for classifier training. Consequently, developing a unified learning strategy that remains effective across different replay generators becomes essential for robust generative replay in forgery detection.

3 Methodology

3.1 Overall Framework

In this paper, we propose a Dual Confusion-Aware REgularization (Dual-CARE) strategy to fully exploit the potential of generative replay for generated deepfake detection. Specifically, we first introduce the Domain-aware Confusion Score (DC Score) to quantify the current state of the domain confusion issue, thus **dual-modulating** the loss tradeoff of both generator and detector training. Then, we train a pair of updating replay generators to generate replay data that simulates all previous-task domains, with a domain-aware fast training strategy that leverages DC Score to enhance generative performance. Finally, we deploy the "domain-safe" samples to directly supervise the incremental deepfake detector, and we introduce Relative Separation Loss (RS Loss) as a trade-off with the direct supervision for the "domain-risky" samples. Our overall framework is shown in Fig. 3.

3.2 Domain-aware Confusion Score

Here, we design a Domain-aware Confusion Score (DC Score) to quantify the domain confusion between current generative forgery data and the deployed replay generator. Since a distinct boundary between domain-risky and domain-safe data is natively absent, the DC Score adaptively measures this confusion to dynamically balance the weights of direct and

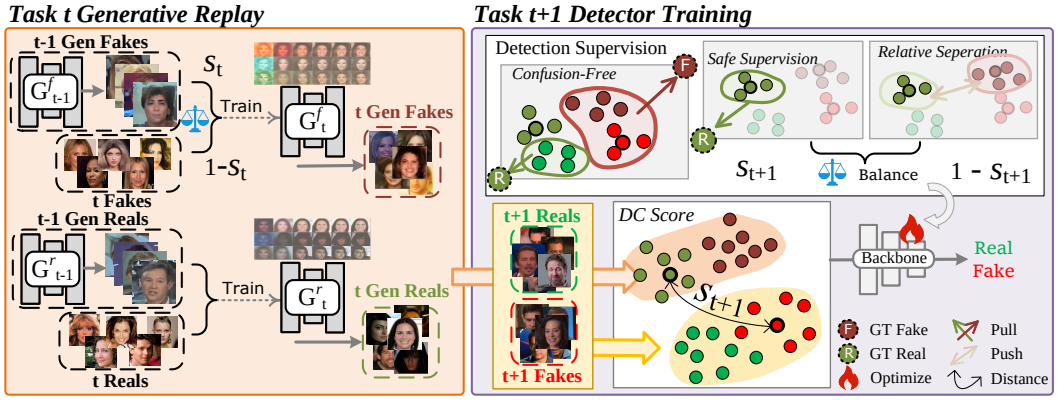


Figure 3: Overall framework of our method. We propose a dual confusion-aware regularization strategy to formulate the training of both generation and detection, thereby enabling the integration of generative replay into generative deepfake detection.

relative constraints. Formally, given a group of images $\mathbf{X}$, their centroid is defined as:

$$Cent(\mathbf{X}) = \frac{1}{|\mathbf{X}|} \sum_{\mathbf{x} \in \mathbf{X}} f(\mathbf{x}), \tag{1}$$

where $f(\cdot)$ is the feature extractor. Following Eq. 1, let $Cent(\mathcal{D}_t^{r\text{-gen}})$ and $Cent(\mathcal{D}_{t+1}^{f})$ be the centroids of the historical generated real samples $\mathcal{D}_t^{r\text{-gen}}$ and the current fake data $\mathcal{D}_{t+1}^{f}$, respectively. Their distance is then quantified via L_2 distance as:

$$s_{t+1} = \|Cent(\mathcal{D}_t^{r\text{-gen}}) - Cent(\mathcal{D}_{t+1}^{f})\|_2. \tag{2}$$

The normalized s_{t+1} is employed as the DC score to directly measure distribution separation, where a smaller value indicates a higher confusion risk.

3.3 Diffusion Replay Generation for IFFD

General Training Pipeline of Diffusion Generator. Given a set of deepfake detection training data as $\mathcal{D} = \{\mathcal{D}^f, \mathcal{D}^r\}$, we deploy Latent Diffusion Model (LDM) as our generator backbone considering the superior performance of the advanced diffusion model. Because $\mathcal{D}^f$ and $\mathcal{D}^r$ provide distinct optimization trajectories for the detector, we deploy two separate generators to model each domain respectively. Learning a single domain is defined as:

$$\mathcal{L}_{G^{(\cdot)}}(\mathcal{D}^{(\cdot)}) = \mathbb{E}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_s, s, \mathbf{c})\|_2^2\right], \tag{3}$$

where $\mathbf{x} \sim \mathcal{D}^{(\cdot)}$, $\epsilon \sim \mathcal{N}(0,1)$, $s \sim \mathcal{U}(1,S)$. Based on Eq. 3, we can obtain a pair of trained generators $\mathbf{G} = \{G^f, G^r\}$.

Generative Replay for All Previous Tasks in Deepfake Detection. Here, we aim to obtain replay samples $\mathbf{R}_t^{\text{gen}}$ for incremental learning at task $(t+1)$ through diffusion-based generative replay. With respect to the number of tasks, we maintain two jointly updated generators $\mathbf{G}_t = \{G_t^f, G_t^r\}$ to reproduce the historical fake and real distributions observed up to task t. In practice, the two generators are implemented using a fixed diffusion backbone with two lightweight LoRA adapters for the fake and real domains, respectively. Only the latest pair of adapters is retained after each incremental task, avoiding the need to store either historical raw samples or an increasing number of task-specific generators.

Specifically, following DDGR (Gao & Liu, 2023), we use the generators $\mathbf{G}_{t-1}$ obtained after task $(t-1)$ to synthesize a replay dataset $\mathcal{D}_{t-1}^{\text{gen}}$, which recursively represents the distributions of all tasks up to task $(t-1)$. The generated replay data are subsequently combined with the current-task dataset $\mathcal{D}_t$ to update $\mathbf{G}_t$. Formally, the training data for each generator can be written as:

$$\hat{\mathcal{D}}_t^{(\cdot)} = \{\mathcal{D}_t^{(\cdot)}, \mathcal{D}_{t-1}^{(\cdot)\text{-gen}}\}, \tag{4}$$

where $(\cdot) \in \{r, f\}$ denotes the real or fake domain. In this manner, $\mathbf{G}_t$ learns the current-task distribution while retaining the generative knowledge accumulated from previous tasks. See Sec. 4.3 for detailed computation analysis.

Domain-aware Fast Training. Since generation performance natively depends on the targeted distribution characteristics, we deploy a confusion coefficient to regularize the generator's training process. Specifically, we first decouple the training gradients induced by the previous and current data, which yields: $\mathcal{L}_{G^{(\cdot)}}(\hat{\mathcal{D}}_t^{(\cdot)}) = \mathcal{L}_{G^{(\cdot)}}(\mathcal{D}_t^{(\cdot)}) + \mathcal{L}_{G^{(\cdot)}}(\mathcal{D}_{t-1}^{(\cdot)\text{-gen}})$. For the real generator G^r, our objective is to approximate multiple real domains as faithfully as possible. The real domains and the synthetic artifacts introduced by the generator itself are not inherently conflicting during the generation process. The potential conflict with fake domains arises primarily during detection, which should therefore be addressed through appropriate design at the detector training stage. Consequently, G^r is expected to be effective in the acquisition of new knowledge while preserving previously learned knowledge. Hence, no additional weighting mechanism is required for G^r.

However, samples generated by G^f inherently carry intrinsic fabrication artifacts, complicating their relationship with newly learned forgery features. To stabilize training and enhance performance, we introduce a confusion coefficient to adaptively balance the training weights:

$$\mathcal{L}_{G^f}(\hat{\mathcal{D}}_t^f) = (1 - s_t) \times \mathcal{L}_{G^f}(\mathcal{D}_t^f) + s_t \times \mathcal{L}_{G^f}(\mathcal{D}_{t-1}^{f\text{-gen}}). \tag{5}$$

As s_t increases, the domain divergence between current-task forgery artifacts and the generator's inherent characteristics becomes increasingly pronounced. In such cases, the gradient conflict introduced by incoming information aggravates, rendering the preservation of previously acquired knowledge paramount to prevent catastrophic overwriting. Accordingly, increasing the penalty weight for prior knowledge facilitates a gradual adaptation process, thereby safeguarding historical performance and guiding the model toward a more favorable local optimum. Conversely, when s_t is small, the domain discrepancy between the current task and the generator is bounded, thereby alleviating the conflicts induced by incoming knowledge. Under this condition, a larger learning weight can be assigned to the new information gradient, which facilitates faster convergence while maintaining stability.

