Break Visual-Linguistic Asymmetry: Unleashing VLM's Cross-Modal Potential for General Face Forgery Detection

Guilin Pang[1], Yiu-ming Cheung[1*], Ruiqi Li[1], and Weifeng Su[2]

[1] Hong Kong Baptist University, Hong Kong SAR, China
{csglpang, ymc, csrqli}@comp.hkbu.edu.hk
[2] Beijing Normal-Hong Kong Baptist University, Zhuhai, China
wfsu@bnbu.edu.cn

Abstract. Face forgery detection (FFD) is essential in preventing the misuse of diverse generation methods like generative adversarial networks (GANs) and diffusion models. Although advanced FFD methods have started to explore the benefits of Vision-Language Model (VLM), they either rely solely on visual modality or optimize visual-linguistic modalities independently, leaving researches on cross-modal interaction largely unexplored. We find that **visual-linguistic asymmetry** inherent in VLM tends to cause cross-modal misalignment, undermining its ability to discriminate and generalize on visually similar real-forged faces. This finding shows that this asymmetry is twofold: (1) cross-modal: visual-linguistic spaces affect VLM's generalizability oppositely. (2) intra-modal: spaces with different-level semantics have inverse effects on VLM's discriminability and generalizability. Building upon these insights, we propose Asy-Det, a parameter-efficient detector that follows an asymmetry-guided visual-linguistic interaction paradigm, to fully unlock the cross-modal potential of VLM for reliable FFD. With only 4.41M trainable parameters, Asy-Det achieves an image-level AUC of 92.94% on unseen Celeb-DF-v2 dataset.

Keywords: Face forgery detection · Visual-linguistic asymmetry

1 Introduction

Face manipulation techniques have rapidly advanced due to breakthroughs in generative adversarial networks (GANs) [4, 39] and diffusion models [28, 49], enabling the creation of highly photorealistic synthetic images. These advances, however, pose severe risks to personal privacy, social media security and beyond. To this end, developing robust face forgery detection (FFD) methods has become crucial, serving a vital role in preventing the misuse of face manipulation technologies. As a result, numerous detection methods have been proposed, most of which rely only on visual modality [5, 8, 12, 20, 26, 40], where a few tune visual and

* Corresponding author

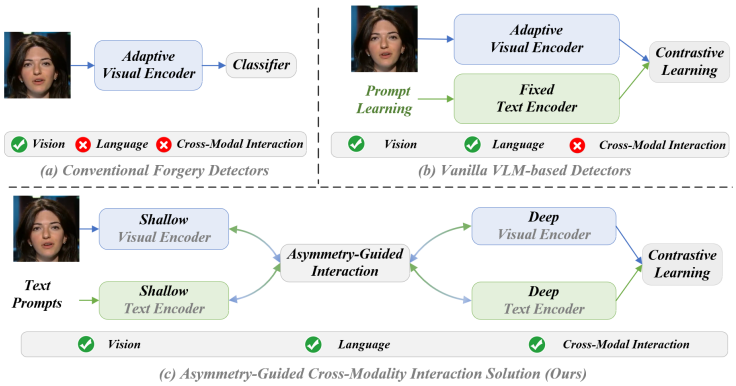

Fig. 1: Comparison with existing face forgery detectors. Current detection methods either (a) rely only on visual information or (b) tune visual and linguistic modalities independently. In contrast, our solution develops an asymmetry-guided visual-linguistic interaction mechanism to ensure cross-modal alignment for generalizable FFD.

linguistic modalities independently [18, 32, 46], as illustrated in Fig. 1. Among these, vision-based detectors predominantly focus on analyzing high-frequency signals [12, 34] or low-level visual artifacts, such as blending boundaries [5, 16] and unnatural textures [22, 38], to distinguish forged faces from genuine ones. However, it is widely acknowledged that these detectors typically face a generalization challenge, i.e., suffering from significant performance degradation when encountering forgeries (e.g., GANs or diffusion models) unseen during training [40, 43].

To address this generalization challenge, advanced solutions [21, 30, 41] have explored the benefits of pre-trained Vision-Language Model (VLM) [27] for FFD, as depicted in Fig. 1 (b). Despite the high versatility of pre-trained VLM, a huge number of their parameters make fine-tuning on the FFD task challenging, given the small scale and limited diversity of forged datasets. Thus, to adapt the pre-trained VLM, current detection methods primarily follow two paradigms: prompt learning [18, 21] and adapters [5, 30], or a combination of both. These adaptation paradigms typically optimize only the extra adapters or prompts while keeping the entire pre-trained VLM frozen to prevent overfitting, thereby achieving reasonable generalizability. Nevertheless, these VLM-based detectors tune visual and linguistic modalities independently without considering their interactive relationships (before contrastive learning), which could potentially curtail their discriminability on visually similar real-forged faces.

To this end, given the scarcity of FFD research on cross-modal interaction, a straightforward approach is to directly apply generic multi-modal interplay methods. However, an unexpected dilemma emerges: those generic methods [10, 14, 44, 45, 50], which employ symmetric and implicit interaction architectures to align cross-modal relationships of VLM, receive unsatisfactory per-

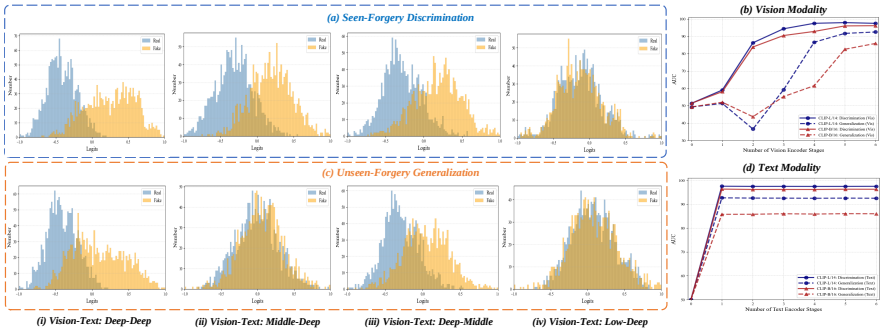

Fig. 2: The impact of visual-linguistic spaces at different depths on VLM's discrimination and generalization. VLM trained on DeepFake (FF++) are tested for generalization on FaceSwap (unseen) and for discrimination on DeepFake (seen). We freeze all VLM parameters and train only two added one-layer linear projectors placed at the end of the visual and linguistic branches, respectively. We obtained Fig. (b) by iteratively removing Transformer layers from the visual encoder while keeping the linguistic branch fixed, Fig. (d) is generated in similar way. In Fig. (a) and (c), we visualize the logit distributions at four meaningful turning points identified in Figs. (b) and (d). The distributions shown in (a-iii) and (c-iii) are from the text modality, while the other six are from the visual modality. Note that 'Low', 'Middle', and 'Deep' correspond to encoder stages 1, 3, and 6 in Fig. (b) and (d).

formance (Table 3) even lagging behind vision-based detectors. We argue that **visual-linguistic asymmetry** inherent to VLM tends to cause cross-modal misalignment in the shared latent space, undermining its discriminability and generalizability on fine-grained FFD task.

