HIERARCHICAL ADAPTATION AND CAUSAL DECONFOUNDING FOR GENERALIZABLE DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing deepfake detectors often suffer from poor generalization to unseen domains and manipulation types, which can be attributed to two intrinsic limitations. First, prevailing adaptation methods apply uniform fine-tuning across all backbone layers, overlooking the hierarchical representational heterogeneity of vision transformers: layers encoding fundamentally different information require distinct adaptation capacities, and a one-size-fits-all strategy can limit cross-distribution transfer. Second, models tend to rely on forgery-irrelevant confounding factors, inducing spurious correlations that obscure reliable forgery-related evidence; such shortcut dependencies can break down when the detector is transferred across datasets or manipulation pipelines. To address these two limitations, we propose a novel detection framework with two complementary modules: Adaptive Subspace Residual Tuning (ASRT) and Confounder-Suppressed Prototype Decomposition (CSPD). Specifically, to resolve the mismatch between uniform adaptation and hierarchical layer characteristics, ASRT exploits the singular-value spectrum of each layer to decompose pretrained parameters into a frozen dominant subspace and a trainable residual subspace. Unlike methods that impose a uniform decomposition rank across layers, it allocates layer-specific adaptation capacity according to each layer's intrinsic representational property, better balancing prior preservation and forgery cue capture. To mitigate reliance on forgery-irrelevant confounders, CSPD constructs a learnable prototype pool to explicitly model recurring forgery-irrelevant nuisance factors and adaptively combines relevant prototypes to estimate and suppress their contributions in the feature space. This causally motivated feature-space residualization reduces spurious dependencies and encourages the detector to focus on forgery-related evidence. Extensive cross-dataset and cross-method experiments demonstrate that our method achieves the highest average AUC across the evaluated methods, showing strong generalization to unseen domains and manipulation types.

1 INTRODUCTION

Recent advances in generative artificial intelligence have substantially improved the realism of synthesized images and videos (Goodfellow et al., 2020; Rombach et al., 2022; Yan et al., 2025b). Their misuse facilitates identity impersonation and threatens privacy and media authenticity (Lee et al., 2024; Busacca & Monaca, 2023), creating an urgent need for deepfake detectors that generalize to unseen forgery types.

In response to this challenge, recent studies have increasingly leveraged large-scale pretrained vision and vision-language models, particularly CLIP (Radford et al., 2021), for deepfake detection. Although CLIP provides transferable representations, its emphasis on high-level semantics limits its ability to capture fine-grained forgery artifacts. Existing methods therefore adapt pretrained models to enhance their sensitivity to forgery-related cues (Zhou et al., 2022; Houlsby et al., 2019; Kong et al., 2025; Yan et al., 2025a; Cui et al., 2025). Nevertheless, their generalization to unseen forgery types remains limited.

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.



(a) Input image (b) Ground-truth mask (c) w/o Deconfounder (d) w Deconfounder

Figure 1: Visual comparison of model feature attribution with and without prototype deconfounding.

Despite their different designs, these approaches largely focus on the adaptation mechanism itself, while giving limited consideration to the distinct representational characteristics of different backbone layers, which encode different types of information: shallow layers primarily capture fine-grained visual patterns, whereas deeper layers encode high-level semantics. Moreover, from a causal perspective, feature extraction often fails to disentangle causal factors from spurious factors. As illustrated in Fig. 1(c), the extracted representation contains not only causal factors associated with the forgery process, but also spurious factors such as background, identity, and texture style. These spurious factors are unrelated to the underlying manipulation process and thus provide no causal evidence for detecting forged content. A detector may therefore rely on spurious correlations instead of learning stable and reliable causal features, which ultimately limits its generalization ability.

To address these limitations, we propose a novel framework for Deepfake detection based on hierarchical adaptive tuning and causal prototype deconfounding. For hierarchical adaptation, we introduce Adaptive Subspace Residual Tuning (ASRT) to target the mismatch between uniform fine-tuning and layer-wise representational heterogeneity. Specifically, ASRT analyzes the singular-value distribution of each weight matrix and decomposes it into a dominant subspace and a residual subspace. The dominant subspace preserves the general visual knowledge learned during pretraining and is kept frozen, whereas the residual subspace serves as a trainable adaptation space for capturing task-specific cues relevant to forgery detection. This design enables layer-wise adaptive parameter updates, allowing shallow and deep layers to learn according to their respective representational properties and thereby make fuller use of their complementary strengths. To mitigate the interference of spurious factors in discriminative representations, we further propose Confounder-Suppressed Prototype Decomposition (CSPD) from a causal perspective to explicitly model and remove forgery-irrelevant confounding components. CSPD is designed to reduce the model's reliance on forgery-irrelevant factors such as background and identity during adaptation. It constructs a learnable prototype pool and applies contrastive constraints in the feature space to model the shared distribution of spurious factors. During prediction, these factors are removed from the representation so that the model focuses on causal features and becomes less susceptible to spurious correlations. As shown in Fig. 1(d), CSPD effectively reduces the model's dependence on irrelevant factors in the feature space, thereby improving generalization to unseen forgery types.

Our main contributions are summarized as follows:

1. We propose ASRT, which exploits layer-wise singular-value distributions to allocate layer-specific adaptation capacity according to the hierarchical representations of vision backbones.

2. We introduce CSPD, a causally motivated prototype-based decomposition that explicitly models and suppresses forgery-irrelevant confounders, encouraging representations to focus on forgery-relevant evidence and reducing spurious correlations.

3. Cross-dataset and cross-method experiments demonstrate that our method achieves state-of-the-art average performance.

2 Related Work

2.1 Traditional Deepfake Detection

Traditional deepfake detectors learn discriminative representations from manipulation artifacts. Early methods exploit frequency anomalies, high-frequency residuals, reconstruction discrepancies, and representation consistency (Qian et al., 2020; Luo et al., 2021; Cao et al., 2022; Ni et al., 2022). Subsequent studies focus on localized manipulation regions, boundary artifacts, identity interference, and complementary forensic evidence (Shuai et al., 2023; Larue et al., 2023; Huang et al., 2023; Luo et al., 2023; Ba et al., 2024; Tian et al., 2024). However, reliance on manipulation-specific artifacts or statistical priors can limit generalization to unseen forgeries. Existing methods address this issue by diversifying the training distribution through self-blending, adversarial augmentation, curriculum learning, or latent-space augmentation (Shiohara & Yamasaki, 2022; Chen et al., 2022; Lin et al., 2024; Yan et al., 2024a). Alternatively, UCF, ProDet, VbSaT, and MAP-Mamba learn transferable representations from common forgery features, generation processes, spatiotemporal cues, or diverse artifacts, respectively (Yan et al., 2023a; Cheng et al., 2024; Yan et al., 2025c; Wang et al., 2026).