Final Summarized Samples for Detector Training. During incremental learning of deepfake detectors, $\mathbf{G}$ will provide the generative replays $\mathbf{R}_t^{\text{gen}} = \{\mathcal{D}_t^{f\text{-gen}}, \mathcal{D}_t^{r\text{-gen}}\}$, which will be combined with the next $(t+1)$-th data for training. During $(t+1)$-th task, the training mini-batch is the combination of generative real/fake replay and $(t+1)$-th data, which can be written as $\mathbf{X}_{t+1} = \{\mathbf{X}_{t+1}^r, \mathbf{X}_{t+1}^f, \mathbf{X}_g^r, \mathbf{X}_g^f\}$. Here, each $\mathbf{X}_{(\cdot)}^{(\cdot)}$ is a group of corresponding $\mathbf{x}_{(\cdot)}^{(\cdot)}$, while $\{\mathbf{X}_{t+1}^r, \mathbf{X}_{t+1}^f\} \subseteq \mathcal{D}_{t+1}$ and $\{\mathbf{X}_g^r, \mathbf{X}_g^f\} \subseteq \mathbf{R}_t^{\text{gen}}$.

3.4 IFFD Training with Generative Replay

Relative Separation Loss. As previously discussed, domain confusion among generative real replays, actual real samples, and actual fake samples may mislead the detector when direct label-based supervision is conducted. However, the similar learning effectiveness between domain-safe&real and domain-risky&real suggests that the domain-risky samples also contain previous information that could be beneficial to mitigating the catastrophic forgetting. Therefore, we propose the Relative Separation Loss (RS Loss) to leverage the valuable previous information. Instead of explicit real-fake supervisions, RS Loss considers from an indirect relative perspective. Specifically, despite the domain-risky $\mathbf{x}_g^r$ may exhibit similarity with current fake samples, its relative relation with $\mathbf{x}_g^f$ should be consistent. This is because both $\mathbf{x}_g^r$ and $\mathbf{x}_g^f$ are generated by the same real domain and the same generative character. As a result, the distinction between them is effectively nullified and filtered, leaving only the t-th forgery character. Consequently, the distribution of $\mathbf{x}_g^f$ and $\mathbf{x}_g^r$ could be encouraged to separate if they correspond to domain-risky samples. Specifically, given a minibatch $\mathbf{X}$, we first calculate the feature centroid of the generated real samples $\mathbf{X}_g^r$ based on Eq. 1, denoted as $Cent(\mathbf{X}_g^r)$. Then, we define RS Loss to maximize the separation between this generated real centroid and each sample-wise generated fake feature $\mathbf{x} \in \mathbf{X}_g^f$.

Table 1: Quantitative comparison (AUC) on Protocol 1 (Mixed-Era Forgery Incremental). L-Bound denotes vanilla incremental learning. Pre Avg. denotes the average performance over previous tasks. AF ↓ denotes Average Forgetting.

Method	Venue	Task	LDM	DFDCP	SDv21	DDPM	DiT	CDF	Pre Avg.	AF ↓	PD ↓
L-Bound	–	T1	99.99	-	-	-	-	-	-	-	-
		T2	80.75	88.76	-	-	-	-	80.75	19.24	15.23
		T3	77.46	52.63	**99.99**	-	-	-	65.04	29.33	23.30
		T4	35.18	55.52	31.75	99.93	-	-	40.81	55.43	44.39
		T5	51.44	50.77	70.63	77.10	99.21	-	62.48	34.68	30.16
		T6	52.73	80.54	81.78	50.34	53.30	**99.56**	63.73	33.84	30.28
LwF	TPAMI' 17	T6	81.31	78.16	89.40	49.11	76.92	98.15	74.98	34.51	21.15
iCaRL	CVPR' 17	T6	83.72	76.31	84.54	59.75	71.91	99.20	75.25	19.45	18.60
DER	CVPR' 21	T6	77.31	80.12	90.36	82.01	83.23	99.07	82.61	30.08	15.75
CoReD	MM' 21	T6	88.58	83.79	94.97	78.79	71.32	99.17	83.49	6.06	17.52
HDP	IJCV' 24	T6	94.31	85.82	93.73	95.10	81.53	99.40	90.10	4.63	16.03
DFIL	MM' 23	T1	99.99	-	-	-	-	-	-	-	-
		T2	97.22	91.07	-	-	-	-	97.22	2.77	5.85
		T3	98.23	75.91	99.98	-	-	-	87.07	8.46	8.62
		T4	94.01	72.43	99.80	**99.98**	-	-	88.75	8.27	8.43
		T5	95.84	66.56	98.61	88.32	96.31	-	87.33	10.42	10.86
		T6	92.37	73.39	99.02	84.27	77.39	99.48	85.29	12.18	12.33
SUR-LID	CVPR' 25	T1	99.99	-	-	-	-	-	-	-	-
		T2	99.31	90.32	-	-	-	-	**99.31**	0.68	5.17
		T3	99.17	87.95	99.96	-	-	-	93.56	1.59	4.30
		T4	**99.50**	80.27	99.54	99.73	-	-	93.10	3.65	5.23
		T5	98.70	75.11	**99.62**	**99.89**	**99.55**	-	93.33	4.17	5.41
		T6	97.40	82.18	**99.03**	**98.35**	92.50	99.22	93.89	4.02	5.21
Dual-CARE (Ours)	–	T1	99.99	-	-	-	-	-	-	-	-
		T2	**99.67**	**91.50**	-	-	-	-	**99.67**	**0.32**	**4.41**
		T3	**99.60**	**91.78**	99.97	-	-	-	**95.69**	**0.05**	**2.87**
		T4	93.63	**90.91**	99.25	**99.98**	-	-	**94.60**	**2.56**	**4.05**
		T5	**99.08**	**86.96**	**99.62**	96.45	95.59	-	**95.53**	**2.33**	**4.45**
		T6	**98.45**	**87.36**	97.71	95.28	**98.07**	98.93	**95.37**	**2.03**	**4.03**

This is achieved by minimizing their average cosine similarity, which can be written as:

$$\mathcal{L}_{\text{rs}} = \frac{1}{|\mathbf{X}_g^f|} \sum_{\mathbf{x} \in \mathbf{X}_g^f} \frac{f(\mathbf{x}) \cdot Cent(\mathbf{X}_g^r)}{\|f(\mathbf{x})\|_2 \cdot \|Cent(\mathbf{X}_g^r)\|_2}. \tag{6}$$

RS Loss can leverage the previous information in the replay real sample by sample. By considering the relative authenticity differences among generated images, we indirectly exploit the information embedded within the generative replay, even when they are domain risky.

Overall Detection Loss. When domain confusion is relatively mild, direct supervision evidently provides strong guidance. Meanwhile, RS Loss facilitates maximal exploitation of the informative content within domain-risky data. Therefore, we propose to adaptively adjust the balance between RS Loss and direct supervision. Hence, by incorporating RS Loss and DC Score, we formulate a unified loss function that simultaneously optimizes both domain-risky and domain-safe samples for generative replay-based deepfake detection. Firstly, the common supervision signal for face forgery detection is a Cross-Entropy loss, which could be written as:

$$\mathcal{L}_{\text{ce}}(\mathbf{x}) = -\left(\hat{y}\log(y_p) + (1-\hat{y})\log(1-y_p)\right), \tag{7}$$

where $\hat{y}$ is the corresponding ground-truth label, y_p is the predicted result from the backbone. Since $\forall\, \mathbf{x} \in \left\{\mathbf{X}_{t+1}^r,\ \mathbf{X}_{t+1}^f,\ \mathbf{X}_g^f\right\}$ $\mathbf{x}$ is *confusion-free* based on the discussion in Sec. 2, we can directly constrained them by $\mathcal{L}_{cf} = \mathcal{L}_{ce}(\mathbf{x})$. Then, we separate $\tilde{\mathbf{x}} \in \mathbf{X}_g^r$ from $\mathbf{X}$ that could *confuse* to adaptively adjust the direct and relative constraints based on DC Score, which can be written as:

$$\mathcal{L}_c = s_{t+1}\mathcal{L}_{ce}(\tilde{\mathbf{x}}) + (1 - s_{t+1})\mathcal{L}_{rs}, \tag{8}$$

which means the safe direct supervision could be conducted with higher weights if the domain distance is relatively large, otherwise a higher L_{rs} should be applied.