Accordingly, we systematically study the discrimination and generalization of pre-trained VLM's latent spaces, revealing that this asymmetry is twofold: (1) **cross-modal asymmetry**: visual-linguistic spaces at same depth affect VLM's generalizability oppositely. For example, from dashed lines in Fig. 2 (b) and (d), one can see that shallow visual spaces impair VLM's generalizability while shallow linguistic spaces greatly promote it, a similar trend holds for deeper spaces. (2) **intra-modal asymmetry**: spaces with different-level semantics have inverse effects on VLM's discriminability and generalizability. For instance, as shown in Fig. 2, shallow textual spaces remarkably improve VLM's discriminability and generalizability, whereas deep textual spaces provide negligible benefit. To further support our assumption, we visualize the logit distributions at four meaningful turning points in Fig. 2 (a) and (c), where a larger separation between real and forged class distributions indicates stronger discriminability and generalizability of the VLM. Building upon these observations, we derive a key insight: **visual-linguistic asymmetry** constitutes the critical bottleneck hindering VLM from fully unlocking their cross-modal potential, and generic cross-modal interaction methods [10, 14, 44, 45, 50] exacerbate this asymmetry, leading them to suffer from a discrimination-generalization dilemma in FFD.

Building upon our insights, we propose Asy-Det, a parameter-efficient detector that adopts an asymmetry-guided vision-language interaction paradigm, thereby fully unleashing the cross-modal capabilities of VLM for generalizable FFD. Asy-Det comprises of two tailored designs: the Meta Forgery Adapter (MFA) and the Asymmetry-Guided Interactor (AGI). Firstly, motivated by intra-modal asymmetry, we introduce a two-phase design that incorporates MFA in each phase to separately optimize detector's discriminability and generalizability. Specifically, MFA in the first phase learn forgery-specific representations from shallow spaces for improving discrimination while those in the last phase excite forgery-general representations in deep spaces to enhance generalization. Secondly, instead of optimizing visual and linguistic modalities independently, we take cross-modal asymmetry into account, a previously overlooked perspective. The AGI is developed to facilitate cross-modal interaction, which allows forgery representations with distinct semantics from different modalities to be mapped into a shared multi-granularity space via an asynchronous registration mechanism. Finally, AGI exploits disparities between visual and linguistic modalities within constructed multi-granularity space through explicit mutual learning, thereby fully unlocking VLM's potential in distinguishing visually similar real-forged faces. With only 4.41M trainable parameters, Asy-Det surpasses current state-of-the-art detectors, achieving an image-level AUC of 92.94% on unseen Celeb-DF-v2 dataset. Our main contributions can be summarized as follows:

- We systematically study the discrimination-generalization disparities between both modalities of VLM, revealing the devil behind ineffective cross-modal interaction, i.e., visual-linguistic asymmetry. To the best of our knowledge, this is the first FFD work to identify this asymmetry and uncover its universality across different forgeries and datasets, laying a solid groundwork for future VLM-based forgery detector evolution.
- We propose Asy-Det, a parameter-efficient detector that follows an asymmetry-guided visual-linguistic interaction paradigm, aiming to fully unlock the cross-modal potential of VLM for generalizable FFD.
- Inspired by visual-linguistic asymmetry, we first propose MFA, which activates forgery-aware features within both modalities, yielding refined representations that effectively differentiate visually similar real and forged faces. Then, we develop AGI to bridge the gap between visual and linguistic modalities, thereby enhancing VLM's capability in cross-modal forgery alignment.

2 Related Work

Face Forgery Detection. Face forgery detection (FFD) has emerged as a prominent research focus in recent years. With the continuous advancement of deep learning techniques, recent FFD methods [2,9,11,15,19,26,31,37,47,48] typically leverage Deep Neural Networks (DNNs) to detect various forgery-related signals, such as high-frequency signals and low-level visual artifacts. For instance, FreqDebias [12] introduces a consistency-driven frequency debiasing framework

that mitigates spectral bias in FFD detectors to improve cross-domain generalization and reduce reliance on dataset-specific spectral cues. However, these detectors typically perform well in within-dataset evaluations, yet often exhibit poor performance in cross-dataset evaluations. This phenomenon is known as the generalization challenge in FFD. To tackle this generalization challenge, numerous methods [5,8,18,21,25,30,32,41,46] have been proposed to adapt large-scale pre-trained Vision-Language Model (VLM) for FFD, through prompt learning, adapters, or a combination of both. For example, UDD [8] exposes position and content biases that cause spurious correlations and proposes two plug-and-play token-level interventions for the VLM visual encoder.

Nevertheless, these VLM-based detectors either solely utilize the visual modality or tune the two modalities independently, without considering the interactive relationships between visual and linguistic modalities, where sufficient sharing of intermediate features has been well-established as crucial for both discrimination and generalization [13]. This design fails to mitigate the visual-linguistic asymmetry that impairs VLMs' discriminability and generalizability for fine-grained FFD, resulting in cross-modal misalignment. Unlike previous approaches, our Asy-Det designs an asymmetry-guided visual-linguistic interaction paradigm to share more forgery-aware features in intermediate embeddings, facilitating cross-modal mutual learning for discriminative and generalizable FFD.

