Appendix: Unleashing Vision-Language Semantics for Deepfake Video Detection

Jiawen Zhu[1] **Yunqi Miao**[2] **Xueyi Zhang**[3] **Jiankang Deng**[4*] **Guansong Pang**[1*]
[1]Singapore Management University, Singapore
[2]The University of Warwick, UK
[3]Nanyang Technological University, Singapore
[4]Imperial College London, UK

Table 1. Data statistics of face-swapping forgery datasets.

Dataset	Sythesis Methods	Real	Fake	Total
FaceForensics++	4	1000	4000	5000
CelebDF v1	1	408	795	1203
CelebDF v2	1	590	5634	6229
DFDC	8	23654	104500	128154
DFD	5	363	3000	3363

Table 2. Data statistics of full-face synthesized data generated based on GAN and Diffusion-based methods.

Dataset	Synthesis Type	Real	Fake	Total
VQGAN	GAN based	590	5634	6229
StyleGAN-XL	GAN based	590	5634	6229
SiT-XL/2	Latent Diffusion	590	5634	6229
DiT	Latent Diffusion	590	5634	6229
PixArt	Latent Diffusion	590	5634	6229

A. Dataset Details

A.1. Data Statistics of Training and Testing

We evaluate our method on five classical face-swapping forgery datasets: FaceForensics++ (FF++) [13], CelebDF v1/v2 [8] (CDF-v1/v2), Deepfake Detection Challenge (DFDC) [6], and DeepfakeDetection (DFD) [1]; as well as full-face synthesized data based on CDF-v2 sourced from the large-scale DF40 dataset [16], where we select five representative GAN and Diffusion-based generative models: VQGAN [7], StyleGAN-XL (StyleGAN) [14], SiT-XL/2 (SiT) [2], DiT [12], and PixArt [3].

To assess the generalization ability, we follow the common cross-dataset evaluation protocol by training the model on the c23-compression version of FF++ and evaluating it on the remaining datasets. Table 1 provides the data statistics of classical face-swapping forgery datasets, while Table 2 shows the full-face synthesized datasets generated by GAN and Diffusion-based generative models.

A.2. Classical Face-Swapping Forgery Datasets

FaceForensics++ (FF++) [13]. FF++ is a widely used benchmark for facial manipulation detection, containing over 1,000 real videos and their corresponding manipulated versions generated with four representative face-swapping and reenactment techniques: DeepFakes, Face2Face, FaceSwap, and NeuralTextures. Multiple compression levels are also provided to simulate real-world media quality.

CelebDF v1/v2 (CDF-v1/v2) [8]. Celeb-DF is a large-scale deepfake video dataset constructed using YouTube celebrity videos and high-quality swapping methods designed to reduce visual artifacts. Version v2 significantly improves visual realism compared to v1, making it more challenging for detection models.

Deepfake Detection Challenge (DFDC) [6]. DFD contains high-quality deepfake videos created with professional actors under controlled conditions. Compared with FF++ and DFDC, DFD offers cleaner visual quality and fewer compression artifacts, providing an ideal benchmark for evaluating fine-grained detection capability.

DeepfakeDetection (DFD) [1]. This dataset is designed for deepfake detection tasks, providing a comprehensive collection of video sequences that can be used to train and evaluate deep learning models for identifying manipulated media. It was downloaded from the official FaceForensics server, which offers high-quality datasets specifically for the purpose of face manipulation detection.

A.3. Full-Face Synthesized Datasets

With the progress of AIGC techniques, full-face synthesis has achieved high perceptual realism without typical blending artifacts found in face-swapping methods. We evaluate on five representative GAN- and diffusion-based generators from DF40 [16], covering different generative families and image priors. DF40 includes fully generated subsets derived from both CelebDF-v2 and FF++. Because

*Corresponding authors: J. Deng (j.deng16@imperial.ac.uk) and G. Pang (gspang@smu.edu.sg)

our model is trained on FF++, we adopt the subset generated from CelebDF-v2 for cross-dataset evaluation to ensure non-overlapping identities and generation patterns.