Therefore, the overall loss can be written as:

$$\mathcal{L}_{overall} = \mathcal{L}_c + \mathcal{L}_{cf}, \tag{9}$$

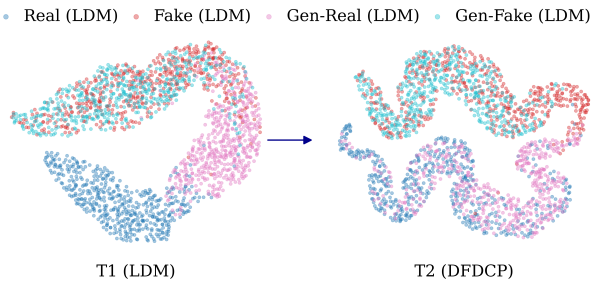


Figure 4: UMAP visualizations. (Left) Initial state: Gen-Real (pink) is ambiguously positioned near Fake (red / cyan). (Right) After learning T2: Our method ($\mathcal{L}_{rs}$) actively separates the Gen-Real cluster, resolving the domain confusion.

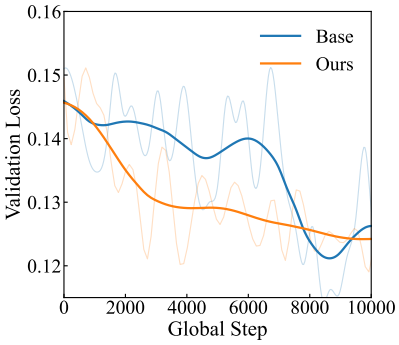


Figure 5: Learning performance with (Ours) or without (Base) Domain-aware Fast Training.

Table 2: Comparison of various generative replay strategies. We evaluate Full Replay, Fake-Only Replay, our dynamic Dual-CARE, and two baselines with fixed confusion score.

Method	Task-Incremental Performance (AUC)								Average Accuracy (ACC)			
	LDM	DFDCP	SDv21	DDPM	DiT	CDF	Avg.	PD ↓	Real	PD ↓	Fake	PD ↓
Lower Bound	52.73	80.54	81.78	50.34	53.30	99.56	69.71	30.28	89.51	10.34	41.28	58.62
Full Replay	96.59	84.61	94.96	98.08	75.67	99.37	91.55	8.44	84.55	15.15	83.96	15.89
Fake-Only Replay	95.39	76.24	98.09	49.94	52.10	99.63	78.56	21.43	82.77	16.98	63.10	36.80
Fixed $s_t = 0.5$	99.77	87.57	97.25	99.40	78.66	99.70	93.72	6.27	94.44	5.36	62.36	37.49
Fixed $s_t = 0.1$	99.59	88.71	96.69	98.86	75.19	99.69	93.12	6.87	**95.38**	**4.52**	61.39	38.51
Ours (Adaptive)	99.69	89.07	98.39	99.47	88.29	99.51	**95.74**	**4.25**	89.60	10.25	**85.29**	**14.56**

which encourages the detector to leverage information from current and replay samples simultaneously.

4 Experimental Results

4.1 Experimental Settings

Datasets. To construct a comprehensive and challenging benchmark for incremental face forgery detection, our experiments utilize a curated selection of datasets, which is designed to simulate a realistic scenario by spanning both classical, widely-used forgery datasets and the latest cutting-edge forgeries generated by advanced diffusion models. The classical datasets include: Celeb-DF-v2 (CDF) (Li et al., 2020b), DeepFake Detection Challenge Preview (DFDCP) (Dolhansky et al., 2020), and the hybrid-category FaceForensics++ (FF++) (Rossler et al., 2019). To address the most recent threats, we further incorporate a suite of modern forgeries, that is, {SDv21 (Rombach et al., 2022b), DiT (Peebles & Xie, 2023)} from DF40 (Yan et al., 2024b) and {LDM (Rombach et al., 2022a), DDPM (Ho et al., 2020)} from DiffusionFace (Chen et al., 2024). This blend of classical and cutting-edge forgery types creates a comprehensive evaluation against evolving threats.
Incremental Protocols. We adopt two complementary incremental protocols. **Protocol 1 (P1)** follows the sequence {LDM, DFDCP, SDv21, DDPM, DiT, CDF}, covering diverse forgery types across different generative paradigms. **Protocol 2 (P2)** follows {DDPM, FF++, DFDCP, CDF}, adapted from the benchmark protocol in SUR-LID for comparison with existing methods. Detailed motivations&protocols can be found in Appendix Sec. B.3.
Implementation Details. We adopt EfficientNetB4 (Tan & Le, 2019) as the backbone and train it using Adam (Kingma & Ba, 2014) for 5 epochs at a learning rate of 0.0002. All methods use 500 replay samples per task. We use frame-level AUC (Yan et al., 2023b) as the primary metric, supplemented by accuracy (ACC). Catastrophic forgetting is measured by $PD = M_0 - M_N$, where M_0 and M_N denote the average metrics in the base and final sessions, respectively. Detailed implementation settings can be found in Appendix Sec. B.

4.2 Effectiveness Comparisons with Baselines

As shown in Tab. 1, our method consistently outperforms general continual learning (LwF (Li & Hoiem, 2017), iCaRL (Rebuffi et al., 2017), DER (Yan et al., 2021)) and state-of-the-art IFFD-specific baselines (CoReD (Kim et al., 2021), HDP (Sun et al., 2025), DFIL (Pan et al., 2023), SUR-LID (Cheng et al., 2025a)) under Protocol 1, achieving the highest average AUC with **no** data replay. This advantage stems from two aspects: (1) generative replay approximates historical distributions to yield diverse rehearsal samples without storing real images; (2) confusion-aware regularization distinguishes reliable from ambiguous signals, facilitating knowledge transfer while minimizing negative interference to enhance stability. Protocol 2 comparisons are provided in the *Appendix* Tab. 6.

4.3 Practical Cost of Generative Replay

We analyze the practical cost of GenReplay from three perspectives, as reported in Tab. 3. **1) Generator training.** Generative replay introduces additional generator training cost at each incremental task. However, this cost is incurred only during training and does not affect the detector's inference efficiency during deployment. Therefore, it represents a reasonable trade-off for achieving privacy-preserving replay and scalable replay diversity.

2) IFFD replay overhead. During training, the additional computational overhead of the proposed Dual-CARE mainly arises from replay-sample generation, which only involves generator inference and is practically minor.

Table 3: Practical Cost Comparisons. n denotes replay size.

Method	GenTrain (h)	DetTrain (h)	Storage (MB)	Inference (s)
Common	–	1.31	$0.3n$	8.03×10^{-4}
Ours (Gen)	1.06	$1.31+1.2\times10^{-5}n$	38.2	8.03×10^{-4}

3) Storage cost. Rather than storing an increasing number of raw replay samples, Dual-CARE maintains only a lightweight LoRA module, which can generate arbitrarily many replay samples without increasing the fixed storage burden. In contrast, sample-based replay requires storage to grow with the number of tasks.

4.4 Ablation Study

To evaluate the detector-side solution to domain confusion, we ablate its components with the generator strategy disabled; thus, "All detector components" denotes the complete detector objective under a fixed generator. As shown in Tab. 4, removing either Gen-Real supervision or $\mathcal{L}_{rs}$ causes substantial performance degradation, demonstrating their complementary roles in suppressing generative artifacts and preserving decision boundaries. All distance combinations outperform the variant without $\mathcal{L}_{rs}$, with DCS-L2 + RS-Cos performing best, suggesting that L2 better captures domain confusion for DCS while cosine similarity better enforces feature separation. Fig. 5 further shows that our domain-aware strategy accelerates convergence when s_t is small, consistent with our design. Additional component-wise ablations, backbone choices, adaptive-weight normalization functions, and variants of $\mathcal{L}_{rs}$ are provided in the *Appendix*.

4.5 Analysis of Generative Replay

Advantage on Replay Diversity. Unlike common replay that requires proportional storage overhead to scale, generative replay inherently yields infinite images via a single generator. In Fig. 6, expanding the replay size consistently boosts the performance of both DFIL and our method, validating the critical impact of replay diversity. In *Appendix*, at an equivalent replay size, our framework demonstrates superior generative diversity across TCE, CLIP-distance, and LPIPS metrics.