Multi-Modal Interaction. Given the paucity of FFD research on visual-linguistic interaction mechanisms (before contrastive learning), a straightforward solution is to directly employ generic multi-modal interplay methods (e.g., CoCoOp [50], MaPLe [14], TCP [45], MMA [44], MMRL [10]) for FFD. CoCoOp [50] introduces conditional prompt learning that generates instance-adaptive prompts conditioned on image features, enabling dynamic adaptation to different visual inputs. MaPLe [14] extends this idea by jointly learning prompts in both vision and language branches, allowing for deeper cross-modal alignment through shared prompt representations. Complementing these prompt-centric strategies, MMA [44] aligns vision and language via uni-modal projections and top-layer adapters, trading dataset-specific discriminability for improved cross-dataset generalizability in few-shot settings.

Despite these works substantially advance VLM adaptation techniques, these methods exhibit unsatisfactory performance in FFD task due to a fundamental limitation: their reliance on symmetric and implicit cross-modal interaction frameworks. In contrast, our Asy-Det follows an asymmetric and explicit interaction principle that encourages vision-language alignment, thereby fully unlocking VLM's potentials for FFD.

3 Proposed Method

3.1 Preliminary

Vanilla Vision-Language Model (VLM): Following advanced face forgery detection (FFD) methods [5, 18, 32], our detector is built upon a widely-used

pre-trained VLM, Contrastive Language-Image Pre-training (CLIP), consisting of a vision encoder V, a text encoder T, and a shared latent space $\mathcal{C}_{vt}$.

Vision Encoder: The vision encoder V comprises a pre-processing operation *PreVis*, Q Transformer layers $\{V_q\}_{q=1}^{Q}$, and a vision projection layer *ProjVis*. Given an image $x \in \mathbb{R}^{3\times H\times W}$, *PreVis* first converts it into N fixed-size patches, and then maps each patch to D-dimensional local representations $f_l^0 \in \mathbb{R}^{N\times D}$. The local patch representations are combined with a global class token f_g^0, forming the image features $[f_g^0, f_l^0] \in \mathbb{R}^{(N+1)\times D}$, which are then passed through Q Transformer layers:

$$[f_g^q, f_l^q] = V_i([f_g^{(q-1)}, f_l^{(q-1)}]), \quad q = 1, 2, \ldots, Q \tag{1}$$

After that, f_g^Q from the last Transformer layer V_Q is fed into *ProjVis* to obtain the output visual features $f_g^{'}$, which is projected into the shared latent space $\mathcal{C}_{vt}$.

Text Encoder: Similarly, the text encoder consists of an embedding operation *EmbedText*, L Transformer layers $\{T_\ell\}_{\ell=1}^{L}$, and a text projection layer *ProjText*. The category prompts are fed into *EmbedText* for tokenization and projection to generate N_t text embeddings $e_{ctx}^0 \in \mathbb{R}^{N_t\times D_t}$ of dimension D_t. Then, two boundary tokens b_{sot}^0 and b_{eot}^0, are appended at the beginning and end of e_{ctx}^0, respectively, producing input features $[b_{sot}^0, e_{ctx}^0, b_{eot}^0] \in \mathbb{R}^{(N_t+2)\times D_t}$ for L Transformer layers:

$$\begin{aligned} w_{\ell-1} &= [b_{sot}^{(\ell-1)}, e_{ctx}^{(\ell-1)}, b_{eot}^{(\ell-1)}] \\ w_\ell &= T_\ell(w_{\ell-1}), \quad \ell = 1, 2, \ldots, L \end{aligned} \tag{2}$$

After passing b_{eot}^L from the last Transformer layer T_L through *ProjText*, the output feature $b^{'}$ is projected into the shared latent space $\mathcal{C}_{vt}$.

Detection with Shared Latent Space: At the end of visual-linguistic encoders, we perform classification using the similarity of features in the shared latent space $\mathcal{C}_{vt}$, which contains the vision feature f_g' and two text features $b_c^{'}, c \in \{0, 1\}$, where $0, 1$ denote "real" and "fake", respectively. The final prediction P_c is derived by calculating cosine similarity between f_g' and $b_c^{'}$, as follows,

$$P_c = \frac{\exp\big(\cos(f_g', b_c^{'})/\tau\big)}{\sum_{c=0}^{\{0,1\}}(\exp\big(\cos(f_g', b_c')/\tau\big)} \tag{3}$$

where τ is a temperature parameter.

3.2 Visual-Linguistic Asymmetry in VLM

Discrimination and generalization are two key objectives in face forgery detection (FFD). A desirable latent space should be discriminative for visually similar forgeries and generalizable to unseen forgeries [40, 43]. However, balancing these two objectives remains challenging. For instance, in practical scenarios, adapting pre-trained VLM for FFD has emerged as a promising direction, yet it still faces

a discrimination-generalization dilemma. That is, the latent space from adapted VLM either overfits to forgery-specific clues, limiting their generalizability, or lacks sufficient separability for reliable discrimination between visually similar real-forged faces. Unfortunately, the primary focus of FFD has concentrated on the design of adaptation methods (i.e., prompt learning and adapters), leaving the investigation into the devil behind the VLM's discrimination-generalization dilemma largely unexplored.

To explore the underlying reasons behind this dilemma, we quantitatively investigate the discrimination-generalization evolution of vision-language spaces from pre-trained VLM. As illustrated in Fig. 2 and Fig. 5, our analysis reveals an asymmetric evolution of discriminability and generalizability in these spaces, which we call *vision-language asymmetry*. Specifically, this asymmetry is twofold:

Cross-modal asymmetry: Visual-linguistic spaces affect VLM's generalizability oppositely. As shown in Fig. 2 (b) and (d), *shallow-text and deep-vision* spaces learn more abstract forgery features that greatly promotes VLM's generalization, whereas *deep-text and shallow-vision* spaces are either harmful or useless to that generalization. Evidently, in VLM, latent spaces from different modalities exhibit inverse generalization evolution processes. This explains the failure of symmetric interaction structures in existing methods [10, 14, 44, 45] and underscores the necessity of asymmetric cross-modal interaction strategies for visual-linguistic alignment.

Intra-modal asymmetry: Latent spaces encoding semantics at different depths exert opposing effects on VLM's discriminability and generalizability. Specifically, as depicted in Fig. 5, shallow textual spaces remarkably improve VLM's discriminability and generalizability, whereas deep textual spaces provide negligible benefit. This intra-modal asymmetry suggests that improvements in generalization typically come at the cost of discrimination (Fig. 5 (a)), and vice versa. To address this, a tailored two-phase adaptation paradigm is essential to separately promote the detector's discriminability on visually similar forgeries and then enhances its generalizability for unseen forgeries.

