Prompt-Conditioned Semantic Bottleneck for Cross-Domain Face Attack Detection

Weihang Wang **Rongjie Liu** **Min Cao***
School of Computer Science and Technology
Soochow University
Suzhou, Jiangsu, China
whwang@suda.edu.cn, 20264227065@stu.suda.edu.cn, mcao@suda.edu.cn

Abstract

Learning visual representations that generalize across domains remains challenging when source-domain supervision is entangled with spurious dataset-specific cues. This issue is particularly pronounced in face attack detection (FAD), where attack artifacts vary substantially across capture devices, manipulation pipelines, and image quality. In this paper, we propose Prompt-Conditioned Semantic Bottleneck, an information-bottleneck-inspired representation learning framework for cross-domain FAD. Unlike prior methods that use prompts mainly for image-text alignment, our method uses label- and attack-type-conditioned prompts to extract attack-aware semantic targets from a frozen VLM. Specifically, the final multimodal token representation, obtained after contextual interactions between visual and prompt tokens, serves as a compact semantic bottleneck. This bottleneck provides offline supervision for a lightweight visual detector, promoting the learning of attack-relevant representations for cross-domain generalization. Extensive experiments on multiple face forgery detection benchmarks demonstrate consistent improvements under cross-dataset evaluation. Ablation studies further show that the gains are not simply due to stronger visual or multimodal models, but arise from attack-aware prompt-conditioned semantic supervision. Importantly, the VLM and prompts are removed during inference, enabling efficient vision-only deployment. Overall, our results demonstrate that prompt-conditioned VLM semantics provide an effective way to improve cross-domain FAD accuracy while maintaining efficient inference. The code is available at https://github.com/WeihangWANG/Semantic-Bottleneck-FAD.

1 Introduction

Face attack detection (FAD) aims to determine whether a face sample is authentic or manipulated, covering presentation attacks and digital face forgeries. Recent diffusion-based methods have advanced multimodally guided face synthesis [52] and concept-level control over the generation process [50], illustrating the expanding capabilities of modern generative models. Despite substantial progress in FAD methods [26, 44, 9, 19, 6, 11, 43], cross-domain FAD remains challenging because attack-related evidence is often subtle and entangled with dataset-specific nuisance factors. For example, attack cues may co-occur with particular capture devices, manipulation pipelines, or image quality patterns in the source dataset [32, 22, 2]. A detector trained with standard supervision can therefore rely on shortcuts that are predictive in the source domain but unstable in unseen target domains, resulting in severe performance degradation under domain shifts. This makes cross-domain FAD not merely a binary classification problem, but a shortcut-sensitive representation learning

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

problem. Hence, the learned representation should preserve attack-relevant cues while suppressing domain-specific biases.

To address this, existing FAD methods either model attack-related visual evidence, such as texture artifacts [17, 57], frequency cues [61, 3, 60], and local inconsistencies [51, 40, 29], or improve cross-domain generalization through adversarial learning [33, 15], meta-learning [34, 24], feature disentanglement [38, 25], and domain augmentation [44]. Recent CLIP- and VLM-based methods [35, 26, 56, 24, 45] introduce language supervision or textual prompts to improve semantic discrimination, showing that multimodal knowledge can benefit FAD. Many prompt-based methods use prompts as textual classifiers, image-text matching anchors, or learnable prompt tokens for multimodal feature learning. Such designs demonstrate the usefulness of language supervision, but they do not directly answer whether prompt-conditioned VLM representations can serve as semantic supervision for learning a more transferable representation. Whether relying on visual cues or prompt-based multimodal prediction, existing methods still learn or use representations that may be tied to source-domain evidence, therefore still face challenges in domain-invariant representations.

The main gap is not merely a stronger visual/multi-modal detector or another prompt-based classifier, but a generalizable task-conditioned semantic target that can guide the model toward transferable attack-relevant cues. In this work, we take a different view: prompts are not used to make predictions, but to specify the semantic query under which a frozen VLM produces offline training targets. This shifts the role of prompts from inference-time decision components to semantic target constructors, enabling us to study whether task-conditioned VLM semantics can guide a visual detector away from source-domain shortcuts and toward attack-relevant cues.

Based on this perspective, we propose Prompt-Conditioned Semantic Bottleneck, a representation learning framework for cross-domain FAD. The nontrivial finding is prompt-conditioned VLM semantics can serve as a semantic bottleneck that biases a visual detector toward attack-relevant factors and away from source-domain shortcuts. Experiments on face anti-spoofing and face forgery detection benchmarks show consistent improvements under cross-dataset evaluation. Furthermore, our method achieves the trade-off among accuracy, efficiency, and generalization capabilities.

Our contributions are summarized as follows:

- We formulate cross-domain FAD as a prompt-conditioned semantic bottleneck learning problem and propose Prompt-Conditioned Semantic Bottleneck, where a frozen VLM provides task-aware semantic targets constructed with label- and attack-type-conditioned prompts.
- We introduce a semantic supervision mechanism that uses precomputed prompt-conditioned VLM semantics to guide a lightweight visual detector toward attack-aware representations, without requiring VLMs or prompts during inference.
- We validate the proposed framework on cross-domain FAS and FFD benchmarks covering presentation attacks and digital face forgeries. Extensive experiments and ablations demonstrate consistent generalization improvements and a favorable trade-off among detection accuracy, inference efficiency, and cross-domain robustness.

2 Related Work

2.1 Cross-Domain Face Attack Detection

Existing vision-based FAD methods exploit various attack-related visual evidence, including texture artifacts, frequency statistics, local inconsistencies, region-level patterns, and attention mechanisms [19, 44, 9, 32, 22, 2]. To improve cross-domain robustness, prior works further introduce adversarial learning, meta-learning, feature disentanglement, domain augmentation, or domain simulation strategies [19, 44]. In terms of data-level, NDA-FAS [44] synthesize attack samples from real-face samples through negative data augmentation. In [63, 48, 5], stylization transfers the target images to the style of the source domain. In addition, pseudo-fake artifacts are generated via GAN or diffusion models in [9, 30]. In terms of feature-level, many methods [27, 38] aim to learn domain-invariant representations to narrow the gap between source and target domains. In particular, adversarial learning is used to encourage domain-invariant representations by aligning feature distributions across domains in [19, 55].