VQGAN [7]. A GAN-based discrete latent-space generator capable of producing high-resolution images with improved perceptual quality. It synthesizes globally coherent facial structures without explicit patch-level inconsistencies.

StyleGAN-XL (StyleGAN) [14]. An improved variant of StyleGAN capable of scaling to diverse large-scale datasets with strong identity realism, making generated faces more diverse and visually convincing.

SiT-XL/2 (SiT) [2]. A diffusion-based generator that leverages scalable transformer architectures for high-fidelity face synthesis, producing smooth textures and natural facial layouts.

DiT [12]. A transformer-based diffusion model that operates directly in latent space. It provides high-quality generative realism with minimal local artifacts, further increasing detection difficulty.

PixArt [3]. A recent high-resolution text-to-image generator demonstrating strong semantic alignment and photorealism. The produced faces lack typical low-level cues, posing challenges to artifact-based detectors.

B. Implementation Details

B.1. Details of Model Configuration

We implement `VLAForge` using OpenCLIP with the publicly available *ViT-L/14* backbone. The parameters of both the visual and text encoders in CLIP are kept frozen throughout all experiments. The ForgePerceiver follows the *vit_tiny_patch16_224* configuration, with all parameters randomly initialized and fully trained from scratch during optimization. The numbers of forgery query tokens and replicated class-token embeddings q are set to 128 by default. The number of fusion weight α is set to 0.5. We adopt the Adam optimizer with an initial learning rate of 2e-5 and a weight decay of 5e-4 to update model parameters. The input images are resized to 224×224, and the batch size is set to 32. To ensure that the model learns to recognize both real and fake faces across diverse identities while mitigating overfitting, training is conducted for 15 epochs on a single NVIDIA GeForce RTX 3090 GPU. We will release the code upon publication to facilitate reproducibility.

B.2. Implementation of Comparison Methods

ForAda [4]. ForAda enhances CLIP for face forgery detection by introducing a task-specific adapter that learns forgery-related visual traces and interacts with CLIP's visual tokens while preserving its inherent generalization capability. The results on classical face-swapping datasets are taken from the original paper, whereas the results on full-face generation datasets are reproduced using the official implementation[1].

RepDFD [9]. Since RepDFD does not provide official code, we reproduce the method based on the implementation details described in the paper. We adopt CLIP-ViT-L/14 pretrained on LAION-400M as the foundation model, with an input resolution of 224 × 224 and an input transformation parameter of $p = 34$. We employed the AdamW optimizer with the learning rate 1.0, and the weight decay was fixed at 0. Besides, the data preprocessing transform was as same as the original CLIP, and the visual prompt was initialized by zero. For the external identity-embedding network, we follow the paper and employ a pre-trained TransFace model [5]. We also experiment with ArcFace [11], but both models occasionally produce invalid or empty embeddings due to low-quality or non-face input regions. We will release our implementation publicly.

C. Additional Results

C.1. Model Complexity of `VLAForge` vs. SotA Methods

Table 3 presents a comparison of model complexity and video-level AUROC across several representative deepfake detection approaches. The results highlight a clear performance–efficiency advantage of `VLAForge`. To be specific, traditional CNN- and transformer-based detectors (*e.g.*, DCL [15], FTCN [17], CFM [10]) contain 19–26M parameters, yet their average AUROC remains limited. This reflects their weak generalization when applied to unseen datasets such as CDF-v2 and DFDC. RepDFD [9] shows the smallest parameter count, but its performance indicates that extremely lightweight models generally sacrifice discriminability, especially in challenging real-world settings. ForAda [4] achieves a stronger balance between model size and performance (5.7M parameters, 91.5 AVG) due to its effective adapter-based design.

In contrast, `VLAForge` uses only 3.28M parameters—smaller than ForAda and significantly smaller than most baselines—yet achieves the highest AUROC on both CDF-v2 (96.8) and DFDC (89.6), yielding the best overall average (93.2). This demonstrates that `VLAForge` provides superior generalization and discriminative power with notably lower parameter complexity, validating the effectiveness of coupling cross-modal semantics with compact forgery-aware learning.