2.2 Vision-Language-Model-Based Deepfake Detection

Large-scale vision-language models such as CLIP (Radford et al., 2021) provide transferable visual–semantic priors and have been increasingly adopted for deepfake detection. Existing methods adapt these priors through input reprogramming (Lin et al., 2025), meta-learning with mixture-of-experts architectures (Lai et al., 2025), parameter-efficient adapters (Cui et al., 2025), and multimodal feature enhancement (Chen et al., 2025b). More closely related to our work, Effort (Yan et al., 2025a) employs orthogonal subspace decomposition to learn forgery patterns while preserving pretrained knowledge, but uses a uniform truncation rank across backbone layers. ResProto-FD (Jing et al., 2026) constructs residual prototypes to capture diverse forgery cues. In contrast, our method determines the trainable subspace separately for each layer, while its prototypes model forgery-irrelevant nuisance components rather than discriminative forgery patterns.

2.3 Causal Representation Learning

Causal representation learning (CRL) (Schölkopf et al., 2021) has recently been explored to improve the generalization of deepfake detectors by separating causal evidence from spurious factors. Bai et al. (2025) mitigate feature-level correlations between irrelevant attributes and class labels. UDD (Fu et al., 2025) reduces content and positional biases in CLIP representations through token-level shuffling and mixing. Nevertheless, existing approaches mainly suppress biases through feature regularization, perturbation, or filtering, without explicitly modeling the distribution of forgery-irrelevant confounding factors.

In summary, existing deepfake detection methods either overlook the hierarchical representational heterogeneity of vision backbones during model adaptation, or fail to explicitly disentangle confounding factors from causal forensic features. To address these two limitations, we propose a hierarchical adaptation and causal deconfounding framework for generalizable deepfake detection.

3 Method

The framework of our proposed method is illustrated in Fig. 2. The input image is converted into patch embeddings and processed by a CLIP-pretrained ViT backbone, which is hierarchically fine-tuned using Adaptive Subspace Residual Tuning (ASRT). Confounder-Suppressed Prototype Decomposition (CSPD) then removes forgery-irrelevant confounding components from the extracted representation, after which a classifier produces the final prediction. The entire framework is jointly optimized with multiple loss functions.

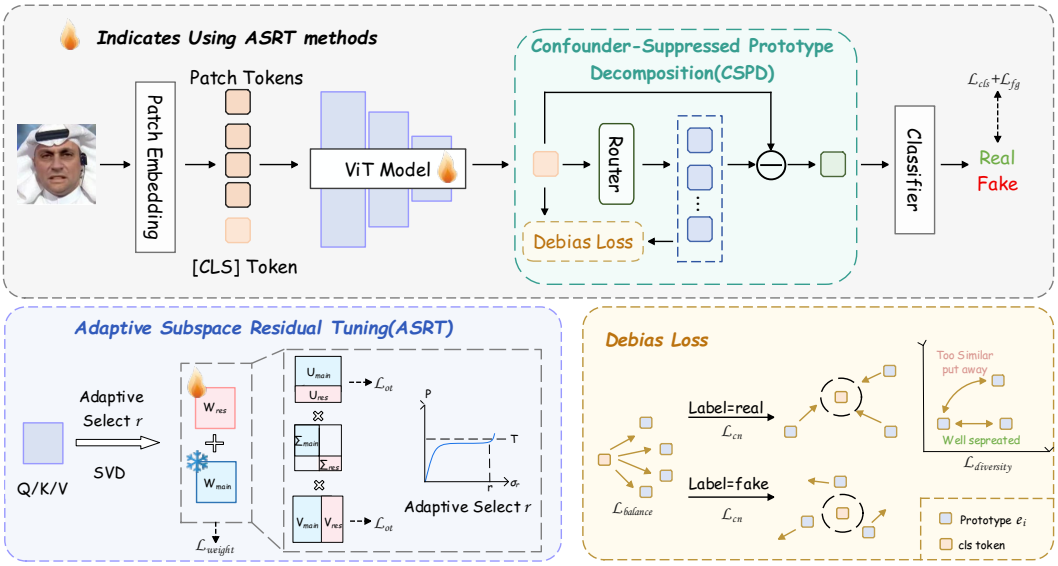


Figure 2: Overview of the proposed framework.

3.1 Adaptive Subspace Residual Tuning

Pretrained vision models learn hierarchical representations: shallow layers mainly capture low-level visual patterns, whereas deeper layers encode more abstract semantic information (Gandelsman et al., 2023; Tong et al., 2024; Yao et al., 2024; Chen et al., 2025a). Accounting for this representational hierarchy is therefore important for capturing fine-grained forgery cues while preserving transferable semantic priors. To address this issue, we propose Adaptive Subspace Residual Tuning (ASRT), which uses the singular-value distribution of each pretrained weight matrix to determine a layer-specific partition between a frozen dominant subspace and a trainable residual subspace, thereby preserving pretrained knowledge while adapting the model to fine-grained forgery cues.

Specifically, let each two-dimensional attention weight matrix in every layer of the pretrained vision backbone be denoted by $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. Applying SVD yields

$$W = U \Sigma V^{\top}. \tag{1}$$

Here, Σ is a diagonal matrix whose singular values are arranged in descending order, while U and V contain the corresponding left and right singular vectors, respectively. Because different network layers focus on different visual properties, their singular-value distributions also differ. To better adapt to the representational characteristics of each layer, we compute the cumulative proportion of the dominant singular values, where the denominator sums over all singular values of W. A threshold T is then used to determine the corresponding subspace dimension adaptively:

$$r = \arg\min \left\{ k \;\middle|\; \frac{\sum_{i=1}^{k} \sigma_i}{\sum_j \sigma_j} \geq T \right\}. \tag{2}$$

Using the resulting subspace dimension r, the weight matrix is decomposed into a dominant component and a residual component:

$$W_{\text{main}} = U_{\text{main}} \Sigma_{\text{main}} V_{\text{main}}^{\top}, \qquad W_{\text{res}} = U_{\text{res}} \Sigma_{\text{res}} V_{\text{res}}^{\top}. \tag{3}$$

