Beyond [CLS] Token: Query-Driven Token-Level Forgery Purification for Generalizable Deepfake Detection

Supplementary Material

1. Classification Token Ablation Study

In our proposed framework, we aggregate query tokens with the last-layer patch tokens into a specific global token that serves as the classification representation. This design is motivated by our observation that the pre-trained [CLS] token tends to emphasize global semantics but may overlook local forgery cues due to the Pre-trained Information Bias (PIB). Note that we have already compared the performances without any losses between the global token and the [CLS] token in the main body of our paper (see the ablation part, first two rows of Table 3). Here, to better understand how different classification tokens affect performance, we conduct extra ablation study to explore the complementary fusion of the [CLS] token and the query-based global token with our proposed losses. Specifically, we compare three simple fusion schemes to our default setting:

- **Global (Default):** using the global token alone for classification.
- **[CLS] + Global (Avg):** averaging the [CLS] token and the global token for classification.
- **[CLS] + Global (Concat):** concatenating the [CLS] token and the global token for classification.
- **[CLS] + Global (Weighted-Avg):** averaging the [CLS] token and the global token with learnable weights for classification.

Table 1 summarizes the results based on Effort [2]. When averaged across all ten test datasets, the performance ranks from best to worst as follows:Global (Default, 0.947) $>$ [CLS] + Global (Concat, 0.942) $>$ [CLS] + Global (Avg, 0.940) $\approx$ [CLS] + Global (Weighted-Avg, 0.940). It is clear that using the global token alone reaches the best average performance. Nevertheless, combining the [CLS] token with the global token through simple fusion schemes (averaging, concatenation, or weighted averaging) results in slightly lower performance. Furthermore, when examining each dataset individually, using the global token alone outperforms in most cases, although there are a few datasets where fusion schemes yield better results.

Based on these observations, we identify two key findings. First, the query-based aggregation design is well suited for the deepfake detection task, enabling the model to better capture local forgery information. Second, although the [CLS] token offers complementary semantic information and can enhance performances on specific test datasets, it also brings in excessive pre-trained global semantic priors that hinder overall performance. These findings justify our choice to use the query-based global token as the primary classification representation, while treating the [CLS] token as an auxiliary semantic source rather than the sole classification token.

2. Detailed Analysis on Losses

2.1. FLCL vs. Mask-Aware Contrastive Learning (MACL, An Oracle Upper Bound)

To further compare, we evaluate our mask-free fake-likelihood contrastive learning loss (FLCL) against an oracle upper bound loss, *i.e.*, the mask-aware contrastive learning loss (MACL), since pixel-level forgery masks are available in a subset of the training data. As for MACL, we use the ground-truth mask to explicitly define the set of forged patch tokens, *i.e.*, only patch tokens within the forged region are considered forged, while those outside are treated as real. The anchors for this contrastive loss are all forged patch tokens, each given equal importance.

Table 2 presents the results about this comparison based on Effort [2]. Note that we only replace the contrastive loss while keeping all other components fixed to our optimal setting. As expected, the oracle MACL achieves the best performance, with an average AUC of 0.919, since it has access to the exact forgery locations. However, our FLCL, which does not rely on mask supervision and instead uses prototype-based fake-likelihood weights to estimate forged regions, recovers most of the performance gain provided by the oracle, with an average AUC of 0.914. These results suggest that FLCL can implicitly localize and highlight genuinely forged patches using only global real/fake supervision, serving as a robust surrogate for an ideal mask-aware objective. This supports our claim that prototype-guided fake-likelihood weighting strategy can serve as an effective substitute for explicit masks.