We also report training and inference time comparisons (mean±std) in Table 3. As shown, our method requires slightly longer runtime than ForAda, but is more efficient than RepDFD. Notably, with only marginal overhead, `VLAForge` achieves notably improvement in performance.

[1] https://github.com/OUC-VAS/ForensicsAdapter

Table 3. Model complexity analysis in Video-level AUROC.

Methods	Param.	Training (ms)	Inference (ms)	CDF-v2	DFDC	AVG
DCL	19.35M	-	-	82.3	76.7	79.5
FTCN	26.6M	-	-	86.9	74.0	80.5
CFM	25.37M	-	-	89.7	80.2	85.0
RepDFD	0.078M	**1832.4±7.5**	**42.3±0.6**	89.9	81.0	85.5
ForAda	5.7M	**933.3±1.4**	**21.2±0.2**	95.7	87.2	91.5
Ours	3.28 M	**1410.0±2.2**	**41.7±0.9**	**96.8**	**89.6**	**93.2**

C.2. Text Prompting Variants Comparison

To verify the importance of ID prior-informed text prompts in `VLAForge`, we evaluate several prompt variants: i) replacing the backbone of CLIP with LLAVA ('LLAVA'); ii)replacing the simple prompts with LLM-generated descriptive prompts ('LLM-Prompts'); iii) replacing CLIP-based identity priors with those extracted using Arc2Face ('Arc2Face-ID'); iv) substituting fixed prompts with learnable prompts (Learnable-Prompt); and v) removing the generic tokens (i.e., 'real/fake') from the prompts (w/o 'Real/Fake'). As shown in Table 4, all variants lead to noticeable performance degradation.

The variant of 'LLAVA' performs slightly worse than CLIP-based results. We attribute this to different training objectives: LLaVA is primarily optimized for visual instruction tuning tasks, while VLAForge benefits more from visual–language alignment, which is more di-015 rectly supported by CLIP's contrastive pretraining. The performance drop with 'LLM-Prompts' suggests that such prompts may lack consistent applicability across samples and exhibit weaker alignment with CLIP's visual representations of facial artifacts. In contrast, the simpler prompts in `VLAForge` provide more stable visual-language alignment. The degradation of 'Arc2Face-ID' likely stems from the fact that face recognition models focus on identity-discriminative features, which are less compatible with CLIP's text embedding space. Moreover, they are more sensitive to image quality, reducing robustness under low-quality or synthetic conditions. The decline in 'Learnable-Prompt' indicates that, although learnable tokens introduce flexibility, they compromise semantic stability, leading to less robust alignment across identities and artifact types. In contrast, fixed prompts serve as consistent semantic anchors that better support identity-aware modulation. Finally, removing the generic tokens ('w/o Real/Fake') results in a significant performance drop, demonstrating that these tokens act as explicit textual anchors that enhance discriminative capability, rather than introducing circular reasoning.

C.3. Visualization Comparison

t-SNE Visualization. A clear distinction can be observed between the feature distributions of '**T2**' and '**T4**'. As shown in Fig. 1, without leveraging cross-modal semantics, the visual-only features learned by '**T2**' fail to form meaningful clusters—samples do not exhibit identity-driven grouping, and real (positive) and fake (negative) samples lack a clear decision boundary, reflecting weak discriminability. In contrast, '**T4**' produces identity-consistent cluster structures, where real samples form compact clusters while fake samples are relatively scattered, indicating richer heterogeneity in forgery artifacts. This demonstrates that `VLAForge` facilitates more discriminative, separable, and semantically well-organized feature representations.

Table 4. Frame (F)- and video (V)-level AUROC using different prompt variants.