The subscript "main" denotes the r-dimensional dominant subspace, whereas "res" denotes the remaining dimensions. Because different layers encode different types of information, their singular-value spectra also differ. Under the same cumulative-energy threshold, the retained dominant rank r therefore need not be identical across layers. Consequently, the residual subspace has a layer-specific dimensionality, resulting in different numbers of trainable residual directions across layers. This mechanism adaptively controls the capacity of the trainable subspace according to each layer's representational characteristics. Under the layer-wise rank allocation, shallow layers yield smaller

dominant ranks r, leaving more trainable residual directions for capturing local forgery traces. In contrast, deeper layers yield larger dominant ranks and fewer trainable residual directions, thereby preserving more pretrained semantic information. The empirical distribution of layer-wise residual ranks aligns with the hierarchical properties described above. A detailed visualization is provided in Appendix A.2 (Fig. 5). ASRT can therefore preserve semantic priors while improving sensitivity to fine-grained manipulation traces.

3.2 Confounder-Suppressed Prototype Decomposition

Although ASRT enhances the extraction of fine-grained forgery cues through layer-adaptive parameter updates, the resulting representation may still contain forgery-irrelevant factors, such as identity, background, and image style. These factors may be spuriously correlated with class labels in the training data, causing the detector to rely on domain-specific information that does not generalize to unseen datasets or manipulation methods.

We analyze this problem from a causal perspective. As illustrated in Fig. 3, the unobserved variable U jointly affects the forgery-related factor Z_d and the forgery-irrelevant factor Z_c. Both factors contribute to the extracted representation f:

$$Z_d = g_d(U, \epsilon_d), Z_c \quad = g_c(U, \epsilon_c), f = h(Z_d, Z_c, \epsilon_f), \tag{4}$$

where g_d, g_c, and h denote the corresponding structural functions, and ϵ_d, ϵ_c, and ϵ_f are mutually independent exogenous disturbances. In Fig. 3, the node Y denotes the model prediction.

Under biased learning, the prediction may depend not only on the forgery-related evidence in Z_d but also on the nuisance information in Z_c. Because the relationship between Z_c and the label may change across datasets or manipulation methods, reliance on such information can reduce generalization. The red bar in Fig. 3(b) schematically represents the suppression of the influence of Z_c on the prediction.

Since Z_d and Z_c are latent and cannot be directly observed, we do not attempt to identify them explicitly. Instead, we adopt an additive approximation in the feature space:

$$f \approx f_d + f_c, \tag{5}$$

where f_d carries the forgery-related evidence associated with Z_d, where f_c carries the forgery-irrelevant nuisance information associated with Z_c. Based on this approximation, we propose Confounder-Suppressed Prototype Decomposition (CSPD) to estimate and suppress the nuisance-related component.

Specifically, let $f \in \mathbb{R}^d$ denote the input representation. CSPD constructs a prototype pool containing N learnable prototype vectors:

$$E = \{e_1, e_2, \ldots, e_N\}, \qquad e_i \in \mathbb{R}^d. \tag{6}$$

Each prototype e_i is intended to capture a recurring nuisance-related pattern. Since the nuisance factors associated with different samples may vary, CSPD adaptively selects and combines the prototypes according to the input representation.

We first compute the routing probabilities as

$$p = \mathrm{softmax}(W_r f), \qquad p_i \in [0, 1], \quad i = 1, \ldots, N. \tag{7}$$

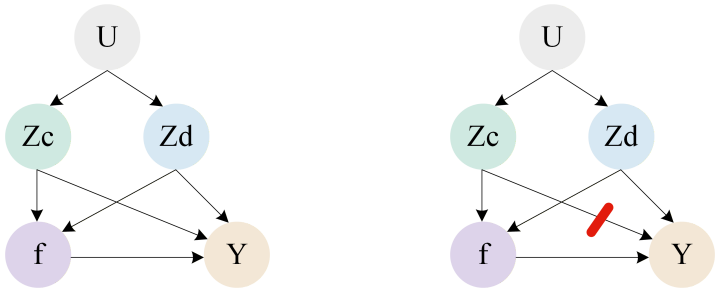
Here, $W_r \in \mathbb{R}^{N \times d}$ denotes the learnable routing matrix, and p_i denotes the routing probability assigned to the i-th prototype for the current sample.

We then select the Top-k prototypes with the largest routing probabilities to form the set $\mathcal{T}$. The selected probabilities are renormalized within $\mathcal{T}$, and the corresponding prototypes are combined to construct a sample-specific nuisance component:

$$\widetilde{p}_i = \frac{p_i}{\sum_{j \in \mathcal{T}} p_j}, \qquad f_c = \sum_{i \in \mathcal{T}} \widetilde{p}_i e_i. \tag{8}$$

Where f_c denotes the nuisance-related component constructed from the selected prototypes. The nuisance-suppressed representation is subsequently obtained by

$$f_d \approx f - f_c. \tag{9}$$



(a) Biased Learning (b) Unbiased Learning

Figure 3: Causal graphs illustrating how spurious correlations arise during task adaptation and how their influence is suppressed.

The resulting representation f_d is passed to the classifier for final prediction. By adaptively selecting and combining sample-relevant prototypes, CSPD estimates and suppresses heterogeneous forgery-irrelevant components, thereby reducing their influence on classification and encouraging the classifier to rely more strongly on forgery-related evidence. The prototype routing adapts to input characteristics rather than following a fixed usage pattern, for which corresponding visualization and statistical analysis are detailed in Appendix A.2(Fig. 6). Overall, CSPD performs a causally motivated feature-space residualization rather than an exact intervention on an identified causal variable.

3.3 Optimization Objectives

We adopt the standard cross-entropy loss $\mathcal{L}_{\text{cls}}$ for binary classification. Beyond this basic supervision, we introduce $\mathcal{L}_{\text{debias}}$ to guide CSPD, $\mathcal{L}_{\text{ot}}$ and $\mathcal{L}_{\text{weight}}$ to regularize ASRT, and $\mathcal{L}_{\text{fg}}$ to balance the learning of authentic and forged samples.