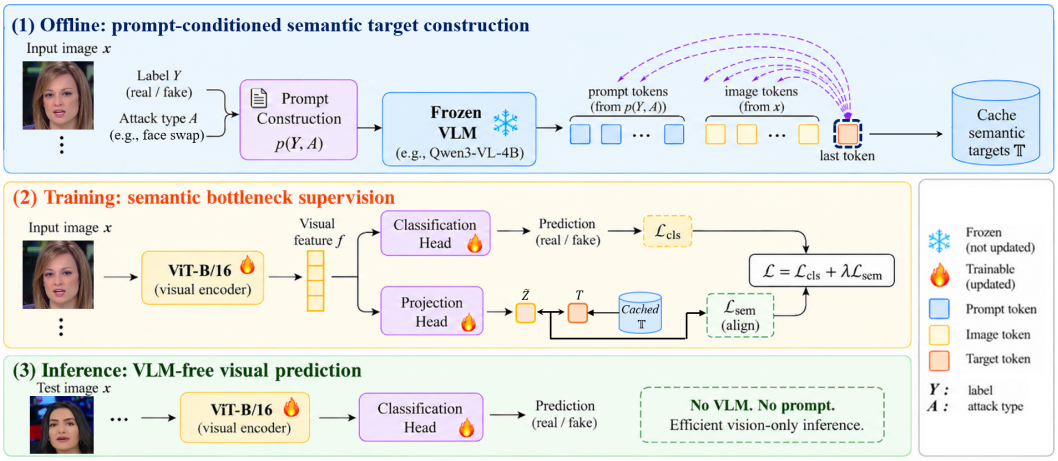


Figure 1: Overview of the proposed prompt-conditioned semantic bottleneck for cross-domain FAD.

2.2 Prompted Vision-Language Models for Face Attack Detection

Vision-language models (VLMs), especially CLIP-style models and recent instruction-following multimodal models, have been increasingly explored for FAD due to their strong semantic representation ability. Existing methods introduce textual descriptions, manually designed prompts, learnable prompt tokens, or attack-type prompts to adapt multimodal representations to spoofing and forgery detection [35, 26, 56, 24]. Moreover, FCG [14] guides CLIP-based models to focus on spatio-temporal inconsistencies in key facial regions like lips, eyes, and nose, enhancing detection of forgery traces. GM-DF [20] leverages CLIP to simultaneously learn domain-specific and domain-invariant representations across multiple datasets, improving robustness to diverse forgery techniques. VLFFD [36] uses forgery masks to prompt multimodal LLMs to generate accurate and diverse forgery descriptions, reducing hallucinations while enhancing interpretability. These studies show that language supervision can improve semantic discrimination when visual artifacts are subtle or domain-dependent in FAD. As prompts directly participate in the classification process, the resulting performance can be sensitive to prompt design, learned prompt tokens, and text-image alignment quality.

Although existing methods have advanced FAD under domain shift, they are still primarily driven by source-domain visual supervision. When attack cues are subtle and heterogeneous, such source-specific correlations can dominate the learned representation, leading to severe performance degradation under cross-domain evaluation.

3 Method

Figure 1 gives an overview of the proposed Prompt-Conditioned Semantic Bottleneck framework. It consists of three stages: offline construction of prompt-conditioned semantic targets, semantic bottleneck supervision for training a lightweight visual detector, and VLM-free visual inference.

3.1 Problem Setup and Motivation

Let X denote the input face image, Y the authenticity label, A the attack type, D the domain-specific nuisance variable, and $Z = f_\theta(X)$ the learned representation of the detector. Ideally, Z should satisfy the following trade-off:

$$\max_{\theta} \; I(Z; Y, A) - \beta I(Z; D),$$

where the first term encourages the retention of authenticity- and attack-relevant information, and the second term penalizes domain-specific information.

Directly optimizing this objective is generally intractable, and standard classification supervision only constrains the prediction $p(Y|Z)$, without explicitly regulating what information is preserved

in Z. In cross-domain FAD, this is problematic because source-domain labels may be spuriously correlated with acquisition conditions, compression artifacts, or manipulation pipelines, allowing the learned representation to capture domain-specific shortcuts. Therefore, improving cross-domain generalization requires an additional supervisory signal that can bias the detector toward attack-relevant semantics rather than source-specific regularities.

Vision-language models provide a natural source of such supervision, as their representations can integrate visual evidence with language-specified semantic context. However, directly using generic VLM features is not sufficient, since they may encode broad visual concepts that are not necessarily aligned with FAD. We therefore extract *prompt-conditioned semantics* from a frozen VLM, where the prompt explicitly specifies the authenticity conclusion and attack type. Different from conventional feature distillation, the VLM is not used as a generic predictor. Instead, it serves as an offline semantic bottleneck that provides an attack-aware target, guiding the visual detector to preserve FAD-relevant cues while suppressing domain-specific shortcuts.

3.2 Prompt-Conditioned Semantic Construction

To this end, we introduce a prompt-conditioned semantic target:

$$T = \Phi_{\mathrm{VLM}}(X, p(Y, A)),$$

where Φ_{VLM} is a frozen VLM and $p(Y, A)$ is a label- and attack-type-conditioned prompt. The prompt is designed to query the VLM from the perspective of FAD, rather than to ask the VLM to perform open-ended prediction or generation. In our implementation, the prompt takes the following form: *"This image is* $[label]$*, the attack type is* $[type]$*, please analyze the image based on this conclusion."* where $[label]$ is the real/fake label and $[type]$ is the corresponding attack type.

The use of Y and A in the prompt structure is important. The authenticity label provides the high-level task conclusion, while the attack type specifies the fine-grained semantic condition under which the image should be interpreted. Thus, the VLM representation is not a generic description of the face image, but a prompt-conditioned semantic summary biased toward FAD-relevant factors. This differs from generic prompts that may lead the VLM to focus on identity, facial attributes, or other generic multimodal representations.

The VLM parameters are fixed throughout the whole process. We perform only a single forward pass and do not use autoregressive decoding. Specifically, we take the last-layer hidden representation of the final multimodal token as the semantic target T. This hidden state is not an additionally appended token. It is produced after self-attention over both visual tokens and prompt tokens. It therefore aggregates visual evidence under the semantic context specified by the prompt, making it suitable as an attack-aware semantic carrier.

In our implementation, we use Qwen3-VL-4B as the VLM. All semantic targets are precomputed offline before training the visual detector. Therefore, the VLM is only used to construct training-time semantic supervision and is not involved in inference. Aligning Z with T then acts as a soft semantic bottleneck: it encourages the visual representation to preserve FAD-relevant information while reducing reliance on source-domain nuisance factors.