3.3 Overview

Motivated by our asymmetric insights presented in Sec. 3.2, we propose an asymmetry-guided detector, termed *Asy-Det*, to fully unleash VLM's cross-modal potential for FFD. As illustrated in Fig. 3, Asy-Det works in a two-phase adaptation fashion and incorporates two tailored designs: Meta-Forgery Adapter (MFA) and Asymmetric-Guided Interactor (AGI). To address intra-modal asymmetry, Asy-Det first identically partitions the vision and language encoders into two equal phases and instantiates MFA into each phase. The first-phase extracts forgery-specific features from shallow spaces to boost VLM's discriminability, while the second-phase learns forgery-general embeddings to enhance its generalizability. Specifically, MFA introduces two parallel branches, i.e., forgery adaptation branch and meta generation branch, where the former excites forgery-aware features at different depths within modalities while the latter propagates these multi-granularity features into an asynchronous registration mechanism.

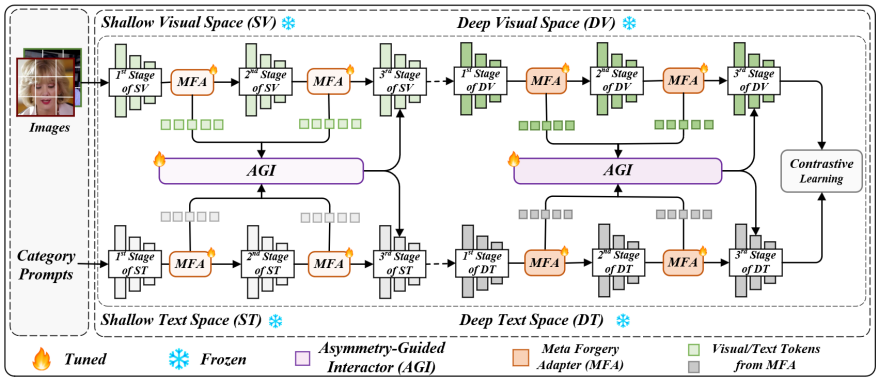

Fig. 3: The overview of the proposed Asy-Det, which comprises a Meta Forgery Adapter (MFA) and an Asymmetry-Guided Interactor (AGI). Notably, to prevent catastrophic alterations of VLM's original visual-language space, only the proposed MFA and AGI are fine-tuned during training, with all parameters from pre-trained VLM kept frozen throughout.

To tackle cross-modal asymmetry, our AGI first integrates multi-granularity features from an asynchronous registration mechanism into a shared and unified latent space. With this latent space, AGI utilizes an explicit mutual-learning strategy to bridge the discrimination-generalization gap between visual and linguistic modalities, fully unlocking the cross-modal potential of VLM in detecting various fake faces generated by diverse forgeries.

3.4 Meta Forgery Adapter

To adapt the pre-trained knowledge from VLM for efficient face forgery discovery, we incorporate MFA into both vision and text modalities to bridge adjacent stages, each contains multiple Transformer layers, as illustrated in Fig. 3. MFA comprises two parallel branches: the forgery adaptation branch discovers forgery-related cues within modalities, while the meta generation branch generates modality-specific tokens for constructing multi-granularity shared spaces at shallow and deep layers of the VLM.

Forgery Adaptation Branch: Taking the shallow vision spaces as an example, the output features $[f_g^q, f_l^q]$ from the q-th Transformer layer V_q (where $q < \frac{Q}{2}$) in Eq. (1) are fed into the forgery adaptation branch, which passes through a lightweight adapter $\mathcal{FA}$ that comprises grouped fully-connected layers and a ReLU activation function to excite forgery-related features:

$$[\hat{f}_g^q, \hat{f}_l^q] = [f_g^q, f_l^q] + \alpha \cdot \mathcal{FA}(V_q([f_g^q, f_l^q])) \tag{4}$$

where α is a learnable scale parameter that balances pre-trained knowledge and forgery-related features in the newly generated features $[\hat{f}_g^q, \hat{f}_l^q]$. By introducing this controllable fusion strategy, the pre-trained knowledge will not be catastrophically modified due to the introduction of forgery-related features, thereby

enhancing the vision-language model's capability in discriminating visually similar real-forged faces.

Meta Generation Branch: Unlike vanilla VLM-based detectors that build a shared space $\mathcal{C}_{vt}$ only at the end of visual-linguistic encoders, meta generation branch constructs two more multi-granularity shared spaces, $\{\mathcal{C}_{vt}^s\}_{s=1}^2$, corresponding to the first and second phases of the encoders. Taking $\mathcal{C}_{vt}^1$ as an example, given that the local patch tokens $f_l^q \in \mathbb{R}^{N\times D}$ contain rich fine-grained forgery cues, we employ them to generate shared vision tokens $f_{ssv}^q \in \mathbb{R}^{N\times D_s}$. Specifically, meta branch begins by discovering and preserving fine-grained forgery cues within the input features through convolution operation $Conv$ and asynchronous register Π:

$$f_{ssv}^q = \Pi(ReLU(Conv(f_l^q))), \quad q \in \{Q/6, Q/3\} \tag{5}$$

A similar process is applied to text encoder to generate shared linguistic token $w_{sst}^q \in \mathbb{R}^{(N_t+2)\times D_s}$:

$$w_{sst}^\ell = \Pi(ReLU(Conv(w_\ell))), \quad \ell \in \{L/6, L/3\} \tag{6}$$

With f_{ssv}^q and w_{sst}^ℓ, we construct shallow shared space $\mathcal{C}_{vt}^1$, containing forgery-related features across four distinct granularities for subsequent asymmetric interaction across visual and linguistic modalities.

3.5 Asymmetry-Guided Interactor

While we effectively activate intra-modal forgery-relevant features through the designed MFA, the visual and linguistic modalities undergo independent fine-tuning without considering their cross-modal correlations. Our insight regarding cross-modal asymmetry, visual and linguistic modalities exhibit drastically different discrimination-generalization evolution processes, shows that train both modalities independently make VLM difficult to bridge significant disparities between vision-language spaces. To align vision-language spaces, unlike existing detectors, we develop an *Asymmetric-Guided Interactor* (AGI) to identify the cross-modal relationships in multi-granularity shared spaces $\{\mathcal{C}_{vt}^s\}_{s=1}^2$ constructed by MFA.