2.2. Effect of RAAL on Attention Behavior

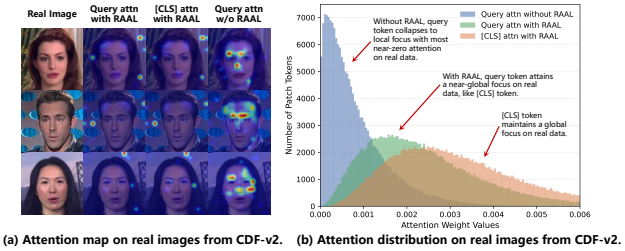
To further validate the effectiveness of the proposed real-aware attention loss (RAAL), we conduct a comparative analysis of attention behaviors on real images from the Celeb-DF v2 dataset. As illustrated in Figure 1(a), without RAAL, query tokens exhibit high activation patterns even on real images, indicating that the model tends to "over-search" for forgery cues in authentic regions, which may lead to false positives. In contrast, with RAAL, the attention maps of query tokens become significantly more suppressed and sparse, closely aligning with those of the [CLS] to-

Table 1. **Classification token ablation study for cross-dataset evaluation.** We conduct a comparison of four distinct classification token variants based on Effort [2], each equipped with our proposed losses. We report the video-level AUC for each dataset, as well as the average performance across all ten test datasets.

Token Used	CDF-v1	CDF-v2	FSh	UADFV	DFD	DFDC	DFDCP	DFo	WDF	FFIW	Avg.
Global (Default)	**0.980**	**0.960**	0.913	**0.989**	0.975	0.869	0.925	**0.992**	**0.917**	**0.950**	**0.947**
[CLS]+Global (Avg)	0.959	0.957	**0.917**	0.984	0.956	0.866	**0.940**	0.986	0.914	0.921	0.940
[CLS]+Global (Weighted-Avg)	0.962	0.957	0.908	0.988	0.961	0.869	0.929	0.988	0.914	0.930	0.940
[CLS]+Global (Concat)	0.960	0.959	0.887	0.986	**0.976**	**0.875**	0.918	0.991	0.916	0.945	0.942

Table 2. **Comparison between our mask-free FLCL loss and the oracle MACL loss.** We report cross-dataset video-level AUC based on Effort [2] for two settings: employing our proposed FLCL loss and using the oracle MACL loss.

Losses	CDF-v2	FSh	DFDC	Avg.
FLCL	0.960	0.913	0.869	0.914
MACL (Oracle)	0.962	0.923	0.873	0.919



(a) Attention map on real images from CDF-v2. (b) Attention distribution on real images from CDF-v2.

Figure 1. **Impact of RAAL on attention behavior for real images from Celeb-DF v2.** (a) Qualitative comparison of attention heatmaps among [CLS] token (with RAAL), query token (with RAAL), and query token (without RAAL). Note how RAAL suppresses redundant activations in the query tokens, making them consistent with the [CLS] token on real samples. (b) Statistical distribution of attention scores. The incorporation of RAAL shifts the query token's attention distribution towards the [CLS] token and effectively reducing forensic noise on authentic images.

ken. This observation is further supported by the attention score distributions in Figure 1(b). These results suggest that RAAL effectively regularizes the query-guided purification process, encouraging the model to focus on genuine forgery signals rather than being distracted by intrinsic image textures or pre-training biases. Therefore, RAAL is consistent with our key insight: real images require global, holistic inspection, whereas fake images are often characterized by localized forgery artifacts.

2.3. Metric Choice and Alternative Real-Preserving Regularization for RAAL

Real-attention alignment learning loss (RAAL) aligns the attention distributions of the query token and the pre-trained [CLS] token on real images, aiming to preserve beneficial real global semantics while enabling the query tokens to focus on forgery-aware cues. In our default setting, we use the Kullback–Leibler (KL) divergence to measure the discrepancy between these two attention maps. Here, we try to investigate how different alignment metrics of RAAL affect performance, and then provide a brief comparison of RAAL with other alternative real-preserving regularizations.

Table 3. **Comparison among three distance/divergence discrepancy metrics used in RAAL.** Based on Effort [2], we report the video-level AUC scores on CDF-v2 and UniFace datasets, as well as the average AUC across both datasets.

RAAL Metrics	CDF-v2	UniFace	Avg.
Manhattan Distance (L1)	0.958	0.982	0.970
Euclidean Distance (L2)	0.958	0.985	0.972
KL Divergence	0.960	0.988	0.974

Metric Choice. Given the attention vectors of the query token and the [CLS] token on real images, denoted as $\mathbf{a}^{\mathrm{Q}}$ and $\mathbf{a}^{\mathrm{CLS}}$, respectively, we consider three metrics to measure their discrepancy:

- **L1 (Manhattan) Distance:**

$$\|\mathbf{a}^{\mathrm{Q}} - \mathbf{a}^{\mathrm{CLS}}\|_1. \tag{1}$$

- **L2 (Euclidean) Distance:**

$$\|\mathbf{a}^{\mathrm{Q}} - \mathbf{a}^{\mathrm{CLS}}\|_2. \tag{2}$$

- **KL Divergence (Ours):**

$$\mathrm{KL}\big(\mathbf{a}^{\mathrm{Q}} \,\|\, \mathbf{a}^{\mathrm{CLS}}\big), \tag{3}$$

  where both vectors are normalized to probability distributions before computation.