3.3 Prompt-Conditioned Semantic Bottleneck Learning

Given the prompt-conditioned semantic target T, we train a lightweight visual detector to absorb such attack-aware semantics. The detector consists of a visual encoder f_θ, a classification head h_θ, and a semantic projection head g_θ. For an input image X, the visual encoder produces representation $Z = f_\theta(X)$, and the classification head predicts the authenticity label $\hat{Y} = h_\theta(Z)$. Since the visual representation Z and the VLM semantic target T have different dimensions, we use a projection head g_θ to map Z into the semantic space $\widetilde{Z} = g_\theta(Z)$.

We then impose a semantic alignment objective between the projected visual representation and the prompt-conditioned target:

$$\mathcal{L}_{\mathrm{sem}} = 1 - \frac{\widetilde{Z}^\top T}{\|\widetilde{Z}\|_2 \|T\|_2}.$$

This loss does not require the visual detector to reproduce the full behavior of the offline VLM. Instead, it constrains the visual representation to approach the attack-aware semantic target specified

by the prompt. In this sense, the semantic alignment acts as a soft bottleneck: the detector is encouraged to encode information that is useful for FAD semantics while avoiding over-reliance on source-domain shortcuts.

The classification objective is defined by the standard cross-entropy loss:

$$\mathcal{L}_{\mathrm{cls}} = \mathrm{CE}(h_\theta(Z), Y).$$

The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda \mathcal{L}_{\mathrm{sem}},$$

where λ balances discriminative supervision and prompt-conditioned semantic supervision. The classification loss maintains source-domain discriminability, while the semantic alignment loss provides an attack-aware regularization signal for learning generalizable attack-aware representations.

3.4 Training and Inference

The proposed framework has different computational behaviors during training and inference. During offline preprocessing, the frozen VLM is used to generate prompt-conditioned semantic targets for all training samples. This step is performed once and does not require autoregressive generation. All semantic targets are precomputed and cached before detector training, which avoids repeated VLM computation. During detector training, the cached semantic targets are used together with source-domain labels to optimize the visual detector with $\mathcal{L}_{\mathrm{cls}}$ and $\mathcal{L}_{\mathrm{sem}}$.

At inference time, both the VLM and prompts are removed. Given a test image, the detector only performs a standard visual forward pass: $\hat{Y} = h_\theta(f_\theta(X))$. During inference, the final visual detector does not require any VLM computation. This design allows the model to benefit from VLM-induced attack-aware semantics while maintaining efficient vision-only inference. As a result, the proposed framework improves the trade-off among detection accuracy, cross-domain generalization, and inference efficiency.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets and Evaluation Protocols

We evaluate cross-domain generalization under two representative FAD settings: face anti-spoofing (FAS) and face forgery detection (FFD). In all experiments, the model is trained only on the designated source datasets and evaluated on unseen target datasets without any target-domain adaptation. This protocol is designed to test whether the learned representation transfers across changes in acquisition conditions, manipulation pipelines, compression levels, and image quality.

For the FAS setting, we follow the standard leave-one-out cross-domain protocol on four widely used datasets: MSU-MFSD (M) [49], CASIA-FASD (C) [59], Replay-Attack (I) [4], and OULU-NPU (O) [2]. For each split, three datasets are used as source domains for training, and the remaining dataset is held out as the unseen target domain for testing. We report HTER and AUC for each target dataset, as well as the mean performance across all leave-one-out splits.

For the FFD setting, we adopt the common cross-dataset protocol by training on FaceForensics++ (FF++, c23/HQ) [32] and directly testing on unseen datasets, including Celeb-DF-v2 (CDFv2) [22], DFDC [8], and DFDC-preview (DFDCP) [7]. This setting evaluates generalization from controlled FF++ manipulations to more diverse real-world forgery distributions. Following prior work, we report video-level AUC and frame-level AUC, respectively.

4.1.2 Implementation Details

Model architectures. We instantiate the frozen VLM with Qwen3-VL-4B and use it only for offline semantic target extraction. For each training sample, we use the prompt template: "This image is *<real/fake>*, the attack type is *<attack type>*. Please analyze the image based on this conclusion." The VLM processes each image-prompt pair with a single forward pass, without autoregressive text decoding. We extract the last-layer hidden representation of the final multimodal token as the prompt-conditioned semantic target and precompute all targets offline.

Table 1: Cross-dataset evaluation on OULU-NPU (O), CASIA-FASD (C), Replay-Attack (I), and MSU-MFSD (M) for FAS. The best results are indicated in bold, and the second-best results are underlined.

Method	Venue	OCI→M		OMI→C		OCM→I		ICM→O		Avg.	
		HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑
SSDG-R[19]	CVPR'20	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54	11.29	95.31
NDA-FAS[44]	TIFS'23	4.29	99.27	13.22	92.89	5.50	96.90	17.45	89.93	10.11	94.74
SSAN-R[48]	CVPR'22	6.67	98.75	10.00	96.67	8.88	96.79	13.72	93.63	9.82	96.46
HPDR[17]	CVPR'24	4.58	96.02	11.30	94.42	11.26	92.49	9.93	95.26	9.27	94.55
IADG[62]	CVPR'23	5.41	98.19	8.70	96.44	10.62	94.50	8.86	97.14	8.40	96.57
UDG-FAS[28]	ICCV'23	5.95	98.47	9.82	96.76	5.86	98.62	10.97	95.36	8.15	97.30
TTDG[63]	CVPR'24	7.91	96.83	8.14	96.49	6.50	97.98	10.00	95.70	8.14	96.75
SAFAS[38]	CVPR'23	5.95	96.55	8.78	95.37	6.58	97.54	10.00	96.23	7.83	96.42
MCF[46]	MM'23	4.00	98.84	8.46	96.90	8.02	97.39	10.70	95.64	7.80	97.19
DiffFAS[9]	ECCV'24	2.86	98.41	10.11	96.32	6.36	97.89	8.11	97.27	6.86	97.47
GAC-FAS[21]	CVPR'24	5.00	97.56	8.20	95.16	**4.29**	98.87	8.60	97.16	6.52	97.19
FSFM[41]	CVPR'25	3.78	99.15	3.16	99.41	4.63	**99.03**	7.68	97.11	4.81	98.68
CFPL-FAS[26]	CVPR'24	3.09	99.45	2.56	99.10	5.43	98.41	**3.33**	**99.05**	3.60	99.00
Ours	-	**1.43**	**99.67**	1.50	**99.88**	5.33	98.51	5.85	98.47	**3.53**	**99.13**

The visual detector uses ViT-B/16 as the backbone with a two-layer MLP classification head. Since the ViT-B feature dimension is 768 and the VLM semantic target dimension is 2560, we use a one-layer MLP projection head to map visual features from 768 to 2560 for semantic alignment. The projection head is used only during training, while inference only requires the visual backbone and classification head.