Impact of Generator Quality. We analyze replay generator quality by comparing GAN, DDIM, and LDM under the P2 protocol (Fig. 7). Results show a direct correlation between generation fidelity and detection accuracy. The outdated GAN architecture performs poorly, whereas diffusion-based LDM and DDIM generate promising replays that significantly enhance incremental learning, leading us to select LDM. Furthermore, the similar performance of LDM and DDIM proves that Dual-CARE is generator-agnostic (given suf-

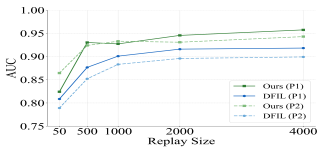


Figure 6: Replay-size comparison under two protocols.

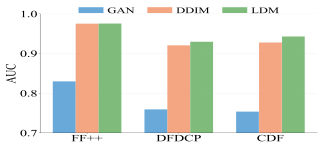


Figure 7: Results of replay generators across datasets.

Table 4: Ablation study on core components and distance metrics. All results are AUCs obtained by the final Protocol 1 model at T6 and evaluated on tasks T1–T5, and we isolated each component by disabling the generator-side strategy. Δ_{Avg} denotes the average performance difference relative to Ours. Within each group, the best and second-best results are highlighted in **bold** and underline.

Variant	LDM	DFDCP	SDv21	DDPM	DiT	Avg.	Δ_{Avg}
Core component ablation							
w/o DC-guided Gen	95.42	83.61	94.16	96.38	**94.99**	92.91	−2.07
w/o Gen-Real Sup.	88.41	79.17	98.52	52.04	53.47	74.32	−20.66
w/o $\mathcal{L}_{rs}$	92.75	**89.29**	**99.34**	90.47	83.04	90.97	−4.01
***w* Det. All**	**99.69**	89.07	98.39	**99.47**	88.29	**94.98**	–
Distance metric ablation							
DCS-Cos + RS-Cos	97.96	**92.59**	97.52	98.20	74.02	92.05	−2.93
DCS-L2 + RS-L2	**99.77**	85.62	96.09	97.12	79.10	91.54	−3.44
DCS-Cos + RS-L2	99.54	86.37	97.31	99.28	82.05	92.91	−2.07
DCS-L2 + RS-Cos	99.69	89.07	**98.39**	**99.47**	**88.29**	**94.98**	–

ficient generation capability), highlighting its scalability and extensive application potential.

4.6 Analysis of Domain Confusion Effect

To analyze domain confusion, we visualize the T1 LDM feature space using UMAP (McInnes et al., 2018). As illustrated in Fig. 4, initial Gen-Real samples reside in an ambiguous region near Fake clusters, posing a high risk of confusion. Conversely, after incrementally learning T2 DFDCP, our $\mathcal{L}_{rs}$-driven adaptive mechanism effectively resolves this ambiguity, pushing Gen-Real features away from Fake clusters and aligning them with the Real cluster. This confirms that our method rectifies distributional misalignment and restores clear feature separation. Additional cross-dataset visualizations are provided in the *Appendix*.

4.7 Privacy Quantitative Analysis

Traditional sample replay poses privacy risks by storing raw facial images. Generative replay avoids this but may risk identity memorization. Hence, we measure feature similarity between synthesized replay samples and original training faces using ArcFace (Deng et al., 2018) and CosFace (Wang et al., 2018). In Fig. 8, similarities are near zero with minimal variance. Despite a high generation fidelity (FID 15.36), the generator produces entirely novel identities by capturing the structural distribution rather than memorizing training samples. This confirms that our generative replay framework effectively mitigates identity leakage, providing a secure rehearsal mechanism for IFFD. Furthermore, in Appendix Tab. 10, we demonstrate that the generative ID cannot be effectively recalled by the training samples, and GenReplay may not be distinguished from Actual Replay.

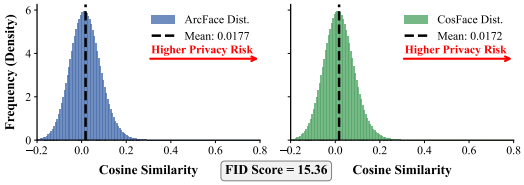


Figure 8: Quantitative analysis of the privacy protection effectiveness of our method.

5 Conclusion

In this paper, we present Dual Confusion-Aware REgularization (Dual-CARE), a generative replay framework for incremental face forgery detection. Dual-CARE characterizes domain confusion between replay generators and emerging forgery models via a Domain Confusion Score, and accordingly modulates both generator updates and detector supervision, with Relative Separation Loss applied to domain-risky samples. Extensive experiments demonstrate improved robustness and accuracy across diverse replay settings while effectively mitigating the adverse effects of domain overlap.

AI Use Disclosure

In this work, we used generative AI tools to assist with translation. We did not use generative AI tools for other tasks requiring mandatory disclosure, including developing theoretical or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing research methodology or experiments, implementing the proposed method, processing datasets, or interpreting experimental results.

Additionally, we used generative AI tools for several tasks for which disclosure is recommended, including drafting and editing parts of the manuscript, improving readability, assisting with manuscript structure, formatting LaTeX and references, and refining the presentation of tables and figures. All AI-assisted content was reviewed and revised by the authors. We take responsibility for the final content of this work, including text or artifacts produced with the aid of generative AI.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *European Conference on Computer Vision*, pp. 139–154, 2018.

Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4113–4122, 2022.

Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 18710–18719, 2022.

Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 1081–1088, 2021.

Zhongxi Chen, Ke Sun, Ziyin Zhou, Xianming Lin, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. Diffusionface: Towards a comprehensive dataset for diffusion-based face forgery analysis. *arXiv preprint arXiv:2403.18471*, 2024.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? *Advances in Neural Information Processing Systems*, 37:21979–21998, 2024.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Li Hao, Jiaxin Ai, Qin Zou, Chen Li, and Zhongyuan Wang. Stacking brick by brick: Aligned feature isolation for incremental face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13927–13936, 2025a.

Jikang Cheng, Ying Zhang, Qin Zou, Zhiyuan Yan, Chao Liang, Zhongyuan Wang, and Chen Li. Ed ˆ4: Explicit data-level debiasing for deepfake detection. *IEEE Transactions on Image Processing*, 2025b.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition Conference*, pp. 19207–19217, 2025.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2021.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4685–4694, 2018.

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.

Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3994–4004, 2023.

Rui Gao and Weiwei Liu. Ddgr: Continual learning with deep diffusion-based generative replay. In *International Conference on Machine Learning*, pp. 10744–10763, 2023.

Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. Learning meta face recognition in unseen domains. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6163–6172, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499, 2023.

Hossein Kashiani, Niloufar Alipour Talemi, and Fatemeh Afghah. Freqdebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8775–8785. IEEE, 2025.

Junsu Kim, Hoseong Cho, Jihyeon Kim, Yihalem Yimolal Tiruneh, and Seungryul Baek. Sddgr: Stable diffusion-based deep generative replay for class incremental object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 28772–28781, 2024.

Minha Kim, Shahroz Tariq, and Simon S Woo. Cored: Generalizing fake media detection with continual representation using distillation. In *ACM International Conference on Multimedia*, pp. 337–346, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5001–5010, 2020a.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020b.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

Jiahao Liang, Huafeng Shi, and Weihong Deng. Exploring disentangled content information for face forgery detection. In *European Conference on Computer Vision*, pp. 128–145. Springer, 2022.

Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3589–3599, 2021.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Kun Pan, Yifang Yin, Yao Wei, Feng Lin, Zhongjie Ba, Zhenguang Liu, Zhibo Wang, Lorenzo Cavallaro, and Kui Ren. Dfil: Deepfake incremental learning by exploiting domain-invariant forgery clues. In *ACM International Conference on Multimedia*, pp. 8035–8046, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE International Conference on Computer Vision*, pp. 4195–4205, 2023.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pp. 86–103. Springer, 2020.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022a.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022b.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE International Conference on Computer Vision*, pp. 1–11, 2019.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in Neural Information Processing Systems*, 30, 2017.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 18720–18729, 2022.

James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, and Zsolt Kira. Always be dreaming: A new approach for data-free class-incremental learning. In *IEEE International Conference on Computer Vision*, pp. 9374–9384, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 2316–2324, 2022.

Ke Sun, Shen Chen, Taiping Yao, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Continual face forgery detection via historical distribution preserving. *International Journal of Computer Vision*, 133(3):1067–1084, 2025.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114, 2019.

Jiahe Tian, Cai Yu, Xi Wang, Peng Chen, Zihao Xiao, Jizhong Han, and Yesheng Chai. Dynamic mixed-prototype model for incremental deepfake detection. In *ACM International Conference on Multimedia*, pp. 8129–8138, 2024.

Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14923–14932, 2021.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3014–3023, 2021.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *IEEE International Conference on Computer Vision*, pp. 22412–22423, 2023a.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv preprint arXiv:2307.01426*, 2023b.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8984–8994, 2024a.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Li Yuan, Chengjie Wang, Shouhong Ding, et al. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024b.

Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. In *International Conference on Machine Learning*, 2025.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2185–2194, 2021.

Appendix

A Related Works

A.1 Face Forgery Detection

Current face forgery detection methods typically leverage available forgery samples to train a generalized model capable of handling unseen forgeries. Various forgery-specific patterns, such as noise Li et al. (2020a), local region Chen et al. (2021); Zhao et al. (2021), and frequency information Qian et al. (2020); Guo et al. (2020); Kashiani et al. (2025), are explored to capture more discriminative forgery cues. To alleviate the performance degradation observed in cross-domain evaluations, researchers propose a range of learning strategies from different perspectives, including contrastive learning Sun et al. (2022), identity information modeling Huang et al. (2023); Dong et al. (2023), disentangled representation learning Liang et al. (2022); Yan et al. (2023a), reconstruction-based learning Cao et al. (2022); Wang & Deng (2021), and data augmentation Chen et al. (2022); Shiohara & Yamasaki (2022); Yan et al. (2024a). Recently, several ViT-based methods such as CLIP Cui et al. (2025) and LoRA-based Effort Yan et al. (2025) are proposed to enhance the generalization capability of forgery detection by leveraging large vision-language models. In summary, many general approaches have been proposed to learn transferable forgery features from limited known data. These methods aim to maintain good performance on unseen samples. However, given the large scale and diversity of existing forgery data, relying on a few known datasets to train a truly universal detector is unrealistic.

A.2 Incremental Learning for Forgery Detection

Incremental learning has been extensively studied across various domains and is typically categorized into parameter isolation De Lange et al. (2021), parameter regularization Aljundi et al. (2018); Kirkpatrick et al. (2017); Li & Hoiem (2017), and data replay Mai et al. (2021); Rebuffi et al. (2017). In the field of face forgery detection, most incremental methods are based on sample replay, where representative samples from previous tasks are stored or reused to mitigate catastrophic forgetting. Representative replay-based approaches in incremental face forgery detection have adopted different strategies to preserve prior knowledge. For example, CoReD Kim et al. (2021) relies on distillation loss to maintain knowledge from previous tasks. Meanwhile, DFIL Pan et al. (2023) improves replay effectiveness by emphasizing both center and hard samples. In addition, HDP Sun et al. (2025) employs refined universal adversarial perturbations as a replay mechanism. Similarly, DMP Tian et al. (2024) constructs mixed prototypes to summarize earlier task distributions. More recently, SUR-LID Cheng et al. (2025a) introduces sparse uniform replay combined with a latent-space incremental detector to better preserve previous knowledge. Although effective in retaining previous knowledge, all these methods depend on explicit access to stored samples, which may raise privacy concerns and limit scalability.

Beyond sample-based replay, generative replay offers a promising alternative, reconstructing past distributions through generation rather than storage. This idea has been widely explored in general continual learning, where generative models synthesize pseudo-data to approximate previous task distributions and thus preserve earlier knowledge without explicit memory buffers. The seminal work Deep Generative Replay Shin et al. (2017) introduces a dual-model framework that reconstructs past data distributions without storing real samples. Building on this idea, data-free class-incremental learning methods Smith et al. (2021) synthesize pseudo-samples through model inversion to enable continual learning under memory and privacy constraints. More recently, diffusion-based generative replay approaches such as DDGR Gao & Liu (2023) and SDDGR Kim et al. (2024) have improved the stability and diversity of generated data, achieving stronger knowledge retention in incremental classification and detection tasks. These advances further demonstrate the substantial and growing potential of generative replay in continual visual learning, though its application to forgery detection remains largely unexplored.

Table 5: Hyper-parameters and implementation details of Dual-CARE.

Name	Description	Value
Backbone	Detector architecture	EfficientNet-B4
Resolution	Input image resolution	256×256
Batch size	Training batch size	32
Optimizer	Optimization algorithm	Adam
Learning rate	Initial learning rate	2×10^{-4}
Weight decay	L2 regularization coefficient	5×10^{-4}
Epochs	Training epochs per incremental task	5
Loss	Classification objective	Cross Entropy
GPU	Training device	1× NVIDIA Tesla A100
Train frames	Sampled frames per video during training	8
Test frames	Sampled frames per video during evaluation	32
Augmentation	Image augmentation strategy	Flip, Rotation, Blur, JPEG Compression
Flip probability	Probability of horizontal flipping	0.5
Rotation range	Rotation angle range	$[-10, 10]^\circ$
JPEG quality	JPEG compression quality range	[40,100]
Generator	Replay generator architecture	Latent Diffusion Model (LDM)
Generator number	Separate generators for real/fake domains	2
Replay strategy	Historical knowledge preservation	Diffusion-based generative replay
Adaptation method	Efficient generator update	LoRA
LoRA rank	Low-rank adaptation dimension	16
LoRA scaling	LoRA scaling factor	16
Replay samples	Generated replay samples per class	500
Diffusion steps	Number of diffusion timesteps	1000
DDIM steps	Sampling steps	250
DDIM η	Stochasticity coefficient	1.0
Latent resolution	Diffusion latent space size	64×64
Batch size	Sampling batch size	10

B Detailed Implementation Settings

First, we provide our detailed hyperparameter settings in Tab. 5. All baselines are reproduced within DeepFakeBench (Yan et al., 2023b) under consistent experimental settings to ensure fair comparison.

B.1 Preprocessing

Data Preprocessing. Following the standard DeepFakeBench (Yan et al., 2023b) protocol, all video frames undergo face detection, extraction, and alignment before being resized to 256×256. For input normalization, we adopt a mean of $[0.5, 0.5, 0.5]$ and a standard deviation of $[0.5, 0.5, 0.5]$ for the three RGB channels. During training, 8 frames are uniformly sampled from each video, while 32 frames are sampled during testing to ensure more stable and reliable performance evaluation.

Data Augmentation. To enhance the generalization ability of the detector, we employ a comprehensive data augmentation pipeline on the current task data using the `Albumentations (Buslaev et al., 2020)` library. The augmentation operations and their corresponding application probabilities are detailed as follows:

- **Spatial Transformations:** Horizontal Flip ($p = 0.5$), Rotation within $\pm 10^\circ$ ($p = 0.5$), and isotropic resizing.
- **Pixel-level Transformations:** Gaussian Blur with a kernel size in the range of $[3, 7]$ ($p = 0.5$).
- **Compression Artifacts:** JPEG compression with a quality range of 40–100, applied with $p = 0.5$.
- **Color Perturbations:** One of Random Brightness/Contrast (limit 0.1), FancyPCA, or HueSaturationValue, selected and applied with probability $p = 0.5$.

It is worth noting that while the current task data are augmented as described above, the generated replay samples are kept with standard normalization only so as to preserve their intrinsic generative distribution characteristics.

B.2 Hyperparameters for Generative Replay

In our generative replay setup, a replay buffer is constructed for each training batch. Specifically, for a current-task batch of size $B_{\text{new}} = 32$, we generate and replay $B_{\text{fake}} = 12$ fake samples and $B_{\text{real}} = 12$ real samples to maintain class balance during the incremental training process.

B.3 Detailed Incremental Protocols.

To comprehensively evaluate model robustness in evolving forgery landscapes, we propose two complementary incremental protocols that capture real-world dynamics and benchmark-level comparability.

- **Protocol 1 (P1): Mixed-Era Forgery Incremental.** It follows the sequence {LDM, DFDCP, SDv21, DDPM, DiT, CDF}, simulating a realistic, temporally chaotic evolution of forgery techniques. It intentionally interleaves classical Face-Swapping (FS) datasets (DFDCP, CDF) with modern Entire Face Synthesis (EFS) forgeries to emulate the heterogeneous and non-sequential emergence of threats in the wild. This setup is designed to rigorously evaluate the resilience of a model to catastrophic forgetting and its adaptability to domain confusion.
- **Protocol 2 (P2): Benchmark-Aligned Incremental.** It employs the sequence {DDPM, FF++, DFDCP, CDF} and extends the benchmark protocol introduced in recent work SUR-LID. To align with prior baselines while avoiding redundant configurations, we replace the initial dataset (SDv21) with DDPM, a diffusion-based forgery type. It ensures consistency with established benchmarks while incorporating emerging generative paradigms for a fair yet forward-looking evaluation.