Table 3 reports results based on Effort [2]. Note that only the distance/divergence metrics of RAAL are changed, with all other components held consistent at their optimal settings. Overall, KL divergence achieves the best AUC, followed by L2 distance and then L1 distance. However, the average AUC differences among the three metrics are relatively small, indicating that RAAL is not overly sensitive

Table 4. **Comparison of alternative real-preserving regularizations.** We compare three variants based on Effort [2]: no real-preserving regularization (None), L2 token-level alignment (L2 Token Alignment, *i.e.*, employing alignment between the `[CLS]` token and the global token on real images with L2 distance), and our proposed attention-level real-attention alignment (RAAL).

Regularization	CDF-v2	FSh	DFDC	Avg.
None	0.952	0.902	0.860	0.904
L2 Token Alignment	0.957	0.900	0.863	0.907
RAAL	0.960	0.913	0.869	0.914

to the exact choice of distance/divergence. We conjecture that KL divergence is slightly more suitable here because it respects the probabilistic nature of attention maps and encourages the query token to match the shape of the `[CLS]` attention distribution rather than only minimizing euclidean distance in the vector space.

Alternative Real-Preserving Regularization. Except the attention-level alignment, we further compare RAAL with simpler real-preserving regularizers that operate in the feature space. Concretely, we consider the following three variants:

- **None:** the detector is trained without any explicit real-preserving constraint.
- **L2 Token Alignment:** we apply an L2 distance penalty between the `[CLS]` token and the global token on real images, encouraging them to stay close in the token-level space.
- **RAAL:** our default attention-level alignment, which minimizes the discrepancy (using KL divergence) between the query-token attention map and the `[CLS]` token attention map on real images.

Table 4 summarizes the results on three test datasets. Note that we only perform ablation on the RAAL part, while all other components remain at their optimal settings. Across all test three datasets, RAAL consistently achieves the best AUC, followed by L2 token alignment, while the model without any real regularization performs the worst. These results suggest that preserving pre-trained real-image semantics is beneficial to some extent. However, aligning attention distributions (RAAL) is more effective than simple token-level alignment, likely because attention-level alignment preserves the spatial structure of real-image evidence and prevents over-constraining the token-level representations.

2.4. Sensitivity Analysis of Loss Weight

We further investigate how the loss weights for FLCL and RAAL affect performance. Specifically, we vary each loss weight around its default value used in the main experiments and report the AUC on three representative datasets (FaceShifter, WDF, and DFDCP), along with their average.

For FLCL, Figure 2 shows how the video-level AUC changes with different FLCL loss weights on the three datasets and their average. Overall, the curves are relatively stable: within a reasonably wide range around our default setting, the AUC only exhibits mild fluctuations, and the average AUC remains close to its peak. When the weight is set too low, the benefit of FLCL is diminished; when it is set too high, the model tends to slightly overemphasize the contrastive signal at the expense of the base classification loss. For RAAL, Figure 3 presents an analogous study by varying the RAAL loss weight. We observe a similar pattern: the AUC curves on FaceShifter, WDF, and DFDCP change smoothly as the weight increases, and the average AUC reaches its maximum near the default choice. These results together demonstrate that both FLCL and RAAL are robust to the choice of loss weights.

3. Attention Weight Visualization

To further illustrate how different fine-tuning strategies attend to image regions, we additionally visualize attention weight maps for Effort, LoRA, FFT, and their counterparts enhanced with our method (Effort + Ours, LoRA + Ours, FFT + Ours). For each method, we select five real and five fake images, and plot the corresponding attention maps derived from the query tokens or last-layer `[CLS]` token. Figure 4 displays the original images alongside the attention maps for all six methods. For each image, the six attention maps are arranged in a consistent order, enabling direct visual comparison between the baseline fine-tuning strategies and their variants (+Ours) on both real and fake examples.