Take $\mathcal{C}_{vt}^1$ as an example, the final outputs of visual forgery features f_{ssv} and text prompt embeddings w_{sst} are obtained by averaging intra-modal tokens with distinct granularities:

$$f_{ssv} = AVG(f_{ssv}^{Q/6}, f_{ssv}^{Q/3}),\ w_{sst} = AVG(w_{sst}^{L/6}, w_{sst}^{L/3}) \tag{7}$$

Then, an interaction block that integrates bidirectional cross-attention units serves as a bridge and allows forgery-aware features to be transmitted into each other, ensuring mutual learning of vision and text modalities. Specifically, we first compute cross-attention matrices $Attn_{t2v} \in \mathbb{R}^{(N_t+2)\times N}$ from text to vision, where w_{sst} are conditioned on f_{ssv}, and then compute another cross-attention matrices $Attn_{v2t} \in \mathbb{R}^{N\times(N_t+2)}$ from vision to text similarly:

$$Attn_{v2t} = w_{sst} \cdot (f_{ssv})^T,\ Attn_{t2v} = f_{ssv} \cdot (w_{sst})^T \tag{8}$$

With the bidirectional attention matrices $Attn_{t2v}$ and $Attn_{v2t}$, we align f_{ssv} with w_{sst} in $\mathcal{C}_{vt}^{1}$, as follows,

$$\hat{f}_{ssv} = \varphi(Attn_{v2t}) \cdot w_{sst} + f_{ssv}, \ \hat{w}_{sst} = \varphi(Attn_{t2v}) \cdot f_{ssv} + w_{sst} \tag{9}$$

where $\hat{f}_{ssv}$ and $\hat{w}_{sst}$ are the aligned vision forgery features and text prompt embeddings, respectively, and φ is the softmax. To incorporate aligned cross-modal context while preserving pre-trained knowledge, $\hat{f}_{ssv}$ and $\hat{w}_{sst}$ are projected back to the original modality-specific spaces through a residual structure:

$$\hat{f}_{l}^{q} = f_{l}^{q} + M_v \cdot \hat{f}_{ssv}, \quad \hat{w}^{\ell} = w^{\ell} + M_t \cdot \hat{w}_{sst}, \quad q = Q/2, \ell = L/2 \tag{10}$$

where $M_v \in \mathbb{R}^{D_s \times D_v}$ and $M_t \in \mathbb{R}^{D_s \times D_t}$ are learnable weight matrices. Combined with the asymmetric cross-modal interaction paradigm, vision and text encoders can guide each other focus on forgery-aware representations, thereby achieving effective discrimination-generalization.

Previous approaches utilize a ***symmetric and implicit*** interaction framework, which not only induces insufficient cross-modal information propagation but also aggravates vision-language misalignment, resulting in suboptimal VLM discriminability and generalizability. Different from these approaches, our AsyDet follows an ***asymmetric and explicit*** interaction principle that encourages mutual learning across modalities. Our solution boasts two pivotal advantages: (1) an explicit interaction mechanism that ensures sufficient forgery-aware information propagation across modalities; (2) an asymmetric paradigm that reconciles disparities between vision-language spaces, thereby fully unlocking the discriminative and generalizable potentials of VLM for FFD.

3.6 Analysis of Optimal Depth Pair

We analyze the optimal choice of the feature depth of both modalities, where the optimal depth pair (q, ℓ) is defined as $(q^*, \ell^*) = \arg\min_{(q,\ell)} \mathbb{E}_{(x,y)\sim\mathcal{D}} \mathcal{L}$, where $\mathcal{L}$ is the classification loss, (x, y) is the image and label pair, $\mathcal{D}$ is the dataset distribution. The optimal depth pair minimizes the generalization error of our face forgery detection task. We theoretically characterize the relationship between the optimal depth pair (q^*, ℓ^*) and the generalization bound of the resulting VLM-based detector. Specifically, we show that the choice of layer depth induces a Rademacher-complexity term that upper-bounds the generalization gap, and that under a complexity budget, the optimal depth configuration is in general asymmetric (i.e., $q^* \neq \ell^*$), consistent with our observed visual-linguistic asymmetry. The full definitions, lemmas, propositions and proofs are deferred to the supplementary material.

4 Experiments

4.1 Experimental Setup

Following prior work [8, 12, 18, 40], we employ standard face forgery detection benchmarks for comprehensive evaluation, including FaceForensics++ [29], DeepFakeDetection [7], Celeb-DF-v2 [17], DFDC-Large [6], Wilddeepfake [51] and

Table 1: Cross-dataset evaluations using the ***IMAGE-LEVEL*** AUC metric on multiple face forgery benchmarks. All detectors are trained on FF++_c23 and evaluated on other datasets. The best results are highlighted in bold and the second is underlined.

Method	Detector	Venue	DFD	CDF-v2	DFDC-L	Avg
Naive	EfficientB4 [36]	ICML 19	0.815	0.749	0.696	0.754
Naive	CLIP-B16 [27]	ICML 21	0.841	0.762	0.722	0.775
Frequency	F^3-Net [26]	ECCV 20	0.798	0.735	0.702	0.745
Frequency	SPSL [20]	CVPR 21	0.812	0.765	0.704	0.760
Frequency	F^2Trans-B [23]	TIFS 23	—	0.831	0.718	—
Frequency	FreqDebias [12]	CVPR 25	0.868	0.836	0.741	0.745
Spatial	ProDet [2]	NIPS 24	—	0.845	0.724	—
Spatial	LSDA [40]	CVPR 24	0.880	0.830	0.736	0.815
Spatial	UDD [8]	AAAI 25	<u>0.910</u>	<u>0.869</u>	0.758	<u>0.846</u>
Spatial	SSG [18]	AAAI 25	—	0.800	<u>0.773</u>	—
Spatial	IDCNet [38]	TIFS 25	0.847	0.809	0.724	0.793
Spatial	ED^4 [3]	TIP 25	0.877	0.839	0.743	0.820
Spatial	Ours	—	**0.938 (↑2.80%)**	**0.929 (↑6.00%)**	**0.835 (↑6.20%)**	**0.901 (↑5.50%)**

DF40 [42], to assess model robustness, cross-forgery and cross-dataset generalization. More experimental setup (i.e., dataset and implementation) details can be found in the supplementary material.