Debiasing Loss. To guide the prototype pool toward diverse forgery-irrelevant factors, we introduce three constraints. The first regulates the similarity between a feature f and its most relevant prototype:

$$\mathcal{L}_{\text{cn}} = \begin{cases} 1 - \max\limits_{j} \cos(f, e_j), & y = 0, \\ \max\left(0, \max\limits_{j} \cos(f, e_j) - m\right), & y = 1, \end{cases} \tag{10}$$

where e_j denotes the j-th prototype and m is a margin. Since authentic images contain no manipulation traces, their representations primarily encode the intrinsic characteristics of authentic images. By increasing the similarity between each authentic sample and its most relevant prototype, we encourage the prototypes to learn these authentic-image characteristics. Because the forged images considered in this work are generated by locally manipulating authentic images, they still retain substantial authentic-image information. We therefore do not enforce complete separation between forged samples and the prototypes. Instead, the margin m limits their maximum similarity, allowing the prototypes to preserve the normal authentic-image characteristics present in forged samples. Since these characteristics may become spuriously correlated with class labels, this asymmetric constraint encourages the prototypes to primarily capture such forgery-irrelevant factors. We further encourage balanced prototype utilization and prevent prototype collapse through

$$\mathcal{L}_{\text{balance}} = 1 - \cos(\overline{p}, u), \tag{11}$$

$$\mathcal{L}_{\text{diversity}} = \frac{1}{N(N-1)} \sum_{i \neq j} \text{ReLU}\left(\cos(e_i, e_j) - \tau\right), \tag{12}$$

where $\overline{p}$ is the batch-averaged routing distribution, u is the uniform distribution, N is the number of prototypes, and τ is a similarity threshold. The complete debiasing loss is

$$\mathcal{L}_{\text{debias}} = \lambda_{\text{cn}}\mathcal{L}_{\text{cn}} + \lambda_{\text{balance}}\mathcal{L}_{\text{balance}} + \lambda_{\text{diversity}}\mathcal{L}_{\text{diversity}}. \tag{13}$$

ASRT Regularization. We regularize the residual subspace using orthogonality and weight constraints:

$$\mathcal{L}_{\text{ot}} = \frac{1}{2}\left\|U_{\text{res}}^{\top}U_{\text{res}} - I\right\|_F + \frac{1}{2}\left\|V_{\text{res}}^{\top}V_{\text{res}} - I\right\|_F, \tag{14}$$

$$\mathcal{L}_{\text{weight}} = -\frac{1}{M}\sum_{l=1}^{M}\text{mean}\left(\sigma^{(l)}\right), \tag{15}$$

where M is the number of adapted layers and $\sigma^{(l)}$ contains the residual singular values of the l-th layer. The orthogonality loss $\mathcal{L}_{\text{ot}}$ encourages different directions in the residual subspace to remain approximately orthogonal, thereby reducing redundancy among them. The weight loss $\mathcal{L}_{\text{weight}}$ maintains sufficient strength in the trainable residual transformations, enabling the model to learn new task-specific features for forgery detection.

Fine-Grained Balance Loss. Let R and F denote the index sets of authentic and forged samples in a training batch, respectively. Since $|R| \neq |F|$, the standard classification loss may be dominated by the more frequent class. We therefore define

$$\mathcal{L}_{\text{fg}} = \frac{1}{|R|}\sum_{i\in R}\text{CE}(q_i, y_i) + \frac{1}{|F|}\sum_{i\in F}\text{CE}(q_i, y_i), \tag{16}$$

where q_i and y_i denote the predicted class probability and ground-truth label of the i-th sample, respectively.

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1\mathcal{L}_{\text{debias}} + \lambda_2\mathcal{L}_{\text{ot}} + \lambda_3\mathcal{L}_{\text{weight}} + \lambda_4\mathcal{L}_{\text{fg}}. \tag{17}$$

4 Experiments

4.1 Experimental Settings

Datasets. We adopt two widely used and standard protocols for evaluation: **Protocol 1:** Under this setting, we evaluate the cross-dataset generalization ability of the proposed method following the protocol of DeepfakeBench (Yan et al., 2023b). FaceForensics++ (FF++) (Rossler et al., 2019) with medium compression (c23) is used for training, while evaluation is conducted on several widely used benchmarks, including Celeb-DF-v2 (Li et al., 2020), DeepFake Detection (DFD) (Google AI, 2019), DeepFake Detection Challenge (DFDC) (Dolhansky et al., 2020), and DFDC Preview (DFDCP) (Dolhansky et al., 2019). **Protocol 2:** Under this setting, we evaluate the cross-method generalization ability of the proposed method on the latest dataset DF40 (Yan et al., 2024b). All test samples are generated within the FF++ domain using different manipulation methods, enabling evaluation under cross-method scenarios.

Implementation details. All experiments are implemented with PyTorch 2.4.1 and conducted on a single NVIDIA RTX 3090 GPU. The model uses pretrained CLIP ViT-L/14 as the backbone. All implementation details follow the standard DeepfakeBench pipeline. The batch size is set to 16 during training and 32 during testing. For each video, 32 frames are uniformly sampled during both training and testing. We use the Adam optimizer (Kingma & Ba, 2014), with a learning rate of 2×10^{-4} and a weight decay of 5×10^{-4}. The relevant hyperparameters are set as $\lambda_{\text{cn}} = 5$, $\lambda_{\text{balance}} = 1$, $\lambda_{\text{diversity}} = 10$, $\lambda_1 = 1$, $\lambda_2 = 0.01$, $\lambda_3 = 0.1$, and $\lambda_4 = 1$. For the evaluation metric, we report frame-level AUC and video-level AUC for comparison with other methods.

4.2 Generalization Experiments

We evaluate the generalization ability of our method under two evaluation protocols. **Protocol 1: Cross-dataset evaluation.** As shown in Table 1, our method achieves the best average performance and ranks first on multiple test datasets, demonstrating strong generalization across different data distributions. **Protocol 2: Cross-method evaluation.** As shown in Table 2, our method achieves the best average performance and ranks among the top two across all tested manipulation methods, demonstrating strong generalization to unseen forgery types. Compared with the competing methods, our framework better accounts for layer-wise representational differences and reduces reliance on forgery-irrelevant information, leading to more stable and transferable representations across datasets and manipulation methods. More experimental results are reported in Appendix A.1.

Table 1: Frame-level AUC results for cross-dataset generalization.