B.4 Generative Model Details

We adopt the Latent Diffusion Model (LDM (Rombach et al., 2022a)) as our primary replay generator, leveraging its compressed latent space to achieve both computational efficiency and high-fidelity synthesis.

Autoencoder Configuration: We employ a VQ-regularized autoencoder (Rombach et al., 2022a) with a downsampling factor of $f = 4$ (referred to as `vq-f4`). The encoder compresses each $256 \times 256 \times 3$ input image into a latent feature map of size $64 \times 64 \times 3$. The autoencoder uses 128 base channels with channel multipliers of $[1, 2, 4]$ and includes two residual blocks at each resolution. The codebook contains $n_{embed} = 8192$ entries, each with an embedding dimension of 3.

Diffusion and Network Configuration: The diffusion process is modeled in the latent space using a UNet (Ronneberger et al., 2015) architecture with the following specifications:

- **Backbone:** A time-conditional UNet with 224 base channels is used as the denoising backbone, comprising two residual blocks per level and channel multipliers of $[1, 2, 3, 4]$.
- **Attention:** Spatial attention is incorporated at resolutions corresponding to downsampling factors of $[8, 4, 2]$ (i.e., 8, 16, and 32), with each attention head configured with 32 channels.
- **Noise Schedule:** A linear noise schedule is used in the forward diffusion process, ranging from $\beta_{start} = 0.0015$ to $\beta_{end} = 0.0195$, and the model is optimized over $T = 1000$ timesteps with an L_2 reconstruction objective.
- **Optimization:** Training is performed using a base learning rate of 2.0×10^{-6} to ensure stable convergence.

Sampling Configuration. During the generative replay phase, we employ the Denoising Diffusion Implicit Model (DDIM (Song et al., 2020)) sampler to accelerate synthesis. The sampling process is configured with the following parameters:

- **Sampling Steps:** We perform inference with 250 steps. Although the model is trained over 1000 timesteps, a strided sampling schedule is used to substantially speed up generation while preserving high fidelity.
- **Stochasticity (η):** We use the DDIM sampler with the stochasticity parameter set to $\eta = 1.0$. This configuration effectively implements strided DDPM (Ho et al., 2020) sampling, ensuring that the generative replay preserves diversity comparable to the original training distribution.

C Further Analysis and Ablations

C.1 Results on Protocol 2

The main paper details our experiments on the more realistic Protocol 1. To ensure fair alignment with established baselines, we further evaluate our method on Protocol 2. The quantitative results for this benchmark-aligned setup are presented in Tab. 6.

Table 6: Quantitative comparison (AUC) on Protocol 2. L-Bound denotes vanilla incremental learning without any strategy.

Method	Task	DDPM	FF++	DFDCP	CDF	Avg.
L-Bound	T1	99.99	-	-	-	99.99
	T2	69.60	95.16	-	-	82.38
	T3	60.58	71.28	92.38	-	74.75
	T4	53.39	65.92	82.23	99.77	75.33
LwF	T1	99.99	-	-	-	99.99
	T2	81.01	85.02	-	-	83.02
	T3	76.12	65.07	93.91	-	78.37
	T4	61.71	61.14	81.35	98.18	75.60
iCaRL	T1	99.99	-	-	-	99.99
	T2	96.56	91.02	-	-	93.79
	T3	91.71	80.07	90.95	-	87.58
	T4	91.79	73.93	87.10	99.09	87.98
DFIL	T1	99.99	-	-	-	99.99
	T2	97.17	93.31	-	-	95.24
	T3	93.51	71.27	91.64	-	85.47
	T4	92.33	68.10	80.74	99.63	85.20
SUR-LID	T1	99.99	-	-	-	99.99
	T2	99.35	92.36	-	-	95.85
	T3	99.13	81.51	92.13	-	90.92
	T4	98.16	78.16	86.37	99.18	90.47
Dual-CARE (Ours)	T1	99.99	-	-	-	99.99
	T2	99.88	94.51	-	-	**97.20**
	T3	99.60	89.28	91.97	-	**93.62**
	T4	98.87	85.90	92.04	98.12	**93.73**

C.2 Comprehensive Ablation on Core Components

In the main paper, we evaluate the "Dual" components separately to investigate their isolated performance. Here, we aim to investigate their mutual influence and the performance as a whole. To systematically evaluate the contribution and synergy of our mechanisms, we extend the leave-one-out analysis from the main text to a comprehensive permutation of all three core components: Domain-aware Fast Training (DFT), Direct Gen-Real Supervision (DGS), and Relative Separation Loss (RSL). As shown in Tab. 7, the results reveal

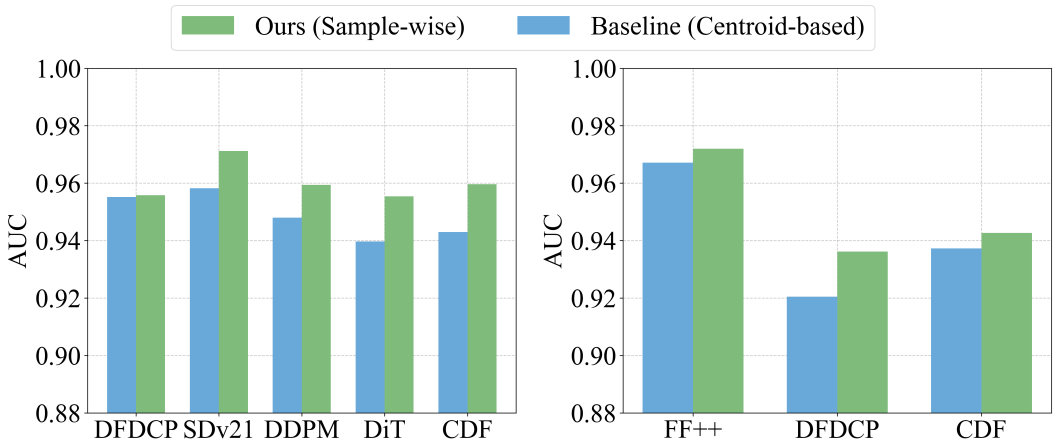


Figure 9: Ablation on Relative Separation Loss. The sample-wise formulation consistently outperforms the centroid-based variant across both Protocol 1 (Left) and Protocol 2 (Right) by preserving fine-grained structural diversity.

a clear hierarchical dependency. Explicit label guidance from Gen-Real Sup. serves as the foundational anchor; omitting it leads to a severe performance collapse, as the detector fundamentally fails to retain historical real concepts. Building upon this established baseline boundary, the introduction of $\mathcal{L}_{rs}$ consistently provides a substantial performance lift by adaptively resolving the domain confusion between generated real and fake samples. Furthermore, the integration of DC-guided Gen ensures that the generative replay maintains stable and high-fidelity distributions across incremental tasks. Ultimately, the synergistic combination of all three components achieves the optimal balance between prior knowledge preservation and generative artifact mitigation.

Table 7: Comprehensive ablation study on the combinations of core components. The synergistic integration of all three modules yields the optimal incremental detection performance.

DFT	**DGS**	**RSL**	LDM	DFDCP	SDv21	DDPM	DiT	Avg.
×	×	×	89.47	81.44	95.91	53.15	61.53	76.30
×	×	✓	94.11	80.63	95.40	53.21	58.62	76.39
×	✓	×	92.06	85.28	96.64	94.65	87.16	91.15
×	✓	✓	95.42	83.61	94.16	96.38	94.99	92.91
✓	×	×	94.94	78.98	96.42	50.12	58.98	75.88
✓	×	✓	92.99	78.60	97.51	56.29	55.42	76.16
✓	✓	×	**99.39**	**88.40**	96.10	**98.38**	86.05	93.66
✓	✓	✓	98.45	87.36	**97.71**	95.28	**98.07**	**95.37**

C.3 Impact of Sample-wise Constraint in $\mathcal{L}_{rs}$

To assess the importance of the fine-grained sample-wise constraint in our Relative Separation Loss ($\mathcal{L}_{rs}$), we compare it against a coarse-grained centroid-based variant. As shown in Fig. 9, the sample-wise formulation consistently outperforms the centroid-based strategy across all stages in both protocols. By enforcing separation at the instance level rather than relying on collapsed domain centers, our approach prevents local decision boundary distortions and preserves structural diversity. This fine-grained supervision proves crucial for mitigating catastrophic forgetting, particularly as domain confusion intensifies in later incremental stages.

