DeFB: Decomposed Feature Learning for Real-Time Multi-Person Eyeblink Detection in Untrimmed In-the-Wild Videos

Jinfang Gan[1], Wenzheng Zeng[1, 2*], Yang Xiao[1†], Xintao Zhang[1], Chaoyang Zheng[1], Ran Zhao[1], Ran Wang[3, 4], Min Du[5], Zhiguo Cao[1]

[1]National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, 430074, China
[2]National University of Singapore
[3]School of Journalism and Information Communication, Huazhong University of Science and Technology
[4]School of Future Technology, Huazhong University of Science and Technology
[5]ByteDance, Beijing, 100089, China

Abstract

Multi-person eyeblink detection in untrimmed in-the-wild videos is a recently emerged and challenging task. Due to its significant spatio-temporal fine-grained characteristics compared to general actions, we empirically find that general action detectors, though effective in general domains, struggle with this task (i.e., Blink-AP<2%). Specialized eyeblink detection methods alleviate it through fine-grained spatio-temporal operations. SOTA method proposes a unified model combining instance-aware face localization and eyeblink detection through joint multi-task learning and feature sharing. While effectiveness, it exhibits two critical limitations that may contribute to its unsatisfactory performance (i.e., Blink-AP=10.11%): (1) Face localization and eyeblink detection require distinct spatio-temporal feature granularities, making joint modeling in a unified feature space suboptimal. (2) Eyeblink task training could be largely affected by unstable face-eye feature learning under the joint training paradigm. To address this, we propose DeFB, a decomposed feature learning paradigm with favorable effectiveness and efficiency: (1) We model faces and eyes in granularity-specific feature spaces, which enhances fine-grained perception while reducing computational costs compared to a unified feature space. (2) To mitigate face-eye feature learning instability, we adopt an asynchronous learning mechanism where eye feature learning refines well-trained coarse face features, with shared queries acting as a bridge between stages to retain the efficient feature sharing of existing unified models. Compared with SOTA method, DeFB doubles the performance (Blink-AP: 24.65% v.s. 10.11%) while boosting efficiency by nearly 35%. DeFB can also be integrated as a plug-in to substantially augment the eyeblink detection capabilities of general action detectors.

Code — https://github.com/jinfanggan/DeFB

1 Introduction

Multi-person eyeblink detection in the wild for untrimmed video (Zeng et al. 2023b) is a recently emerged and challenging task with vital applications such as fatigue detection (Rosenfield 2011; Bergasa et al. 2006) and emotion analysis. The task requires face detection, tracking, and spatio-temporal fine-grained eyeblink detection in an instance-aware manner from untrimmed in-the-wild videos, which is challenging as it not only needs to overcome complex environments such as extreme poses, lighting variations, occlusion, and multi-person interactions but also demands an order-of-magnitude finer spatio-temporal granularity that is far beyond general actions (as shown in Fig. 1 (a)), while balancing inference time especially in crowd scenes. Such difficulty is further validated by the unsatisfactory perfor-

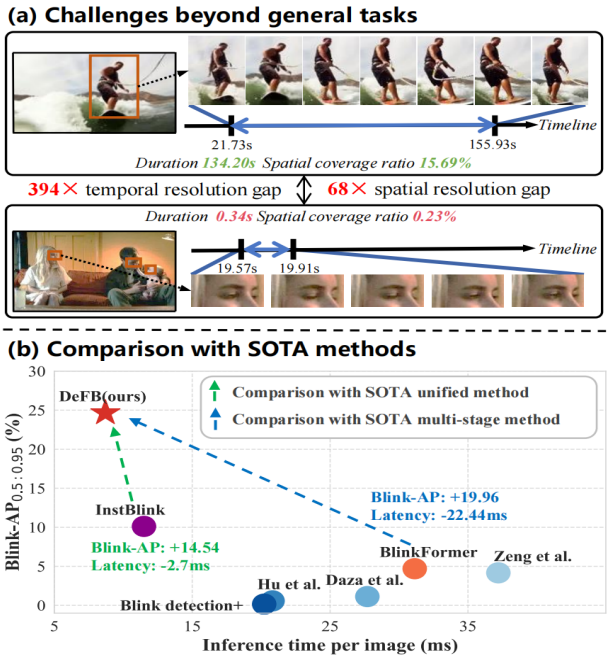

Figure 1: (a) Eyeblink detection faces more pronounced spatio-temporal challenges than general actions. (b) DeFB achieves a superior accuracy-efficiency balance compared to other SOTA methods.

*Project lead: Wenzheng Zeng (HUST→NUS).
†Corresponding author: Yang Xiao (Yang_Xiao@hust.edu.cn).

mance of existing general action detection detectors (Zhang, Wu, and Li 2022; Shi et al. 2023; Yang et al. 2024; Kim et al. 2025) (i.e.,Blink-AP$<$2%) and SOTA multi-person eyeblink detection method (Zeng et al. 2023b) (Blink-AP $\approx$ 10%).