Method	Venue	CDF-v2	DFDCP	DFDC	DFD	Avg
UCF (Yan et al., 2023a)	ICCV'23	0.753	0.759	0.719	0.807	0.7595
ED (Ba et al., 2024)	AAAI'24	0.864	0.851	0.721	–	–
ProDet (Cheng et al., 2024)	NeurIPS'24	0.845	0.812	0.724	0.858	0.8098
LSDA (Yan et al., 2024a)	CVPR'24	0.830	0.815	0.736	0.880	0.8153
RepDFD (Lin et al., 2025)	AAAI'25	0.823	0.826	0.770	0.848	0.8168
UDD (Fu et al., 2025)	AAAI'25	0.869	0.856	0.758	0.910	0.8483
Effort (Yan et al., 2025a)†	ICML'25	0.879	0.864	0.824	0.923	0.8725
CRDA (Chou et al., 2026)	AAAI'26	0.854	0.797	0.742	–	–
ResProto-FD (Jing et al., 2026)	AAAI'26	0.838	–	0.795	0.914	–
Ours	–	**0.903**	**0.881**	**0.851**	**0.948**	**0.8958**

† indicates results reproduced by us.

Table 2: Video-level AUC results for cross-method generalization.

Method	Venue	UniFace	e4s	FaceDan	InSwap	SimSwap	BleFace	Avg
UCF (Yan et al., 2023a)	ICCV'23	0.831	0.731	0.862	0.809	0.647	0.827	0.7845
IID (Huang et al., 2023)	CVPR'23	0.839	0.766	0.844	0.789	0.644	0.789	0.7785
LSDA (Yan et al., 2024a)	CVPR'24	0.872	0.694	0.721	0.855	0.793	0.875	0.8017
ProDet (Cheng et al., 2024)	NeurIPS'24	0.908	0.771	0.747	0.837	0.844	0.929	0.8393
CDFA (Lin et al., 2024)	ECCV'24	0.762	0.631	0.803	0.772	0.757	0.756	0.7468
Effort (Yan et al., 2025a)	ICML'25	0.962	**0.983**	0.926	0.936	0.926	0.873	0.9343
VbSaT (Yan et al., 2025c)	CVPR'25	0.960	0.980	0.912	0.937	0.931	0.906	0.9377
FA† (Cui et al., 2025)	CVPR'25	0.951	0.969	0.946	0.948	0.913	0.880	0.9345
MAP-Mamba (Wang et al., 2026)	TIFS'26	0.815	0.779	0.826	0.855	0.813	0.896	0.8307
TSRL (Lei et al., 2026)	CVPR'26	0.954	**0.983**	0.948	0.937	0.904	0.902	0.9380
Ours	–	**0.972**	0.981	**0.971**	**0.954**	**0.945**	**0.937**	**0.9600**

† indicates results reproduced by us.

4.3 Ablation Studies

We present ablations on the two core modules, *i.e.* ASRT and CSPD, and training objectives in this section, while the remaining ablation studies are provided in Appendix A.2.

Module-wise analysis. Table 3 reports the ablation results for different module combinations. When used alone, CSPD improves performance by suppressing forgery-irrelevant components, whereas ASRT adaptively partitions the dominant and residual subspaces according to the representational characteristics of each layer, enabling layer-specific parameter updates. Their combination achieves the best overall performance: ASRT produces layer-adapted forgery representations, which CSPD further refines by removing forgery-irrelevant components, resulting in more reliable and transferable features.

Table 3: Frame-level AUC ablation results for ASRT and CSPD.

ASRT	CSPD	CDF-v2	DFDCP	DFDC	DFD	Avg
–	–	0.749	0.702	0.704	0.827	0.7455
–	✓	0.770	0.756	0.725	0.854	0.7763
✓	–	0.887	0.868	0.833	0.939	0.8818
✓	✓	**0.903**	**0.881**	**0.851**	**0.948**	**0.8958**

Ablation of Objective Terms. Table 4 analyzes the contribution of each objective term. $\mathcal{L}_{\text{debias}}$ suppresses forgery-irrelevant factors. $\mathcal{L}_{\text{weight}}$ maintains sufficient residual adaptation strength by encouraging non-negligible singular values. $\mathcal{L}_{\text{ot}}$ promotes orthogonality among residual subspace directions, reducing redundancy and increasing representation diversity. $\mathcal{L}_{\text{fg}}$ separately normalizes the losses of authentic and forged samples, thereby reducing the dominance of the more frequent class. The consistent performance gains confirm the complementary roles of these objective terms.

Comparison of adaptation methods. Table 5 compares ASRT with representative adaptation methods under the same experimental setting. The proposed ASRT achieves the best performance on all

Table 4: Frame-level AUC results for different objective-term combinations.

$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{debias}}$	$\mathcal{L}_{\text{weight}}$	$\mathcal{L}_{\text{ot}}$	$\mathcal{L}_{\text{fg}}$	CDF-v2	DFDCP	DFDC	Avg
✓					0.874	0.837	0.803	0.8380
✓	✓				0.878	0.862	0.835	0.8583
✓	✓	✓			0.887	0.877	0.828	0.8640
✓	✓	✓	✓		0.897	0.867	0.837	0.8670
✓	✓	✓	✓	✓	**0.903**	**0.881**	**0.851**	**0.8783**

test sets as well as the best average result. Unlike alternative adaptation strategies, ASRT explicitly accounts for layer-wise representational differences by determining the trainable subspace from the singular-value distribution of each layer. This enables a better balance between preserving pretrained knowledge and adapting to fine-grained forgery cues, leading to stronger cross-dataset generalization.

Table 5: Video-level AUC comparison of different adaptation methods.

Method	DFDCP	DFDC	DFD	Avg.
MoE-ffd (Kong et al., 2025)	0.850	–	0.936	–
Effort (Yan et al., 2025a)	0.909	0.843	0.965	0.9057
ASRT (Ours)	**0.913**	**0.856**	**0.976**	**0.9150**

Comparison of confounder-suppression strategies. As shown in Table 6, we compare CSPD against two alternative deconfounding strategies under the identical ASRT backbone and evaluation protocol. "Mask + Adversarial" randomly masks feature dimensions during training and applies an adversarial classifier to the removed features, where as "Query + Attention" uses query-based cross-attention to aggregate learnable nuisance tokens and subtracts the resulting nuisance representation. CSPD models forgery-irrelevant information through a combination of multiple learnable prototypes. Each prototype represents a recurring forgery-irrelevant pattern, and their weighted combination constitutes the forgery-irrelevant component of the current sample. The prototypes shared across samples provide a relatively stable structural reference for estimating forgery-irrelevant components. This enables CSPD to identify and reconstruct sample-relevant nuisance components more precisely, thereby reducing the risk of inadvertently removing discriminative forgery cues during nuisance suppression.