	Method	CDF-v2	DFDC	DFD	VQGAN	SiT
F-level	**LLaVA**	89.8	85.8	92.1	98.1	76.6
	LLM-Prompt	89.5	86.3	92.7	97.5	75.1
	Arc2Face-ID	90.6	86.4	92.5	98.0	76.9
	Learn-Prompt	83.8	82.3	90.1	96.6	76.4
	w/o **Real/Fake**	89.5	86.3	92.7	97.5	75.1
	`VLAForge`	**91.2**	**87.0**	**93.6**	**98.4**	**77.4**
V-level	**LLaVA**	95.3	88.4	96.2	99.3	84.2
	LLM-Prompt	94.9	89.0	96.7	99.0	84.2
	Arc2Face-ID	95.9	89.0	96.8	99.4	84.4
	Learn-Prompt	89.7	85.1	94.4	98.7	84.0
	w/o **Real/Fake**	94.9	89.0	96.7	99.0	84.2
	`VLAForge`	**96.8**	**89.6**	**97.2**	**99.7**	**85.9**

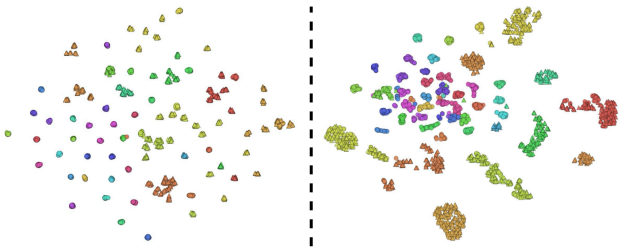

Figure 1. T-sNE visualization comparison of the DFD features between visual-only model ('**T2**' in Table 3) and complete `VLAForge` ('**T4**' in Table 3).

VLA attention maps. Fig 2 provides additional qualitative results illustrating VLA attention maps across a broader set of identities. Unlike the examples in the main text—where multiple frames from the same identity were compared to show temporal consistency—each identity here is represented by a single frame. Nonetheless, a consistent pattern emerges: without identity priors (top row), the attention maps remain sparse and unstable, often highlighting fragmented or irrelevant regions. In contrast, with identity priors injected into the text prompts (bottom row), the maps become substantially more coherent and identity-consistent, focusing on semantically meaningful facial regions. These results further validate that identity-conditioned textual semantics significantly enhance the spatial precision and reliability of VLA-based forgery indication.

Figure 2. More visualization of VLA attention maps with (w.) and without (w/o.) injecting identity prior into text prompts.).

Table 5. Frame (F)- and video (V)-level AUROC using different random seeds.

	SEED	CDF-v2	DFDC	DFD	VQGAN	SiT-XL/2
F-level	**1024**	91.2	87.0	93.6	98.3	77.4
	0000	91.4	87.9	92.8	98.6	76.9
	1111	90.7	86.7	93.2	97.9	77.6
	AVG	**91.1**	**87.2**	**93.2**	**98.3**	**77.3**
V-level	**1024**	96.8	89.6	97.2	99.7	85.9
	0000	97.0	89.9	96.4	99.6	85.4
	1111	95.9	89.2	97.5	99.4	86.3
	AVG	**96.5**	**89.6**	**97.0**	**99.6**	**85.8**

C.4. Hyperparameter Sensitivity Analysis

Table 5 reports the frame-level and video-level AUROC scores of `VLAForge` under three different and commonly used random seeds to assess its robustness and training stability. Overall, the results demonstrate high consistency across seeds, with only minor performance fluctuations, confirming that the proposed method is not sensitive to random initialization. By default, we apply a fixed random seed of 1024 to ensure training stability and reproducibility.

References

[1] Deepfake detection. https://ai.googleblog.com/2019/09/contributing-data-to-deepfakedetection.html, 2021. Accessed 2021-11-13. 1

[2] Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021. 1, 2

[3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 1, 2

[4] 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*, pages 19207–19217, 2025. 2

[5] Jun Dan, Yang Liu, Haoyu Xie, Jiankang Deng, Haoran Xie, Xuansong Xie, and Baigui Sun. Transface: Calibrating transformer training for face recognition from a data-centric perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20642–20653, 2023. 2

[6] 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. 1

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1, 2

[8] 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*, pages 3207–3216, 2020. 1

[9] 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*, pages 5262–5270, 2025. 2

[10] 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. 2

[11] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *European Conference on Computer Vision*, pages 241–261. Springer, 2024. 2

[12] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1, 2

[13] 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*, pages 1–11, 2019. 1

[14] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 1, 2

[15] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2316–2324, 2022. 2

[16] 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, 2024. 1

[17] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the*

IEEE/CVF international conference on computer vision, pages 15044–15054, 2021. 2