Training configuration. We train the model for 100 epochs on a single NVIDIA RTX 4090 GPU with batch size 12. We use AdamW with a base learning rate of 1×10^{-5}, a minimum learning rate of 1×10^{-6}, and weight decay 0.05. A 10-epoch warm-up is followed by a cosine learning-rate schedule with two 45-epoch cycles. For video inputs, we randomly sample 4 frames per video during training. All face regions are detected by RetinaFace, cropped, and resized to 224×224.

4.2 Comparison with State-of-the-art Methods

4.2.1 Evaluation on Face Anti-spoofing Benchmarks

As shown in Table 1, we evaluate cross-domain FAS under the standard leave-one-out protocol on four datasets. These datasets differ in acquisition devices, illumination conditions, subject distributions, and presentation attack media such as printed photos, replayed videos, and display devices. Therefore, this setting evaluates whether a detector can generalize to unseen presentation-attack conditions rather than relying on source-specific spoofing artifacts.

Our method achieves strong generalization across different target domains, obtaining an average HTER of 3.53% and an average AUC of 99.13%. In particular, it achieves the lowest HTER on OCI→M and OMI→C, with 1.43% and 1.50% HTER, respectively. Compared with recent prompt-based FAS methods that enhance multimodal feature learning through prompt design, such as CFPL-FAS [26], our method achieves stronger overall cross-domain performance, with clear gains on OCI→M and OMI→C while remaining competitive on OCM→I and ICM→O. This suggests that the proposed semantic bottleneck helps the visual detector learn more generalizable attack-aware representations, rather than relying solely on source-domain visual cues or generic multimodal representations.

4.2.2 Evaluation on Face Forgery Detection Benchmarks

Table 2 and Table 3 report cross-dataset video-level and frame-level AUC results for face forgery detection, respectively. All methods are trained on FF++ and directly evaluated on unseen target datasets, including CDFv2, DFDC, and DFDCP. FF++ contains several commonly studied manipulation types, such as DeepFakes, Face2Face, FaceSwap, and NeuralTextures, and provides a relatively controlled source-domain training setting. In contrast, the target datasets exhibit different forgery generation processes, subject distributions, compression levels, and visual qualities. For example, CDFv2 contains higher-quality in-the-wild deepfake videos, while DFDC and DFDCP introduce

Table 2: Cross-dataset evaluation results in video-level AUC (%) for the FFD task. All methods are trained on FF++ and evaluated on other datasets. The best results are indicated in bold, and the second-best results are underlined.

Method	Venue	Video-level AUC			
		CDFv2	DFDC	DFDCP	Avg
MCF [46]	MM'23	80.25	73.61	82.55	78.80
TALL [53]	ICCV'23	90.79	76.78	-	83.83
AUNet [1]	CVPR'23	92.77	73.82	86.16	84.25
SLF [5]	CVPR'24	89.00	-	-	89.00
NACO [58]	ECCV'24	89.50	76.70	-	83.10
LSDA [55]	CVPR'24	91.10	77.00	-	84.05
MLR [16]	CVPR'24	91.56	75.17	-	83.36
FPG [51]	MM'24	94.49	74.75	87.24	85.49
RepDFD [23]	AAAI'25	89.94	80.99	95.03	88.65
FSFM [41]	CVPR'25	91.44	83.47	89.71	88.21
TFCU [13]	CVPR'25	93.18	**86.05**	-	89.62
GM-DF [20]	MM'25	94.80	84.00	**96.30**	<u>91.70</u>
FCG [14]	CVPR'25	<u>95.00</u>	81.80	-	88.40
Ours	-	**95.28**	<u>84.99</u>	<u>96.01</u>	**92.09**

more diverse identities, scenes, and manipulation pipelines. This protocol therefore evaluates whether a detector trained on FF++ can generalize beyond source-specific forgery artifacts to heterogeneous real-world target distributions.

Our method achieves consistently strong performance across target datasets. At the video level, it obtains the best average AUC of 92.09%, outperforming recent CLIP-based methods such as FCG [14] and GM-DF [20]. Our method ranks first on CDFv2. On DFDC and DFDCP, our method reaches 84.99% and 96.01% AUC, respectively, ranking second among the compared methods. At the frame level, our method achieves the highest average AUC of 89.48%, ranking first on CDFv2 and DFDCP and remaining competitive on DFDC. The consistent improvements at both video and frame levels indicate that prompt-conditioned semantic supervision helps the visual detector learn generalizable attack-aware representations, leading to robust generalization from FF++ to unseen forgery datasets with different manipulation pipelines and visual qualities. Moreover, this performance is achieved without deploying the VLM during inference, highlighting the favorable trade-off between cross-domain accuracy and inference efficiency.

4.2.3 Evaluation on Generation-based Benchmarks

As shown in Table 4, we further evaluate the proposed method on GAN- or diffusion-based forgeries. The model is trained on FF++ and evaluated on the DF40-EFS benchmark, including two GAN-based generators, VQGAN and StyleGAN2, and four emerging diffusion-based generators, DDIM, RDDM, DiT, and SiT.

Compared with recent state-of-the-art methods, our approach remains highly competitive on GAN-based forgeries. More importantly, it demonstrates stronger generalization to emerging diffusion-based forgeries, achieving the best reported results on RDDM, DiT, and SiT, while remaining competitive on DDIM. These results suggest that the attack-aware semantic knowledge elicited from Qwen3-VL has been effectively transferred to the lightweight ViT-B detector, enabling it to generalize to previously unseen diffusion-generated forgeries without requiring the VLM or prompts at inference.

4.3 Ablation Studies

4.3.1 Effect on Training Objective

We ablate different training objectives to validate the effectiveness of the proposed semantic bottleneck supervision. As shown in Table 5, using semantic alignment alone is insufficient for learning discriminative forgery representations. By jointly optimizing classification and semantic bottleneck alignment, our method consistently achieves the best results, improving the average AUC by 3.18% over classification only. This demonstrates that the semantic bottleneck provides complementary

Table 3: Cross-dataset evaluation results in frame-level AUC (%) for the FFD task. All methods are trained on FF++ and evaluated on other datasets. The best results are indicated in bold, and the second-best results are underlined.