Table 6: Frame-level AUC comparison of different confounder suppression strategies.

Method	CDF-v2	DFDCP	DFDC	DFD	Avg
Mask + Adversarial	0.882	0.866	0.836	0.941	0.8813
Query + Attention	0.898	0.854	0.838	0.943	0.8833
CSPD (Ours)	**0.903**	**0.881**	**0.851**	**0.948**	**0.8958**

5 Conclusion

We propose a generalizable deepfake detection framework that combines ASRT with CSPD. ASRT performs layer-specific adaptation based on singular-value distributions, preserving pretrained knowledge while improving sensitivity to fine-grained forgery cues. CSPD learns a prototype pool to model and suppress forgery-irrelevant components, reducing the influence of spurious correlations. Extensive cross-dataset and cross-method experiments demonstrate the effectiveness and generalization ability of the proposed framework. One limitation is that ASRT determines its adaptation space primarily from the pretrained weight structure, without considering individual inputs or target domains. Moreover, CSPD may not capture all complex nuisance factors. Future work will explore dynamic subspace adaptation and more effective prototype constraints.

AI USE STATEMENT

During the preparation of this manuscript, we used generative AI tools for language polishing and limited code generation. Their use in writing was restricted to improving grammar, sentence structure, and overall readability. These tools were not used to formulate research ideas, analyze data, draw conclusions, or generate substantive scientific content. All AI-assisted text was reviewed and revised by the authors, and all AI-generated code was manually inspected and tested. The authors take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 provides detailed descriptions of ASRT and CSPD, including the subspace decomposition, prototype construction, and training objectives. Section 4.1 describes the datasets and evaluation protocols. Appendix A.1 provides the implementation details, hyperparameter settings, and evaluation metrics, while Appendices A.2 and A.3 present additional generalization and ablation results, respectively. These methodological descriptions and experimental settings are provided to facilitate reproduction of the reported results.

REFERENCES

Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. Exposing the deception: Uncovering more forgery clues for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 719–728, 2024.

Ningning Bai, Xiaofeng Wang, Ruidong Han, Jianpeng Hou, Qin Wang, and Shanmin Pang. Towards generalizable face forgery detection via mitigating spurious correlation. *Neural Networks*, 182:106909, 2025.

Angela Busacca and Melchiorre Alberto Monaca. Deepfake: Creation, purpose, risks. In *Innovations and economic and social changes due to artificial intelligence: the state of the art*, pp. 55–68. Springer, 2023.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4113–4122, 2022.

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multimodal language models see better when they look shallower. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6688–6706, 2025a.

Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18710–18719, 2022.

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. X2-dfd: A framework for explainable and extendable deepfake detection. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 83792–83839. Curran Associates, Inc., 2025b. doi: 10.52202/085713-2809.

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.

Yuxuan Chou, Tao Yu, Wen Huang, Tao Dai, Shu-Tao Xia, et al. Improving deepfake detection with reinforcement learning-based adaptive data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3381–3389, 2026.

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

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

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.

Hongyan Fei, Zexi Jia, Chuanwei Huang, Jinchao Zhang, and Jie Zhou. Exploring specular reflection inconsistency for generalizable face forgery detection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KwXkLYmZvR.

Xinghe Fu, Zhiyuan Yan, Taiping Yao, Shen Chen, and Xi Li. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3040–3048, 2025.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Google AI. Contributing data to deefakedetection. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html, 2019. Accessed: 2021-11-13.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4490–4499, 2023.

Jiuyao Jing, Yu Zheng, and Chunlei Peng. Resproto-fd: Visual-language residual prototype sets for generalized face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5548–5556, 2026.

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

Chenqi Kong, Anwei Luo, Peijun Bao, Yi Yu, Haoliang Li, Zengwei Zheng, Shiqi Wang, and Alex C Kot. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *IEEE Transactions on Dependable and Secure Computing*, 2025.

Yingxin Lai, Hongyang Wang, Jing Yang, Xiangui Kang, Bin Li, Linlin Shen, and Zitong Yu. Gmdf: Generalized multi-scenario deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4300–4309, 2025.

Nicolas Larue, Ngoc-Son Vu, Vitomir Struc, Peter Peer, and Vassilis Christophides. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21011–21021, 2023.

Hao-Ping Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2024.

Zhanhe Lei, Zhongyuan Wang, Jikang Cheng, Baojin Huang, Yuhong Yang, Zhen Han, Chao Liang, and Dengpan Ye. Tutor-student reinforcement learning: A dynamic curriculum for robust deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41436–41445, June 2026.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3207–3216, 2020.

Kaiqing Lin, Yuzhen Lin, Weixiang Li, Taiping Yao, and Bin Li. Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5262–5270, 2025.

Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In *European conference on computer vision*, pp. 104–122. Springer, 2024.

Anwei Luo, Chenqi Kong, Jiwu Huang, Yongjian Hu, Xiangui Kang, and Alex C Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:1168–1182, 2023.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16317–16326, 2021.

Long Ma, Zhiyuan Yan, Jin Xu, Yize Chen, Qinglang Guo, Zhen Bi, Yong Liao, and Hui Lin. From specificity to generality: Revisiting generalizable artifacts in detecting face deepfakes. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 69306–69344. Curran Associates, Inc., 2025. doi: 10.52202/085713-2330. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/64569060d7e0966dc769d22813179c93-Paper-Conference.pdf.

Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12–21, 2022.

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.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18720–18729, 2022.

Chao Shuai, Jieming Zhong, Shuang Wu, Feng Lin, Zhibo Wang, Zhongjie Ba, Zhenguang Liu, Lorenzo Cavallaro, and Kui Ren. Locate and verify: A two-stream network for improved deepfake detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7131–7142, 2023.

Jiamu Sun, Zhiyuan Yan, Ke-Yue Zhang, Taiping Yao, and Shouhong Ding. Dfd-hr: Generalizable deepfake detection via hierarchical routing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13984–13995, June 2026.

Jiahe Tian, Peng Chen, Cai Yu, Xiaomeng Fu, Xi Wang, Jiao Dai, and Jizhong Han. Learning to discover forgery cues for face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:3814–3828, 2024.

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

Chi Wang, Ziwen He, Xinjue Hu, Weinan Guan, Wei Wang, and Zhangjie Fu. Map-mamba: Multi-artifacts perception mamba for generalizable face forgery detection. *IEEE Transactions on Information Forensics and Security*, 2026.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF 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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 4534–4565, 2023b.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF 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, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37:29387–29434, 2024b.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. In *International Conference on Machine Learning*, pp. 70268–70288. PMLR, 2025a.