4.2 Comparison with Existing Detectors

Cross-Dataset Generalization Evaluation. The results in Table 1 validate the superiority of our proposed Asy-Det. Our method dominates across all test scenarios, achieving 92.94% AUC on CDF-v2 with a substantial 6.00% margin over UDD [8], the runner-up detector. This result highlights a fundamental limitation in existing methods like UDD [8]: while they extract spatial features effectively, they lack mechanisms to handle visual-linguistic asymmetry through cross-modal interaction. Compared to frequency-based approaches (SPSL [20], F^3-Net [26]), our method achieves moderate success on DFD while demonstrating significantly improved performance on CDF-v2 and DFDC-L. These methods depend on frequency-domain artifacts that prove brittle when confronting diverse forgery types and post-processing variations. Our Asy-Det, conversely, maintains consistent excellence with 5.50% average AUC gains across all benchmarks. This strong cross-dataset generalization stems from our asymmetry-guided interaction framework, which aligns visual and linguistic modalities while exploiting their complementary strengths that fully realizing VLM's discriminative potential for detecting realistic forgeries.

Cross-Manipulation Generalization Evaluation. The cross-forgery generalization results are shown in Table 2. Asy-Det establishes new state-of-the-art performance in nearly all cross-manipulation scenarios. The most dramatic improvement occurs when training on FSh and testing on DF, where we exceed SSG [18] by 1.64% AUC. Additionally, an intriguing pattern emerges when com-

Table 2: Cross-manipulation performance comparison using ***IMAGE-LEVEL*** AUC. Notably, 'DF' in the first row indicates the detector is trained on Deepfakes; the entries below show its performance when tested on three other manipulation methods.

Detector	Deepfakes (DF)			FaceSwap (FS)			FaceShifter (FSh)			Avg
	DF	FS	FSh	DF	FS	FSh	DF	FS	FSh	
DCL [33]	**99.98**	61.01	68.45	74.80	**99.90**	64.86	63.98	58.43	99.49	76.77
IID [11]	99.51	63.83	73.49	75.39	99.73	66.18	65.42	59.50	99.50	78.06
SSG [18]	99.36	94.94	81.51	98.31	99.59	85.21	89.99	81.22	**99.82**	92.22
Asy-Det	99.25	**95.47**	**88.86**	**98.66**	99.79	**88.77**	**91.63**	**84.71**	99.08	**94.02**

Table 3: Cross-dataset comparison with generic multi-modal interaction approaches using the ***IMAGE-LEVEL*** AUC metric on three face forgery benchmarks.

Method	DFD	CDF-v2	DFDC-L	Avg
CoCoOp [50]	0.828	0.757	0.702	0.762
MaPLe [14]	0.898	0.798	0.751	0.816
MMA [44]	0.887	0.854	0.773	0.838
TCP [45]	0.841	0.762	0.722	0.775
MMRL [10]	0.871	0.827	0.728	0.809
Ours	**0.938** (↑4.00%)	**0.929** (↑7.50%)	**0.835** (↑6.20%)	**0.901** (↑6.30%)

paring within-manipulation versus cross-manipulation performance. Detectors like DCL [33] achieve near-perfect results on their training manipulation (e.g., 99.98% on DF when trained on DF) yet suffer dramatic drops when tested on unfamiliar methods. This suggests DCL [33] memorizes method-specific features rather than learning generalizable forgery concepts. However, our Asy-Det breaks this pattern, delivering strong cross-manipulation transfer while remaining competitive on within-manipulation tasks. A potential reason is that our detector enables bidirectional knowledge flow between visual and linguistic streams that learns forgery-invariant representations spanning multiple generation techniques.

Comparison with Generic Multi-Modal Interaction Methods. We compare our method with generic multi-modal interaction techniques ([10,14,44, 45,50]) by directly applying them to face forgery detection. The results in Table 3 shows that these approaches underperform our method. This empirical evidence validates our central thesis that symmetric interaction paradigms inadequately address VLM's inherent visual-linguistic asymmetry in the forgery detection context. Our Asy-Det substantially outpaces MMA [44], the strongest generic baseline, by 6.30% on average gain. These margins underscore a key insight: recognizing and compensating for visual-linguistic asymmetry is essential for harnessing VLM's full cross-modal capabilities in distinguishing subtle facial manipulations.

Robustness Evaluation. Real-world forgery detection confronts corrupted inputs from various perturbations. As shown in Fig. 4, we test the robustness of our method by introducing four types common perturbations (Gaussian noise,

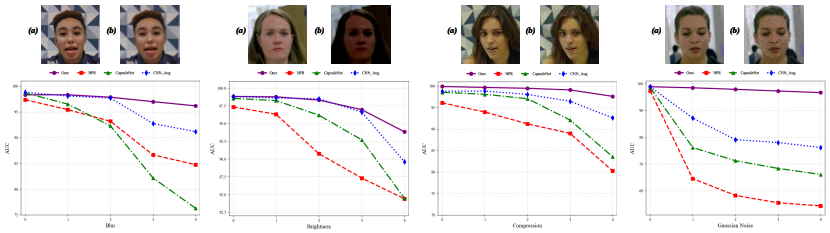

Fig. 4: Robustness evaluations. We report AUCs for Asy-Det (purple) and three SOTA detectors (NPR [35] (red), CapsuleNet [24] (green), CNN_Aug [1] (blue)) across four unseen corruption types at five severity levels (level 4 = most severe).

Table 4: Ablation studies regarding the effectiveness of the proposed modules. Notably, "T-Param" denotes the total number of trainable parameters. All models are trained on the FF++-C_{23} dataset and evaluated across various other datasets with metrics presented in the order of ***IMAGE-LEVEL*** AUC, AP, and EER.