4. Efficiency and Parameter Usage Analysis

We compare the computational and training costs of various fine-tuning strategies, both with and without the incorporation of our method, on the FF++ (c23) dataset [1]. Specifically, We evaluate three backbone fine-tuning schemes: Effort, LoRA, and FFT, along with their variants enhanced by our method. For each approach, we report (i) the number of trainable parameters, (ii) per-image GFLOPs, and (iii) training time per epoch, as summarized in Table 5. All training time measurements were conducted on identical hardware and under consistent experimental settings. As shown, integrating our method into each fine-tuning strategy results in only a marginal increase in both trainable parameters and training time per epoch, while the overall GFLOPs remain virtually unchanged.

References

[1] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforen-

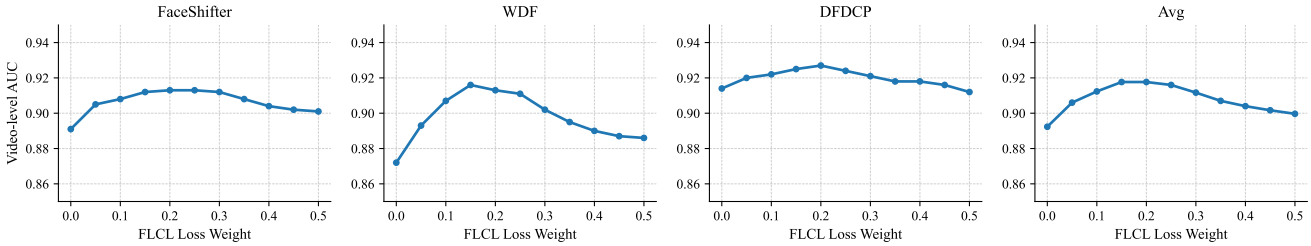


Figure 2. **Sensitivity analysis of FLCL loss weight.** We report video-level AUC on three test datasets, along with their average.

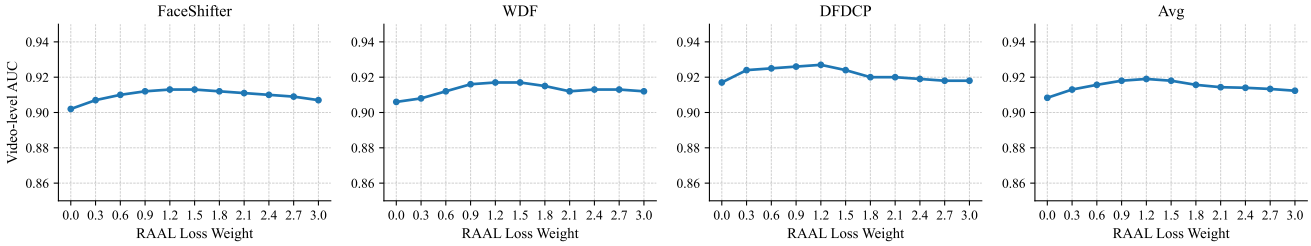


Figure 3. **Sensitivity analysis of RAAL loss weight.** We report video-level AUC on three test datasets, along with their average.

sics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019. 3

[2] Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024. 1, 2, 3

Table 5. **Comprehensive analysis of computational efficiency and model parameter complexity.** We compare trainable parameters, per-image GFLOPs, and training time per epoch on FF++ (c23) for Effort, LoRA, FFT, and their variants enhanced with our method.

Setting	Effort	Effort+Ours	LoRA	LoRA+Ours	FFT	FFT+Ours
#Params	0.19M	4.39M	1.77M	5.97M	305.04M	309.24M
GFLOPs	155.6	155.6	155.7	155.7	155.6	155.6
Training Time Per Epoch	8min32s	8min51s	10min21s	10min32s	10min42s	10min51s

Figure 4. **Attention map visualizations for Effort, LoRA, FFT, and their variants (Effort + Ours, LoRA + Ours, FFT + Ours)** For each image, the original input is shown alongside attention maps from all six methods, allowing direct comparison of their attention distribution.