Method	Venue	Frame-level AUC			
		CDFv2	DFDC	DFDCP	Avg
MCF [46]	MM'23	73.16	69.63	75.78	72.86
SFDG [47]	CVPR'23	75.83	73.64	-	74.74
NoiseDF [42]	AAAI'23	75.89	-	63.89	69.89
UCF [54]	ICCV'23	82.40	-	80.50	81.45
GS [12]	ICCV'23	84.97	-	81.65	83.31
IID [18]	CVPR'23	83.80	-	81.23	82.52
LSDA [55]	CVPR'24	83.00	73.60	81.50	79.37
RepDFD [23]	AAAI'25	80.00	77.34	90.57	82.64
FSFM [41]	CVPR'25	85.05	80.20	85.50	83.58
GM-DF [20]	MM'25	89.20	**84.70**	88.20	87.37
Forensics Adapter [6]	CVPR'25	90.00	84.30	89.00	87.77
Ours	-	**91.92**	83.06	**93.46**	**89.48**

Table 4: Cross-generation evaluation results on the generation-based benchmark DF40. All methods are trained on FF++ and evaluated on other datasets. The best results are indicated in bold, and the second-best results are underlined.

Method	VQGAN	StyleGAN2	DDIM	RDDM	DiT	SiT	Avg
LSDA [55]	71.19	98.69	78.55	77.25	64.68	70.01	76.73
RepDFD [23]	80.50	82.2	-	-	58.7	58.2	69.9
Forensics Adapter [6]	93.90	92.50	-	-	62.0	69.0	79.35
MAP-Mamba [39]	85.80	92.70	**96.70**	82.70	-	-	89.47
VLAForge [64]	**99.70**	**99.70**	-	-	80.30	85.90	91.40
Ours	99.46	99.58	91.24	**93.38**	**88.34**	**91.19**	**93.87**

manipulation-aware semantic guidance, leading to more generalizable representations across unseen domains.

4.3.2 Effect on Prompt Conditioning

We investigate the impact of different prompt settings on semantic bottleneck learning. As shown in Table 6, incorporating both class labels and attack-type information achieves the best performance, outperforming the prompt-free and label-only settings by 3.18% and 2.61% in average AUC, respectively. These results suggest that attack-aware prompt conditioning facilitates the extraction of task-relevant semantic knowledge from the VLM, enabling more effective knowledge transfer and improved cross-domain generalization.

4.3.3 Effect on Semantic Bottleneck Construction

We compare different pretrained visual representations to investigate whether the improvement comes from stronger backbone features. As shown in Table 7, ViT-L and CLIP Vision features achieve comparable performance, while directly using Qwen3-VL hidden representations without bottleneck alignment does not provide additional benefits. In contrast, the proposed attack-conditioned semantic bottleneck achieves consistent improvements across all benchmarks, demonstrating that the gain comes from learning task-relevant semantic knowledge rather than simply adopting stronger pretrained representations.

4.4 Visualization Analysis and Computational Efficiency

4.4.1 CAM Visualizations

Figure 2 presents CAM visualizations on DeepFake samples from FF++, CDFv2, and DFDC. The model consistently focuses on facial regions containing manipulation-related cues, including blending

Table 5: Ablation on the training objective.

Training objective	CDFv2	DFDC	DFDCP
Classification only	92.26	83.86	92.43
Bottleneck alignment only	65.84	59.83	50.43
Classification + Bottleneck alignment	**95.28**	**84.99**	**96.01**

Table 6: Ablation on prompt designs for the semantic bottleneck.

Prompt design	CDFv2	DFDC	DFDCP	Avg.
Without prompt	91.43	82.21	93.08	88.91
Label-only	92.76	81.72	93.95	89.48
Label + attack-type	**95.28**	**84.99**	**96.01**	**92.09**

Table 7: Ablation on semantic bottleneck target construction.

Target backbone	Semantic bottleneck	CDFv2	DFDC	DFDCP	Avg.
ViT-L	×	93.21	82.88	93.15	89.75
CLIP-L	×	93.40	82.82	93.48	89.90
Qwen3-VL	×	91.43	82.21	93.08	88.91
Qwen3-VL	✓	**95.28**	**84.99**	**96.01**	**92.09**

boundaries, abnormal texture patterns, and local inconsistencies around facial components. Despite significant domain shifts in acquisition conditions, visual quality, and manipulation pipelines, the activated regions remain aligned with forgery evidence rather than irrelevant identity or background information. These observations suggest that the proposed prompt-conditioned semantic bottleneck provides effective semantic guidance, enabling the visual detector to capture more transferable forgery-related patterns across domains.

4.4.2 Computational Efficiency

Table 8 compares the inference efficiency and detection performance of different methods on FAS. CLIP-based methods usually require large backbones and introduce considerable inference cost. For example, CLIP [31] and CFPL-FAS [26] contain 169.57M and 155.08M parameters, with 51.81G and 45.46G FLOPs, respectively. FLIP [35] achieves the lowest HTER, but it uses additional CelebA-Spoof training data and still requires 149.62M parameters and 51.81G FLOPs.

In contrast, our deployed detector uses only a ViT-B visual backbone and does not require any VLM, prompt, or text-image matching during inference. Without using extra training data, our method achieves 3.53% HTER with 87.87M parameters and 16.87G FLOPs, which is about one-third of the FLOPs required by CLIP-based inference. These results show that prompt-conditioned semantic supervision improves cross-domain FAD while preserving deployment efficiency, yielding a favorable trade-off among accuracy, generalization, and inference cost.

5 Conclusion

In this paper, we proposed Prompt-Conditioned Semantic Bottleneck, a framework for cross-domain face attack detection. Unlike prior methods that use prompts mainly for image-text alignment, our method uses label- and attack-type-conditioned prompts to extract attack-aware semantic targets from a frozen VLM. Specifically, the final multimodal token representation, obtained after contextual interactions between visual and prompt tokens, serves as a compact semantic bottleneck. This bottleneck provides offline supervision for a lightweight visual detector, promoting the learning of attack-relevant representations for cross-domain generalization. Extensive experiments on multiple face forgery detection benchmarks demonstrate consistent improvements under cross-dataset evaluation. Ablation studies further show that the gains are not simply due to stronger visual or multimodal models, but arise from attack-aware prompt-conditioned semantic supervision. Importantly, the VLM