C.4 Robustness of Domain Confusion Score Normalization

The Domain Confusion Score (s_t) is essential for dynamically balancing the supervision signals. To assess the method's sensitivity to different normalization functions, we compare our default hyperbolic tangent (`tanh`) with Sigmoid and Linear Scaling (`d/5`). As shown in Table 8, all variants produce reasonable results, but `tanh` yields the highest average AUC. This advantage stems from `tanh`'s smooth and bounded mapping of feature distances, which more effectively accommodates the varying magnitudes of domain shifts encountered throughout incremental learning.

Table 8: Performance comparison of normalization functions under Protocol 2.

Method	Task	DDPM	FF++	DFDCP	CDF	Avg.	s_t
tanh	T1	99.99	-	-	-	99.99	
	T2	99.88	94.51	-	-	97.20	0.9997
	T3	99.60	89.28	91.97	-	93.62	0.9999
	T4	98.87	85.90	92.04	98.12	93.73	0.9769
sigmoid	T1	99.99	-	-	-	99.99	-
	T2	97.99	94.97	-	-	96.48	0.9875
	T3	98.41	88.29	89.79	-	92.16	0.9971
	T4	98.83	79.24	89.23	98.88	91.54	0.9013
d/5	T1	99.99	-	-	-	99.99	-
	T2	98.55	94.97	-	-	96.76	0.8077
	T3	99.86	86.25	91.13	-	92.41	0.9999
	T4	99.91	77.91	90.41	98.47	91.67	0.4694

C.5 Performance Comparison with Xception and ResNet34 Backbones

To evaluate the generality of our framework beyond a specific feature extractor, we further test two additional backbones: Xception (Chollet, 2017) and ResNet34 (He et al., 2016). As shown in Fig. 10, we compare Dual-CARE with state-of-the-art incremental forgery detection methods, SUR-LID (Cheng et al., 2025a) and DFIL (Pan et al., 2023). Across both Protocol 1 and Protocol 2, Dual-CARE consistently outperforms all competitors, regardless of the backbone architecture. These results demonstrate that our Dual Confusion-Aware Regularization strategy is model-agnostic and remains robust across diverse network designs.

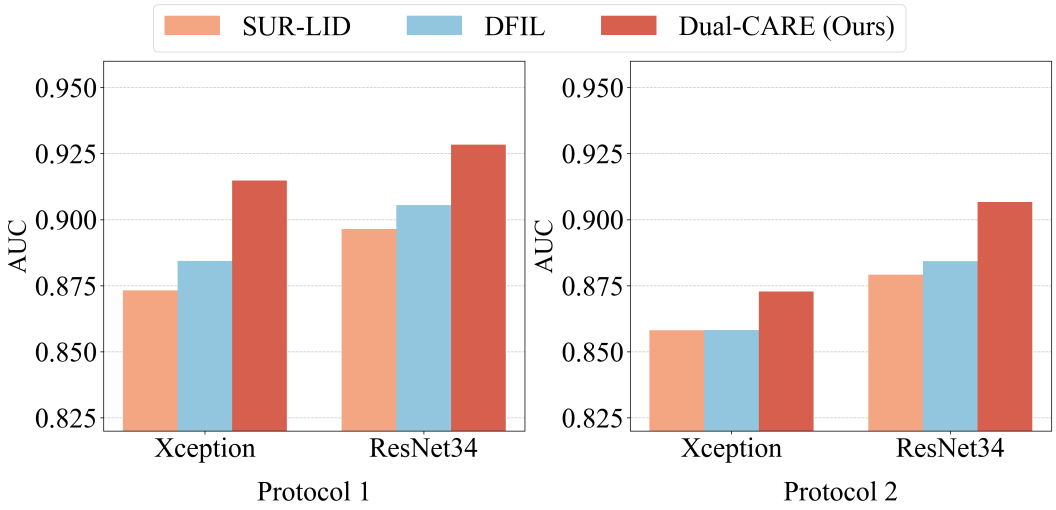


Figure 10: Performance comparison with Xception and ResNet34 backbones.

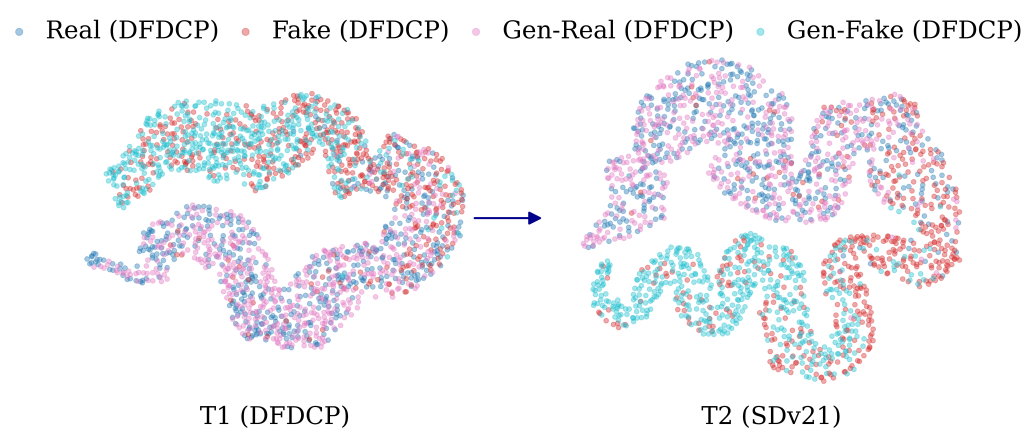


Figure 11: UMAP (McInnes et al., 2018) visualization of the Domain-Safe scenario (DFDCP).

C.6 Visualization of Domain-Safe Scenario

Complementing the domain-risky visualization in the main paper, we further illustrate a "domain-safe" scenario using the DFDCP dataset. As shown in Fig. 11, unlike the LDM case, the generated real samples for DFDCP initially align closely with the actual real samples, indicating minimal risk of domain confusion. Notably, this favorable alignment is well preserved even after learning the subsequent task (SDv21 (Rombach et al., 2022b)). This result demonstrates that our adaptive Dual-CARE strategy effectively identifies domain-safe samples and applies appropriate direct supervision to maintain their distributional integrity, rather than enforcing unnecessary separation.

D Sample Visualizations of Generative Replay

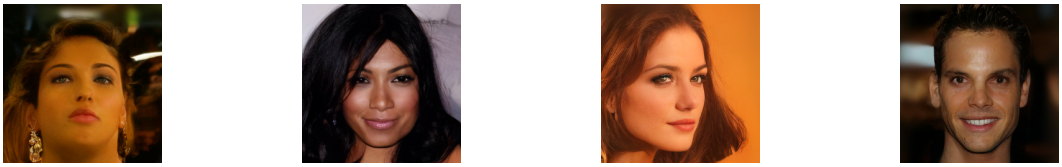
We visualize replay samples across learning stages to demonstrate the robust memory retention of our mechanism. Our approach maintains a single, continuously updated model. As shown in Fig. 12, we track the generator's outputs from initial training on DiffusionFace (DDPM) through subsequent updates on FaceForensics++ (FF++) and DFDCP.

E Additional Experimental Results

E.1 Generative Replay Size and Diversity

For fair comparison, Dual-CARE uses the same replay budget as sample-replay baselines in all main experiments, i.e., 500 replay samples per previous task. As shown in Fig. **??**, increasing the replay size consistently improves the incremental detection performance, indicating that broader coverage of previous task distributions helps alleviate catastrophic forgetting. Unlike sample replay, which requires storing more historical images as the replay size increases, generative replay can synthesize additional samples from a fixed-size generator without increasing raw-data storage. We further compare the diversity of sample replay and generative replay under the same replay budget in Tab. 9. Generative replay achieves consistently higher TCE, CLIP distance, and LPIPS scores than sample replay, demonstrating its advantage in capturing richer distributional, semantic, and perceptual variations. These results further confirm that generative replay provides broader coverage of historical task distributions than repeatedly using a fixed set of stored samples.