Model	MFA	AGI	DFD			CDF-v2			DFDC-L			T-Param
			AUC	AP	EER (↓)	AUC	AP	EER (↓)	AUC	AP	EER (↓)	
M_1	×	×	80.36	97.38	26.17	75.03	85.14	33.14	72.16	72.60	33.69	1.38 M
M_2	✓	×	92.69	99.12	13.77	87.94	93.76	20.79	80.67	84.15	26.78	2.76 M
M_3	×	✓	91.24	98.94	15.98	87.92	93.45	20.79	80.45	83.38	26.57	1.64 M
M_4	✓	✓	**93.85**	**99.25**	**12.95**	**92.94**	**96.43**	**15.16**	**83.60**	**86.50**	**24.79**	**4.41 M**

motion blur, compression, brightness) at five severity levels (0=clean, 4=extreme). While most detectors maintain performance on clean data, they degraded under corruption: at level 4 Gaussian noise, NPR [35] plummet below 65% AUC. Our Asy-Det exhibits superior resilience against corruption, retaining approximately 97% AUC under identical severe noise conditions. This robustness arises from our asymmetry-guided framework extracting complementary forgery features from both visual textures and linguistic semantics.

4.3 Ablation Study

The Effect of Each Component. We analyze Asy-Det's architecture through four progressive variants in Table 4. Our baseline M_1 leverages category prompts on the pre-trained CLIP-Large, freezing all parameters bar two projection layers and treating each modality independently. M_1 struggles across all test sets, achieving only 80.36% AUC on DFD and 75.03% on CDF-v2. This weakness confirms that ignoring cross-modal dynamics leaves visual and linguistic representations misaligned, weakening the detector's ability to distinguish real from visually similar forged faces. Conversely, M_3 integrates only Asymmetry-Guided Interactor (AGI) with the baseline. AGI improves CDF-v2 to 87.92% AUC by enforcing cross-modal alignment through bidirectional knowledge transfer. This validates AGI's role in compensating for cross-modal asymmetry. The complete M_4 configuration includes both MFA and AGI, achieving 92.94% on CDF-v2

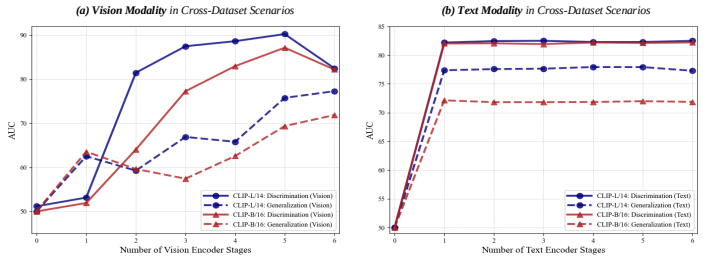

Fig. 5: The impact of visual-linguistic spaces at different depths on VLM's intra-dataset discrimination and cross-dataset generalization. VLM trained on FF++ are tested for generalization on Celeb-DF-v2 and for discrimination on FF++.

(17.91% over baseline) and consistent excellence across benchmarks. This synergy demonstrates that intra-modal and cross-modal asymmetry corrections reinforce each other. Remarkably, M_4 adds merely 4.41M trainable parameters, and maintains practical efficiency while delivering state-of-the-art detection accuracy.

Universality of Visual-Linguistic Asymmetry. We further investigate the universality of visual-linguistic asymmetry generalize under cross-dataset scenario. As depicted in Fig. 5, the asymmetry patterns appear: shallow visual layers harm cross-dataset generalization while shallow linguistic layers enhance it. This cross-dataset consistency confirms asymmetry is not an artifact of a particular dataset's characteristics but an intrinsic property of VLM applied to forgery detection. For the "Intra-Modal" case, the asymmetry appears similarly. These reproducible observations across diverse test distributions validate our fundamental assumption: any VLM-based forgery detector must explicitly resolve the visual-linguistic asymmetry to achieve robust generalization.

5 Conclusion

In this paper, we have addressed a fundamental yet long-neglected challenge in face forgery detection: the **visual-linguistic asymmetry** inherent in VLM that undermines their cross-modal potential for FFD. Through systematic investigation, we have revealed that this asymmetry is twofold: (1) *cross-modal asymmetry*, where visual and linguistic spaces exert opposite effects on VLM's generalizability; and (2) *intra-modal asymmetry*, where spaces encoding different-level semantics have inverse impacts on VLM's discriminability and generalizability. These findings elaborate why the existing VLM-based detectors, which either optimize modalities independently or employ symmetric interaction mechanisms, struggle to achieve a balanced trade-off between discrimination and generalization. Accordingly, we have proposed Asy-Det, which comprises a MFA to mitigate intra-modal asymmetry and an AGI to alleviate cross-modal asymmetry.

With only 4.41M trainable parameters, Asy-Det achieves state-of-the-art performance across multiple benchmarks, including 92.94% AUC on Celeb-DF-v2.

Acknowledgements

This work was supported by the RGC Senior Research Fellow Scheme under Grant SRFS2324-2S02, and in part by RGC General Research Fund under Grant 12202924, the Guangdong Provincial Key Laboratory of IRADS under Grant 2022B1212010006, and Guangdong Higher Education Upgrading Plan (2021–2025).

References

1. Ba, Z., Liu, Q., Liu, Z., Wu, S., Lin, F., Lu, L., Ren, K.: Exposing the deception: Uncovering more forgery clues for deepfake detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 719–728 (2024) 13
2. Cheng, J., Yan, Z., Zhang, Y., Luo, Y., Wang, Z., Li, C.: Can we leave deepfake data behind in training deepfake detector? Advances in Neural Information Processing Systems **37**, 21979–21998 (2024) 4, 11
3. Cheng, J., Zhang, Y., Zou, Q., Yan, Z., Liang, C., Wang, Z., Li, C.: Ed ^4: Explicit data-level debiasing for deepfake detection. IEEE Transactions on Image Processing (2025) 11
4. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8188–8197 (2020) 1
5. Cui, X., Li, Y., Luo, A., Zhou, J., Dong, J.: Forensics adapter: Adapting clip for generalizable face forgery detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19207–19217 (2025) 1, 2, 5
6. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020) 10
7. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C.: The deepfake detection challenge (dfdc) dataset. In: arXiv preprint arXiv:2006.07397 (2020) 10
8. Fu, X., Yan, Z., Yao, T., Chen, S., Li, X.: Exploring unbiased deepfake detection via token-level shuffling and mixing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3040–3048 (2025) 1, 5, 10, 11
9. Guo, X., Song, X., Zhang, Y., Liu, X., Liu, X.: Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 105–116 (2025) 4
10. Guo, Y., Gu, X.: Mmrl: Multi-modal representation learning for vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25015–25025 (2025) 2, 3, 5, 7, 12
11. Huang, B., Wang, Z., Yang, J., Ai, J., Zou, Q., Wang, Q., Ye, D.: Implicit identity driven deepfake face swapping detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4490–4499 (2023) 4, 12