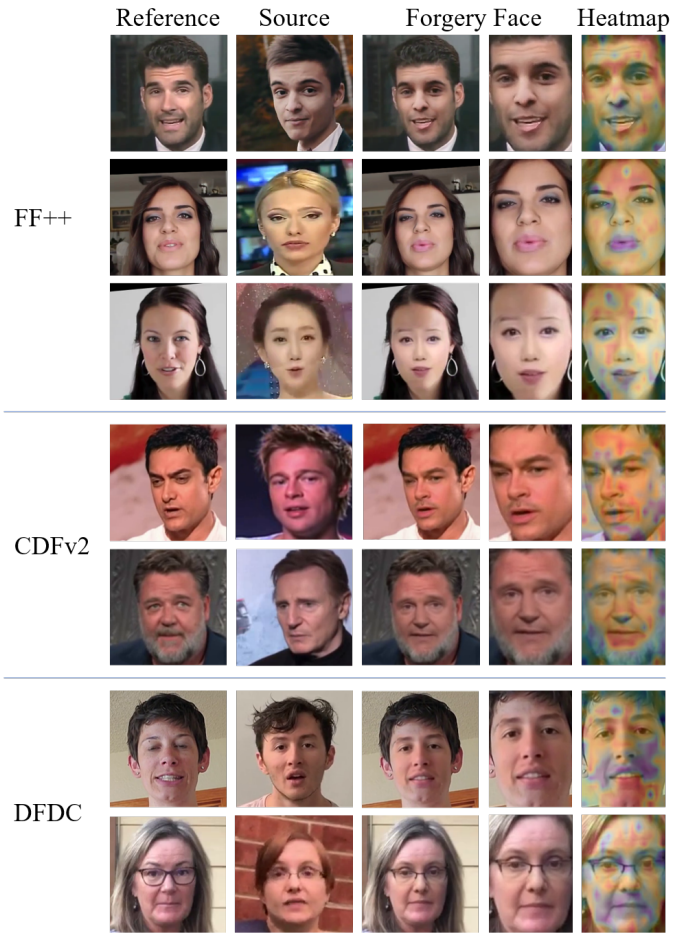


Figure 2: Qualitative visualization of forgery localizations. From left to right: reference frame, source frame, forgery face, aligned face crop for detection, and the corresponding forgery localization heatmap.

Table 8: Inference efficiency and classification performance on FAS. FLOPs are measured during inference. The best results are highlighted in bold, and the second-best results are underlined.

Method	Venue	Backbone	Parameters↓	FLOPs↓	HTER(%)↓
CLIP [31]	PMLR'21	CLIP	169.57M	51.81G	4.96
FSFM [41]	CVPR'25	MAE	**86.19M**	16.94G	4.81
CFPL-FAS [26]	CVPR'24	CLIP	155.08M	45.46G	3.60
FLIP [35]	ICCV'23	CLIP	149.62M	51.81G	**3.06**
Ours	-	ViT-B	87.87M	**16.87G**	3.53

and prompts are removed during inference, enabling efficient vision-only deployment. Overall, our results suggest that prompt-conditioned VLM semantics provide an effective way to improve the trade-off among detection accuracy, cross-domain generalization, and inference efficiency in FAD.

Limitations. Our current framework uses the final VLM token as the semantic bottleneck representation, which aggregates contextual information through interactions with textual and visual tokens while maintaining a compact supervision target. Nevertheless, designing more structured, multi-granularity semantic targets to better characterize diverse manipulation patterns remains an interesting direction for future work.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China under Grant 62476188.

References

[1] Weiming Bai, Yufan Liu, Zhipeng Zhang, Bing Li, and Weiming Hu. Aunet: Learning relations between action units for face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24709–24719, 2023.

[2] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 612–618, 2017.

[3] Junyi Cao and Chao Ma. Towards generalized face anti-spoofing from a frequency shortcut view. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1005–1015. IEEE, 2025.

[4] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *International Conference of Biometrics Special Interest Group (BIOSIG)*, pages 1–7, 2012.

[5] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1133–1143, 2024.

[6] Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19207–19217, 2025.

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

[8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv:2006.07397*, 2020.

[9] Xinxu Ge, Xin Liu, Zitong Yu, Jingang Shi, Chun Qi, Jie Li, and Heikki Kälviäinen. Difffas: face anti-spoofing via generative diffusion models. In *European Conference on Computer Vision (ECCV)*, pages 144–161, 2024.

[10] Anjith George and Sébastien Marcel. Cross modal focal loss for RGBD face anti-spoofing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7882–7891. Computer Vision Foundation / IEEE, 2021.

[11] Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–116, 2025.

[12] Ying Guo, Cheng Zhen, and Pengfei Yan. Controllable guide-space for generalizable face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20818–20827, 2023.

[13] Zonghui Guo, Yingjie Liu, Jie Zhang, Haiyong Zheng, and Shiguang Shan. Face forgery video detection via temporal forgery cue unraveling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7396–7405, 2025.

[14] Yue-Hua Han, Tai-Ming Huang, Kai-Lung Hua, and Jun-Cheng Chen. Towards more general video-based deepfake detection through facial component guided adaptation for foundation model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22995–23005, 2025.

[15] Yan He, Fei Peng, Rizhao Cai, Zitong Yu, Min Long, and Kwok-Yan Lam. Category-conditional gradient alignment for domain adaptive face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 19:10071–10085, 2024.

[16] Cheng-Yao Hong, Yen-Chi Hsu, and Tyng-Luh Liu. Contrastive learning for deepfake classification and localization via multi-label ranking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17627–17637, 2024.

[17] Chengyang Hu, Ke-Yue Zhang, Taiping Yao, Shouhong Ding, and Lizhuang Ma. Rethinking generalizable face anti-spoofing via hierarchical prototype-guided distribution refinement in hyperbolic space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1032–1041, 2024.

[18] 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 (CVPR)*, pages 4490–4499, 2023.

[19] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8484–8493, 2020.

[20] Yingxin Lai, Hongyang Wang, Jing Yang, Xiangui Kang, Bin Li, Linlin Shen, and Zitong Yu. Gm-df: Generalized multi-scenario deepfake detection. In *ACM International Conference on Multimedia (ACM MM)*, page 4300–4309, 2025.

[21] Binh M Le and Simon S Woo. Gradient alignment for cross-domain face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 188–199, 2024.

[22] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3207–3216, 2020.

[23] 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 *AAAI Conference on Artificial Intelligence (AAAI)*, pages 5262–5270, 2025.

[24] Kun-Hsiang Lin, Yu-Wen Tseng, Kang-Yang Huang, Jhih-Ciang Wu, and Wen-Huang Cheng. Instructflip: Exploring unified vision-language model for face anti-spoofing. In *ACM International Conference on Multimedia (ACM MM)*, page 2987–2996, 2025.