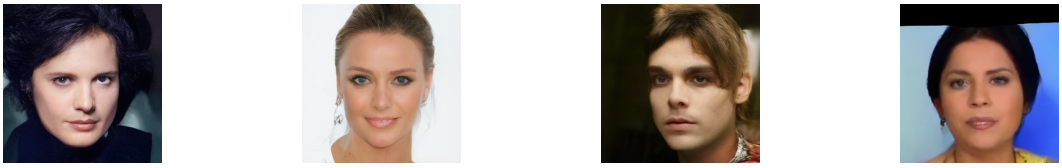
(a) Stage 1: DiffusionFace (Real)



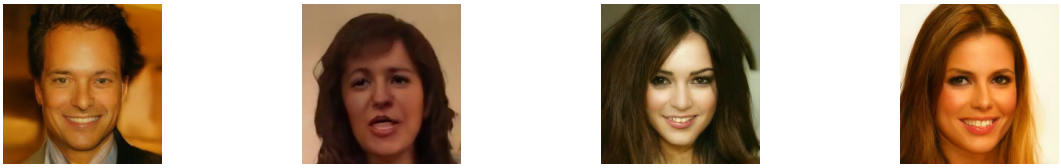
(b) Stage 1: DiffusionFace (DDPM)



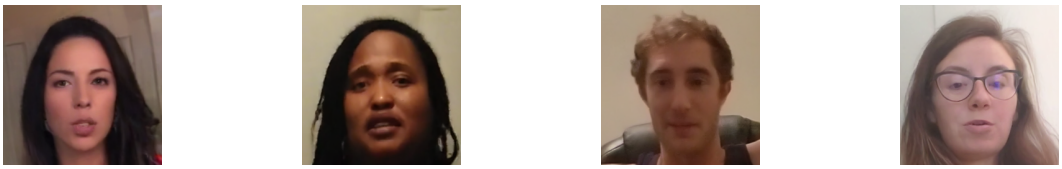
(c) Stage 2: Updated on FaceForensics++ (Real)



(d) Stage 2: Updated on FaceForensics++ (Fake)



(e) Stage 3: Updated on DFDCP (Real)



(f) Stage 3: Updated on DFDCP (Fake)

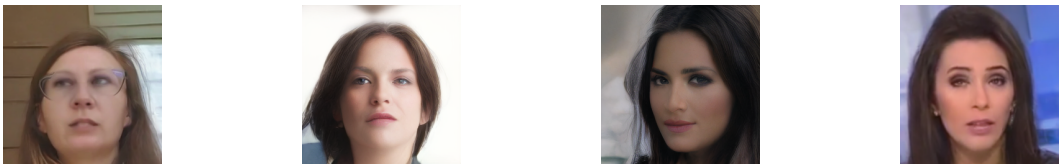


Figure 12: LDM replay samples on (a-b) DiffusionFace, (c-d) FaceForensics++, and (e-f) DFDCP. A single updated generator retains high fidelity across tasks.

E.2 DC Score and Distance Metric

DC Score is quantified in the current-detector f_t feature space because confusions between old and new tasks in IFFD are manifested in the representation geometry of f_t. Thus, our goal is not to estimate absolute distribution discrepancy, but to measure replay confusion perceived by f_t. The dynamic feature space is appropriate for IFFD task transitions, as replay confusion also changes across tasks. Compared with a fixed distance metric, centroid distance under f_t can more directly reflect the actual conflict from task t to t+1. More complex distribution metrics could be explored. However, our goal is to design an intuitive, lightweight, and detector-aligned control signal. DC Score only relies on feature centroid

Table 9: Quantitative comparison of replay diversity under the replay budget of 500 samples. TCE, CLIP Dist., and LPIPS measure distributional, semantic, and perceptual diversity, respectively. Higher values indicate greater diversity.

Replay Strategy	TCE ↑	CLIP Dist. ↑	LPIPS ↑
Sample Replay	38.46	0.3786	0.5161
Generative Replay	**39.64**	**0.4205**	**0.5600**

Table 10: Quantitative privacy evaluation of generative replay. ID Recall measures whether generated identities can be matched to the training identities using face-recognition models. Replay Distinguishment reports the accuracy of a classifier for distinguishing generated replay from actual replay. Lower ID Recall and test accuracy closer to chance indicate stronger privacy preservation.

Metric	Result (%)
ID Recall with ArcFace ↓	0.08
ID Recall with CosFace ↓	0.05
Replay Distinguishment – Train Acc.	96.88
Replay Distinguishment – Test Acc. ↓	46.50
(Random Guess)	50.00

distance, introduces trivial overhead, and already consistently and robustly improves over fixed-s_t variants.

As for L2 over Cosine Similarity: CosSim mainly captures angular consistency; L2 is additionally sensitive to feature magnitude and centroid displacement, making it more suitable for measuring confusion across evolving tasks.

E.3 Fidelity of Non-Diffusion Replays

Our goal is not to reproduce each historical manipulation process at the **pixel level**, but to **distill the common forgery-aware representations**, thus mitigating forgetting. In other words, replay samples do not need to exactly match the original samples, as long as they retain the discriminative forgery-aware information of previous tasks.

Empirically, although replays are generated by a unified LDM generator, we still achieve stable retention on traditional forgery datasets (e.g., DFDCP, CDF in Tab. 1). It suggests that the replay does not simply degenerate into "LDM-style artifacts", but can effectively preserve the common forgery information from prior tasks.

E.4 Further Privacy Evidence

We provide additional quantitative evidence for the privacy preservation of generative replay in Tab. 10. First, we evaluate whether the identities of generated replay samples can be matched to identities in the original training set using ArcFace and CosFace. The low identity-recall rates obtained by both recognition models indicate that the replay generator does not directly reproduce training identities. We further train a binary classifier to distinguish generated replay samples from historical samples. Although the classifier achieves high accuracy on the training split, its performance drops to near-random accuracy on the held-out test split, suggesting that generated and actual replay samples do not exhibit distinguishable patterns. In summary, these results show that generative replay preserves task-relevant distributions without directly exposing or memorizing historical identities.

Table 11: Quantitative comparison in terms of AUC (%) under the incremental protocol adopted by prior IFFD methods. The best result in each column is highlighted in bold.

Method	FF++	DFDCP	DFD	CDF2	Avg.
DMP	**91.61**	84.86	91.81	91.67	**89.99**
DevDF	90.71	**90.31**	93.12	85.15	89.82
Ours	83.89	82.35	**94.81**	**98.33**	89.84

Table 12: Cross-domain generalization performance in terms of AUC (%) on unseen forgery datasets. The best results are highlighted in bold.

Method	DFD	DDIM	UADFV	WDF
SUR-LID	73.31	91.20	95.76	68.88
Ours	**77.97**	**92.45**	**96.25**	**70.73**

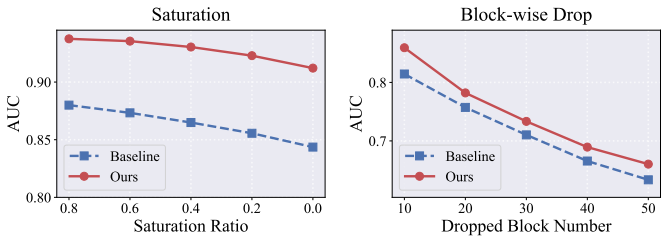


Figure 13: Robustness evaluation under saturation changes and block-wise drop with different perturbation strengths.

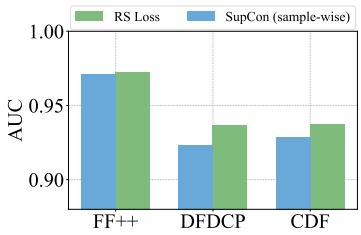


Figure 14: Comparison between RS Loss and SupCon.

E.5 Protocol-Aligned Performance

To facilitate direct comparison with existing IFFD methods, we further evaluate Dual-CARE under the incremental protocol adopted by prior work. As shown in Tab. 11, our method achieves competitive average performance while obtaining the best results on DFD and CDF2, demonstrating its effectiveness under established benchmark settings.

E.6 Adaptation to IFFD

Our strategy uses the DC Score to uniquely address domain confusion caused by rapid generator iterations in forgery detection. By specifically regulating domain-risky samples, it ensures clear separation between real and fake distributions, a task-specific adaptation missing in generic frameworks.

E.7 Robustness

We evaluate the robustness of Dual-CARE under common image perturbations, including saturation changes and block-wise drop. As shown in Fig. 13, our method maintains stable detection performance across different perturbation strengths, demonstrating its resistance to practical image degradations and local information loss.

E.8 Cross-domain Performance

We further evaluate cross-domain generalization on several unseen forgery datasets. As reported in Tab. 12, Dual-CARE consistently outperforms SUR-LID across all datasets, demonstrating that the proposed confusion-aware replay strategy not only preserves prior knowledge but also improves generalization to unseen forgery distributions.

E.9 RS Loss over Sample-level Loss

RS Loss is distribution-level to measure overall confusion. It is less affected by sample-wise outliers and noisy replay artifacts, and is therefore better suited to capture the overall drift among domains. In Fig. 14, comparisons between RS Loss and SupCon demonstrate our claims.