Zhiyuan Yan, Junyan Ye, Weijia Li, Zilong Huang, Shenghai Yuan, Xiangyang He, Kaiqing Lin, Jun He, Conghui He, and Li Yuan. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. *arXiv preprint arXiv:2504.02782*, 2025b.

Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12615–12625, 2025c.

Huanjin Yao, Wenhao Wu, Taojiannan Yang, YuXin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense connector for mllms. *Advances in Neural Information Processing Systems*, 37:33108–33140, 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022.

A APPENDIX

A.1 ADDITIONAL GENERALIZATION EXPERIMENTS

Cross-dataset evaluation. As shown in Table 7, our method achieves the best average performance and ranks first on multiple test datasets, demonstrating strong generalization across different data distributions. **Cross-method evaluation.** As shown in Table 8, our method achieves the best average performance and ranks among the top two across all tested manipulation methods, demonstrating strong generalization to unseen forgery types. Compared with the competing methods, our framework better accounts for layer-wise representational differences and reduces reliance on forgery-irrelevant information, leading to more stable and transferable representations across datasets and manipulation methods.

Table 7: Video-level AUC results for cross-dataset generalization.

Method	Venue	CDF-v2	DFDCP	DFDC	DFD	Avg.
UCF (Yan et al., 2023a)	ICCV'23	0.837	0.770	0.742	0.867	0.8040
IID (Huang et al., 2023)	CVPR'23	0.838	0.689	0.700	0.939	0.7915
CFM (Luo et al., 2023)	TIFS'23	0.897	0.802	0.706	0.952	0.8393
CDFA (Lin et al., 2024)	ECCV'24	0.938	0.881	0.830	0.954	0.9008
LSDA (Yan et al., 2024a)	CVPR'24	0.875	0.812	0.701	0.881	0.8173
χ^2-DFD (Chen et al., 2025b)	NeurIPS'25	0.955	<u>0.912</u>	0.853	0.957	<u>0.9193</u>
UDD (Fu et al., 2025)	AAAI'25	0.931	0.881	0.812	0.955	0.8948
Effort (Yan et al., 2025a)	ICML'25	0.956	0.909	0.843	0.965	0.9183
ResProto-FD (Jing et al., 2026)	AAAI'26	0.915	–	0.819	0.958	–
DFD-HR (Sun et al., 2026)	CVPR'26	<u>0.960</u>	0.901	<u>0.865</u>	<u>0.980</u>	<u>0.9265</u>
Ours	–	**0.961**	**0.915**	**0.873**	**0.984**	**0.9333**

Table 8: Frame-level AUC results for cross-method generalization.

Method	Venue	UniFace	e4s	FaceDan	FSGAN	InSwap	SimSwap	Avg
UCF (Yan et al., 2023a)	ICCV'23	0.787	0.692	0.800	0.881	0.768	0.649	0.7628
LSDA (Yan et al., 2024a)	CVPR'24	0.854	0.684	0.759	0.832	0.810	0.727	0.7777
ProDet (Cheng et al., 2024)	NeurIPS'24	0.845	0.710	0.736	0.865	0.788	0.778	0.7870
CDFA (Lin et al., 2024)	ECCV'24	0.765	0.674	0.754	0.848	0.720	0.761	0.7537
FSgD (Ma et al., 2025)	NeurIPS'25	0.918	0.875	0.830	0.863	0.874	**0.910**	0.8783
χ^2-DFD (Chen et al., 2025b)	NeurIPS'25	0.852	0.912	0.838	0.899	0.784	0.849	0.8557
FA† (Cui et al., 2025)	CVPR'25	0.913	0.930	0.904	0.947	0.901	0.856	0.9085
ResProto-FD (Jing et al., 2026)	AAAI'26	–	<u>0.948</u>	–	<u>0.949</u>	0.893	0.857	–
SRI-Net (Fei et al., 2026)	ICLR'26	<u>0.920</u>	0.894	**0.953**	0.942	**0.911**	0.833	<u>0.9088</u>
Ours	–	**0.947**	**0.968**	<u>0.931</u>	**0.959**	<u>0.910</u>	<u>0.900</u>	**0.9358**

† indicates results reproduced by us.

A.2 ADDITIONAL ABLATION STUDIES

Ablation of the debiasing constraints. Table 9 analyzes the contributions of the individual constraints in $\mathcal{L}_{\text{debias}}$. $\mathcal{L}_{\text{cn}}$ guides the model to suppress forgery-irrelevant factors, $\mathcal{L}_{\text{balance}}$ encourages balanced prototype utilization, and $\mathcal{L}_{\text{diversity}}$ preserves diversity among prototypes. The consistent performance gains from these constraints and the best result achieved by their combination demonstrate their complementary roles in modeling nuisance factors.

Ablation of the Top-k value. Table 10 evaluates the effect of the number of prototypes selected for each sample. Performance peaks at $k = 2$, suggesting that a single prototype is insufficient to capture sample-specific nuisance variations, whereas selecting more prototypes introduces less relevant information. This result supports the design of CSPD, which captures complementary nuisance factors while limiting irrelevant noise.

Ablation of the number of prototypes. Table 11 evaluates the effect of the prototype-pool size. Performance peaks at $N = 20$. A smaller pool has insufficient capacity to represent diverse nuisance

Table 9: Frame-level AUC results for different combinations of debiasing constraints.

$\mathcal{L}_{\text{cn}}$	$\mathcal{L}_{\text{balance}}$	$\mathcal{L}_{\text{diversity}}$	CDF-v2	DFDC	DFD	Avg
			0.853	0.814	0.923	0.8633
✓			0.894	0.836	0.934	0.8880
✓	✓		0.898	0.846	0.940	0.8947
✓	✓	✓	**0.903**	**0.851**	**0.948**	**0.9007**

Table 10: Frame-level AUC results for different Top-k values.

Top-k	CDF-v2	DFDCP	DFDC	DFD	Avg.
$k = 1$	0.888	0.874	0.828	0.942	0.8830
$k = 2$	**0.903**	**0.881**	**0.851**	**0.948**	**0.8958**
$k = 4$	0.899	0.880	0.844	0.941	0.8910
$k = 6$	0.899	0.869	0.845	0.940	0.8883

patterns, whereas a larger pool introduces redundancy and increases optimization difficulty. These results show that a moderately sized prototype pool provides a better balance between nuisance coverage and the preservation of discriminative forgery information.