[25] Xun Lin, Ajian Liu, Zitong Yu, Rizhao Cai, Shuai Wang, Yi Yu, Jun Wan, Zhen Lei, Xiaochun Cao, and Alex Kot. Reliable and balanced transfer learning for generalized multimodal face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[26] Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, and Zhen Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 222–232, 2024.

[27] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Yuan Xie, and Lizhuang Ma. Dual reweighting domain generalization for face presentation attack detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 867–873, 2021.

[28] Yuchen Liu, Yabo Chen, Mengran Gou, Chun-Ting Huang, Yaoming Wang, Wenrui Dai, and Hongkai Xiong. Towards unsupervised domain generalization for face anti-spoofing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20654–20664, 2023.

[29] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17395–17405, 2024.

[30] Hina Otake, Yoshihiro Fukuhara, Yoshiki Kubotani, and Shigeo Morishima. Detect fake with fake: Leveraging synthetic data-driven representation for synthetic image detection. In *European Conference on Computer Vision (ECCV)*, pages 316–332, 2024.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[32] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2019.

[33] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10023–10031, 2019.

[34] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Regularized fine-grained meta face anti-spoofing. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11974–11981, 2020.

[35] Koushik Srivatsan, Muzammal Naseer, and Karthik Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19685–19696, 2023.

[36] Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19576–19586, 2025.

[37] Rui Sun, Fei Wang, Xiaolu Yu, Xinjian Gao, and Xudong Zhang. Robust multimodal face anti-spoofing via frequency-domain feature refinement and aggregation. *Pattern Recognit. Lett.*, 197:31–36, 2025.

[38] Yiyou Sun, Yaojie Liu, Xiaoming Liu, Yixuan Li, and Wen-Sheng Chu. Rethinking domain generalization for face anti-spoofing: Separability and alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24563–24574, 2023.

[39] 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*, 21:1184–1197, 2026.

[40] Chien-Yi Wang, Yu-Ding Lu, Shang-Ta Yang, and Shang-Hong Lai. Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20281–20290, 2022.

[41] Gaojian Wang, Feng Lin, Tong Wu, Zhenguang Liu, Zhongjie Ba, and Kui Ren. Fsfm: A generalizable face security foundation model via self-supervised facial representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24364–24376, 2025.

[42] Tianyi Wang and Kam Pui Chow. Noise based deepfake detection via multi-head relative-interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 14548–14556, 2023.

[43] Weihang Wang, Fei Wen, Haoyuan Zheng, Rendong Ying, and Peilin Liu. Conv-mlp: A convolution and MLP mixed model for multimodal face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.*, 17:2284–2297, 2022.

[44] Weihang Wang, Peilin Liu, Haoyuan Zheng, Rendong Ying, and Fei Wen. Domain generalization for face anti-spoofing via negative data augmentation. *IEEE Transactions on Information Forensics and Security (TIFS)*, 18:2333–2344, 2023.

[45] Xudong Wang, Ke-Yue Zhang, Taiping Yao, Qianyu Zhou, Shouhong Ding, Pingyang Dai, and Rongrong Ji. TF-FAS: twofold-element fine-grained semantic guidance for generalizable face anti-spoofing. In *European Conference on Computer Vision (ECCV)*, pages 148–168, 2024.

[46] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Liang Liu, Yabiao Wang, and Chengjie Wang. Toward high quality facial representation learning. In *31st ACM International Conference on Multimedia (ACM MM)*, pages 5048–5058, 2023.

[47] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7278–7287, 2023.

[48] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4123–4133, 2022.

[49] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security (TIFS)*, 10(4):746–761, 2015.

[50] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8496–8504, 2025.

[51] Ruiyang Xia, Dawei Zhou, Decheng Liu, Lin Yuan, Shuodi Wang, Jie Li, Nannan Wang, and Xinbo Gao. Advancing generalized deepfake detector with forgery perception guidance. In *ACM International Conference on Multimedia (ACM MM)*, pages 6676–6685, 2024.

[52] Zhiyuan Xia, Haojie Li, Jingyu Lin, Yiguo Qiao, and Cunjian Chen. Mtvdiff: Multimodal conditional latent diffusion for enhanced thermal-to-visible face translation. In *Computer Vision – ECCV 2026*, pages 36–52, Cham, 2026. Springer Nature Switzerland.

[53] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22658–22668, 2023.

[54] 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 (ICCV)*, pages 22412–22423, 2023.

[55] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8984–8994, 2024.

[56] Jeongmin Yu, Susang Kim, Kisu Lee, Taekyoung Kwon, Won-Yong Shin, and Ha Young Kim. Multi-view slot attention using paraphrased texts for face anti-spoofing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21117–21128, 2025.

[57] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5294–5304. Computer Vision Foundation / IEEE, 2020.

[58] Daichi Zhang, Zihao Xiao, Shikun Li, Fanzhao Lin, Jianmin Li, and Shiming Ge. Learning natural consistency representation for face forgery video detection. In *European Conference on Computer Vision (ECCV)*, pages 407–424, 2024.

[59] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *International Conference on Biometrics (ICB)*, pages 26–31, 2012.

[60] Tianyi Zheng, Bo Li, Shuang Wu, Ben Wan, Guodong Mu, Shice Liu, Shouhong Ding, and Jia Wang. Mfae: Masked frequency autoencoders for domain generalization face anti-spoofing. *IEEE transactions on information forensics and security*, 19:4058–4069, 2024.

[61] Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, Junyu Dong, et al. Freqblender: Enhancing deepfake detection by blending frequency knowledge. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 44965–44988, 2024.

[62] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Ran Yi, Shouhong Ding, and Lizhuang Ma. Instance-aware domain generalization for face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20453–20463, 2023.

[63] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Shouhong Ding, and Lizhuang Ma. Test-time domain generalization for face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 175–187, 2024.

[64] Jiawen Zhu, Yunqi Miao, Xueyi Zhang, Jiankang Deng, and Guansong Pang. Unleashing vision-language semantics for deepfake video detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 42953–42963, 2026.

A Qualitative Analysis

t-SNE visualizations. We also visualize the learned representations on unseen FFD target datasets using t-SNE. For each dataset, we randomly sample 400 real and 400 fake videos and project their features into two dimensions. As shown in Figure 3, real and fake samples exhibit clearer separation in the target domains. These qualitative results are consistent with our quantitative findings and provide additional evidence that the proposed semantic bottleneck improves attack-relevant representation learning.