12. Kashiani, H., Talemi, N.A., Afghah, F.: Freqdebias: towards generalizable deepfake detection via consistency-driven frequency debiasing. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8775–8785. IEEE (2025) 1, 2, 4, 10, 11
13. Kempf, E., Schrodi, S., Argus, M., Brox, T.: When and how does clip enable domain and compositional generalization? arXiv preprint arXiv:2502.09507 (2025) 5
14. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023) 2, 3, 5, 7, 12
15. Kong, C., Luo, A., Bao, P., Yu, Y., Li, H., Zheng, Z., Wang, S., Kot, A.C.: Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. IEEE Transactions on Dependable and Secure Computing (2025) 4
16. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5001–5010 (2020) 2
17. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df (v2): a new dataset for deepfake forensics. arXiv preprint arXiv (2019) 10
18. Lin, K., Lin, Y., Li, W., Yao, T., Li, B.: Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5262–5270 (2025) 2, 5, 10, 11, 12
19. Lin, Y., Song, W., Li, B., Li, Y., Ni, J., Chen, H., Li, Q.: 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) 4
20. Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., Yu, N.: Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 772–781 (2021) 1, 11
21. Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., Zhao, Y.: Forgery-aware adaptive transformer for generalizable synthetic image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10770–10780 (2024) 2, 5
22. Liu, J., Xie, J., Wang, Y., Zha, Z.J.: Adaptive texture and spectrum clue mining for generalizable face forgery detection. IEEE Transactions on Information Forensics and Security **19**, 1922–1934 (2024). https://doi.org/10.1109/TIFS.2023.3344293 2
23. Miao, C., Tan, Z., Chu, Q., Liu, H., Hu, H., Yu, N.: F 2 trans: High-frequency fine-grained transformer for face forgery detection. IEEE Transactions on Information Forensics and Security **18**, 1039–1051 (2023) 11
24. Nguyen, H.H., Yamagishi, J., Echizen, I.: Capsule-forensics: Using capsule networks to detect forged images and videos. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2307–2311. IEEE (2019) 13
25. Pang, G., Zhang, B., Teng, Z., Qi, Z., Fan, J.: Mre-net: Multi-rate excitation network for deepfake video detection. IEEE Transactions on Circuits and Systems for Video Technology **33**(8), 3663–3676 (2023) 5
26. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European Conference on Computer Vision. pp. 86–103. Springer (2020) 1, 4, 11

27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PmLR (2021) 2, 11
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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) 1
29. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11 (2019) 10
30. Shao, R., Wu, T., Nie, L., Liu, Z.: Deepfake-adapter: Dual-level adapter for deepfake detection. International Journal of Computer Vision **133**(6), 3613–3628 (2025) 2, 5
31. Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18720–18729 (2022) 4
32. Sun, K., Chen, S., Yao, T., Zhou, Z., Ji, J., Sun, X., Lin, C.W., Ji, R.: Towards general visual-linguistic face forgery detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19576–19586 (2025) 2, 5
33. Sun, K., Yao, T., Chen, S., Ding, S., Li, J., Ji, R.: Dual contrastive learning for general face forgery detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2316–2324 (2022) 12
34. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5052–5060 (2024) 2
35. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28130–28139 (2024) 13
36. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019) 11
37. Tian, J., Yu, C., Wang, X., Chen, P., Xiao, Z., Dai, J., Han, J., Chai, Y.: Real appearance modeling for more general deepfake detection. In: European Conference on Computer Vision. pp. 402–419. Springer (2024) 4
38. Wang, Z., Chen, Y., Yao, Y., Han, M., Xing, W., Li, M.: Idcnet: Image decomposition and cross-view distillation for generalizable deepfake detection. IEEE Transactions on Information Forensics and Security (2025) 2, 11
39. Xia, W., Yang, Y., Xue, J.H., Wu, B.: Tedigan: Text-guided diverse face image generation and manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2256–2265 (2021) 1
40. Yan, Z., Luo, Y., Lyu, S., Liu, Q., Wu, B.: 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 (2024) 1, 2, 6, 10, 11
41. Yan, Z., Wang, J., Wang, Z., Jin, P., Zhang, K.Y., Chen, S., Yao, T., Ding, S., Wu, B., Yuan, L.: Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. arXiv preprint arXiv:2411.15633 **2** (2024) 2, 5

42. Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., Luo, D., Wang, C., Ding, S., Wu, Y., et al.: Df40: Toward next-generation deepfake detection. Advances in Neural Information Processing Systems **37**, 29387–29434 (2024) 11
43. Yan, Z., Zhang, Y., Fan, Y., Wu, B.: Ucf: Uncovering common features for generalizable deepfake detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22412–22423 (2023) 2, 6
44. Yang, L., Zhang, R.Y., Wang, Y., Xie, X.: Mma: Multi-modal adapter for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23826–23837 (2024) 2, 3, 5, 7, 12
45. Yao, H., Zhang, R., Xu, C.: Tcp: Textual-based class-aware prompt tuning for visual-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23438–23448 (2024) 2, 3, 5, 7, 12
46. Yu, P., Fei, J., Gao, H., Feng, X., Xia, Z., Chang, C.H.: Unlocking the capabilities of vision-language models for generalizable and explainable deepfake detection. arXiv e-prints pp. arXiv–2503 (2025) 2, 5
47. Yu, Y., Ni, R., Yang, S., Zhao, Y., Kot, A.C.: Narrowing domain gaps with bridging samples for generalized face forgery detection. IEEE Transactions on Multimedia **26**, 3405–3417 (2024). https://doi.org/10.1109/TMM.2023.3310341 4
48. Zhang, T., Yu, P., Xia, Z., Dai, L., Zhou, X., Gao, H.: Fine-grained dino tuning with dual supervision for face forgery detection. arXiv preprint arXiv:2511.12107 (2025) 4
49. Zhao, W., Rao, Y., Shi, W., Liu, Z., Zhou, J., Lu, J.: Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8568–8577 (2023) 1
50. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022) 2, 3, 5, 12
51. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: Wilddeepfake: A challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2382–2390 (2020) 10