Table 11: Frame-level AUC results for different numbers of candidate prototypes.

N	CDF-v2	DFDCP	DFDC	DFD	Avg.
10	0.902	0.878	0.844	0.944	0.8920
20	0.903	0.881	0.851	**0.948**	**0.8958**
30	**0.905**	0.870	0.846	0.941	0.8905
40	0.886	**0.882**	**0.856**	**0.948**	0.8930

Ablation of the dominant-subspace ratio. Fig. 4 analyzes the effect of the dominant-subspace ratio $T = 1 - a(0.1)^n$, where a larger T retains a larger dominant subspace. The best generalization performance is achieved at $a = 2$ and $n = 6$. Performance degrades when the ratio is either too small or too large: the former weakens the preservation of pretrained priors, whereas the latter restricts the residual capacity for capturing task-specific forgery cues. This inverted-U trend confirms that ASRT effectively balances prior preservation and adaptive capacity.

Layer-wise effective-rank analysis. Fig. 5 shows the normalized effective ranks of the Q, K, and V activations across different backbone layers. The effective ranks generally increase with network depth, indicating that deeper-layer representations span broader effective subspaces across samples. This trend is consistent with the layer-wise rank allocation of ASRT: deeper layers preserve larger dominant subspaces and consequently have smaller trainable residual subspaces.These depth and projection-dependent patterns reveal substantial layer-wise representational heterogeneity in the backbone, providing empirical motivation for assigning adaptation capacity separately to different layers.

Prototype usage analysis. To examine the input-dependent routing behavior of the learned prototypes, we analyze the selection frequencies of 20 prototypes on FaceForensics++. As shown in Fig. 6, real samples exhibit a highly concentrated routing pattern, with most selections assigned to Prototypes 5 and 12. In contrast, fake samples activate a broader set of prototypes, including Prototypes 3, 6, 11, 13, and 20. This class-dependent difference indicates that the routing module does not apply a fixed prototype to all inputs. Instead, it adaptively selects different prototype combinations according to the manipulation-related traces present in each image. The diverse routing patterns of fake samples further suggest that the learned prototypes capture complementary forgery characteristics, enabling the model to represent heterogeneous manipulation artifacts.

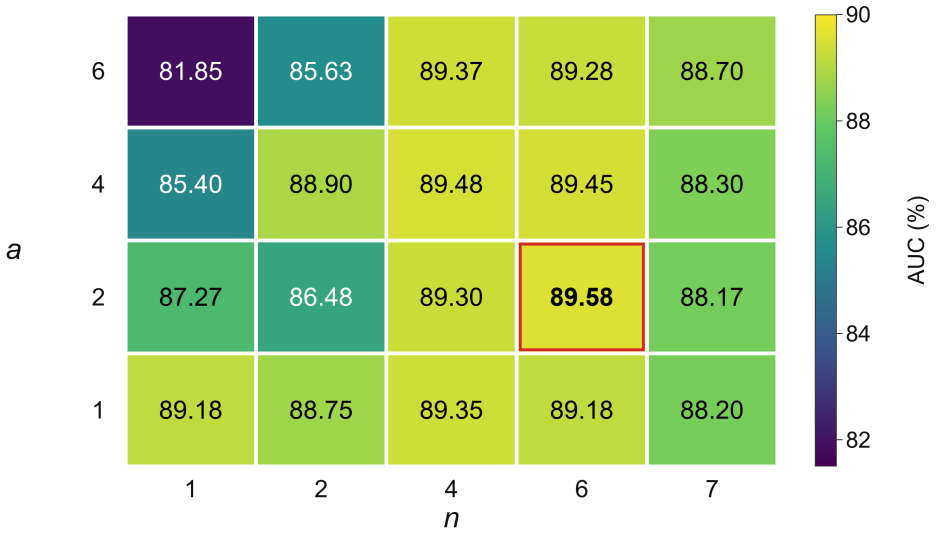


Figure 4: Ablation results for combinations of the parameters a and n.

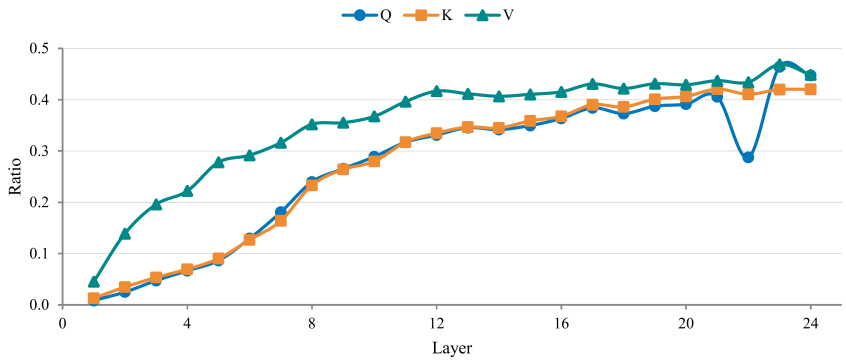


Figure 5: Layer-wise normalized effective ranks of the mean-pooled query (Q), key (K), and value (V) activations of the trained model on the FaceForensics++ test set. The effective rank is computed from the entropy of the singular-value distribution of each centered activation matrix and normalized by its maximum possible rank. Higher values indicate that the representations span a broader effective subspace across samples.

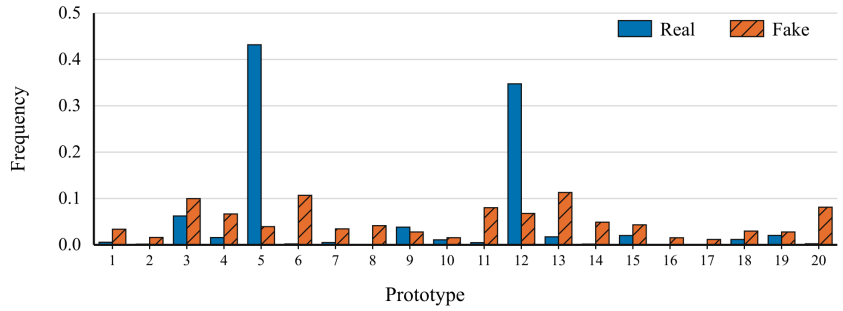


Figure 6: Prototype selection frequencies for real and fake samples on FaceForensics++