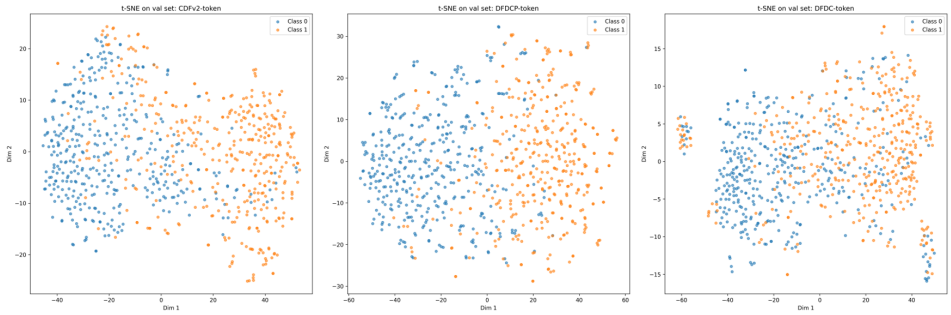


Figure 3: T-SNE visualizations for face forgery detection, illustrating the feature distributions of our method on the CDFv2, DFDC, and DFDCP datasets, respectively.

B Prompt Design and Offline Semantic Teacher

All training samples follow a unified prompt template: "This image is [real/fake], the attack type is [attack type], please analyze the image based on this conclusion." Figure 4 presents the system instruction and prompt template used in our method. The face examples are sampled from the CDFv2 dataset for illustration.

For FAS, the attack types are print photo and replay video, while FFD uses DeepFakes, FaceSwap, Face2Face, FaceShifter, and NeuralTextures. Real samples are assigned with "None". Therefore, the prompt structure remains consistent across tasks, and only source-domain attack annotations are used. No target-domain labels, attack annotations, or prompts are involved.

We adopt Qwen3-VL-4B as the offline semantic teacher. The VLM is frozen and only used to extract the final hidden token after image-prompt interaction as the multimodal semantic target. No autoregressive decoding is performed in our implementation. These extracted semantic targets are cached and used during ViT-B training, while the VLM and prompts are completely removed during inference.

C Offline VLM Overhead

Although our framework uses Qwen3-VL-4B to construct semantic targets, this computation is performed only once during offline preprocessing. For each training sample, the frozen VLM requires a single forward pass without autoregressive decoding, costing approximately 1.05 TFLOPs per sample with a throughput of 0.61 samples/s on an NVIDIA RTX 4090 GPU. The extracted semantic targets are cached and reused during detector training. Therefore, the VLM cost is an offline one-time overhead, while inference remains as efficient as a standard vision-only detector.

D Additional FPAD Evaluation

We further evaluate broader unseen attack types for face presentation attack detection (FPAD). We conduct experiments on the WMCA dataset under the leave-one-out protocol, which includes more diverse unseen physical attacks.

For this multimodal FPAD setting, we keep the same prompt structure and loss functions as the main method. Specifically, the real/fake label and attack type are used as structured prompt conditions.

Instruction

You are an expert in face forgery detection. Given the provided authenticity label, inspect the facial image for forgery evidence, including but not limited to:
(1) face-replacement or cut-and-paste traces;
(2) abnormal geometric consistency, such as facial proportions, pose, or deformation;
(3) boundary or contour anomalies, such as misalignment, splicing, or screen/frame edges.
Ignore identity, attractiveness, demographics, and background unless they are relevant to authenticity. No textual response is required; only perform the conditioned visual-semantic analysis.

Input image



Prompt

This image is a **real face**, the attack type is **none**. Please analyze the image based on this conclusion.

Input image



Prompt

This image is a **fake face**, and the forgery type is **face-swapping**. Please analyze the image based on this conclusion.

Figure 4: Illustration of the system instruction and label-conditioned prompt template. The template conditions the frozen VLM on authenticity labels and attack semantics, while the final hidden representation from a single forward pass is used as the offline semantic supervision signal. The face examples are sampled from the CDFv2 dataset for illustration.

Table 9: Comparison of ACER (%) across different unseen presentation attack types on the WMCA dataset under the LOO protocol.

Method	Fake Head	Glasses	Paper Mask	Rigid Mask	Flexible Mask	Replay	Print	Mean $\pm$ Std
CDCN [57]	2.50	19.80	4.50	9.70	7.90	8.60	2.80	7.99 $\pm$ 5.50
CMFL [10]	2.50	33.50	1.80	1.70	12.40	1.00	0.70	7.60 $\pm$ 11.20
Conv-MLP [43]	**0.20**	32.50	0.90	2.30	12.60	0.80	**0.10**	7.05 $\pm$ 11.16
DRWT-RDIA [37]	1.10	18.40	0.50	4.10	12.90	0.60	0.80	5.49 $\pm$ 7.23
Ours	1.30	16.80	0.87	**0.00**	10.25	**0.00**	1.74	**4.42 $\pm$ 6.05**

The downstream detector is trained with the loss functions. The only difference is that, the input of the model is multimodal, including RGB, depth, and infrared images.

Table 9 shows the result across methods under different presentation attack types. Our method achieves the lowest mean ACER of 4.42% among the compared methods. The overall result shows that the proposed prompt-conditioned semantic bottleneck also generalizes well to more diverse 2D/3D physical attacks for real-world cases.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's main claims, as stated in the abstract and introduction, are clearly articulated and are well-supported by both the analysis and experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses its limitations in Section 5, including evaluation on broader real-world scenarios and offline VLM overhead.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results, theorems, or proofs. The information bottleneck perspective is used as motivation rather than as a formal theoretical guarantee.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the datasets, evaluation protocols, model architecture, prompt template, training losses, and key hyperparameters needed to reproduce the main results. Additional implementation details, including offline VLM target extraction and qualitative analyses, are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the code for reproduction in the supplementary material. All datasets used in our experiments are open source and clearly referenced at the corresponding points in the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the training and testing protocols, cross-dataset settings, optimization details, model backbones, sampling strategy, and evaluation metrics. The appendix further provides details on prompt templates and offline semantic target extraction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For feature discrepancy analysis, the paper reports mean and standard deviation over three random seeds. For standard benchmark comparisons, we follow common FAS and FFD evaluation protocols and report deterministic metrics such as HTER and AUC.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports inference FLOPs and parameter counts for efficiency comparison. It also provides the offline VLM preprocessing cost, including the use of an NVIDIA RTX 4090 GPU and empirical throughput.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research fully adheres to the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses the positive impact of improving robustness against face spoofing and deepfake attacks. It also acknowledges potential risks such as in real-world deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets used in the paper are properly cited and appropriately credited, with licenses and terms of use correctly respected in the manuscript.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.