In this work, we first rethink the design of the SOTA framework and identify its key limitations, which motivate our solution. Specifically, the SOTA method InstBlink (Zeng et al. 2023b) addresses this task with a unified model that simultaneously performs instance-aware face detection, tracking, and eyeblink detection. By designing a unified feature space for joint multi-task learning and execution, it achieves significant improvements in effectiveness and efficiency over sequential frameworks (isolated face tracking followed by single-person eyeblink detection) (Zeng et al. 2023a; Liu, Xu, and Lu 2023; Daza et al. 2021). Despite the effectiveness of this unified design, its performance remains suboptimal (Blink-AP $\approx$ 10%) due to two critical shortcomings: (1) Face localization and eyeblink detection demand distinct spatiotemporal feature granularities, rendering unified-space joint modeling suboptimal: over-modeling for coarse-grained face localization introduces redundancy and reduces efficiency, while unified features fail to support the fine-grained spatiotemporal modeling required for eyeblink detection. (2) Eyeblink feature learning is significantly affected by unstable face-eye feature learning under the joint training paradigm, which undermines the effectiveness of eyeblink feature learning, particularly in early training stages. Beyond ablation studies, we have designed interpretable metrics to support this observation in Sec. 3.2.

To address this, we propose DeFB, a decomposed feature learning paradigm with favorable effectiveness and efficiency. Given the granularity differences between face localization and eyeblink detection, we use global coarse-grained features for face localization. These features are obtained via spatial modeling on the highest-scale image features from the backbone. For eyeblink detection, we employ fine-grained eye features derived through our dense spatiotemporal modeling module, which builds on these global features. We further adopt an asynchronous learning strategy for face and eyeblink features, treating eyeblink feature learning as a refinement of well-trained coarse face features and designing shared queries to serve as a bridge between the two stages. This design yields two benefits: (1) Eyeblink feature learning becomes more effective by avoiding interference from unstable face-eye features in joint learning; (2) The framework retains efficient feature sharing through both this refinement process and shared queries, as in existing unified models, ensuring high inference efficiency.

Extensive experiments show DeFB effectively addresses multi-person eyeblink detection challenges (Fig. 1 (b)), achieving substantial speed and accuracy gains over SOTA methods. It can also be seamlessly integrated into general action detectors to enhance their eyeblink detection capabilities. The main contributions of this paper are as follows:

• We rethink the design of the recent SOTA unified model from a novel viewpoint, where we uncover two crucial limitations: the conflict on feature granularity and unstable joint learning. We design metrics to support our viewpoint in addition to the ablation studies for a deeper understanding.

• We propose DeFB, a decomposed feature learning framework for unified multi-person eyeblink detection, which models faces and eyes in granularity-specific feature spaces with an asynchronous optimization strategy, which significantly enhance fine-grained perception, learning stability, and inference efficiency.

• Our method significantly outperforms current SOTA methods in both effectiveness and efficiency, and can be seamlessly plugged into general action detectors as a high-performance plug-in.

2 Related Works

2.1 Multi-person eyeblink detection methods

Multi-person eyeblink detection in unconstrained videos (Zeng et al. 2023b) is a recently emerged challenging task, differing from prior single-person, constrained, or trimmed settings (Pan et al. 2007; Drutarovsky and Fogelton 2014; Fogelton and Benesova 2016; Radlak, Bozek, and Smolka 2015; Daza et al. 2024, 2020). It requires instance-aware face detection, tracking, and eyeblink detection in untrimmed in-the-wild videos without constraints. Existing methods fall into two categories: (1) Multi-stage methods (Zhao et al. 2025; Zeng et al. 2023a; Liu, Xu, and Lu 2023; Daza et al. 2021; Hu et al. 2020; Soukupová and Cech 2016), which follow a sequential pipeline (face detection $\rightarrow$ tracking $\rightarrow$ eye region extraction $\rightarrow$ eyeblink detection) but suffer from error propagation and instance count-sensitive latency. (2) Unified model, which uses a shared feature space for joint face localization and eyeblink detection, outperforming multi-stage methods via multi-task learning and feature sharing. We identify limitations in the SOTA unified model (Zeng et al. 2023b) and propose DeFB, which doubles performance while improving efficiency by 35%.

2.2 Action localization

Action localization includes spatio-temporal methods (Gu et al. 2018; Jhuang et al. 2013; Soomro, Zamir, and Shah 2012; Köpüklü, Wei, and Rigoll 2019; Yang and Kun 2023; Dang et al. 2024) (detecting action tubes) and temporal detectors (Zhang, Wu, and Li 2022; Shi et al. 2023; Yang et al. 2024; Kim et al. 2025) (predicting temporal boundaries). Effective for general actions, they lack inherent support for persistent instance tracking across frames. Multi-person eyeblink detection, by contrast, demands consistent instance identity and fine-grained per-instance analysis, relying on subtler spatio-temporal patterns. Even with external face trackers for instance awareness (Guo et al. 2022), SOTA temporal action detectors show limited eyeblink detection performance (i.e.,Blink-AP$<$2%), indicating value in task-specific designs. Our DeFB not only achieves strong performance but also serves as a plug-in to compensate for the lack of instance-awareness and fine-grained spatio-temporal feature modeling in generic action detection heads, significantly boosting their eyeblink detection accuracy.

2.3 End-to-end query-based methods

End-to-end object detection models (Carion et al. 2020; Zhu et al. 2021; Li et al. 2022; Meng et al. 2021; Zhao et al.

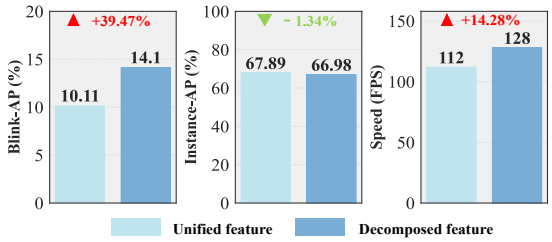

Figure 2: Comparison of the performance and efficiency of InstBlink between unified and decomposed feature spaces.

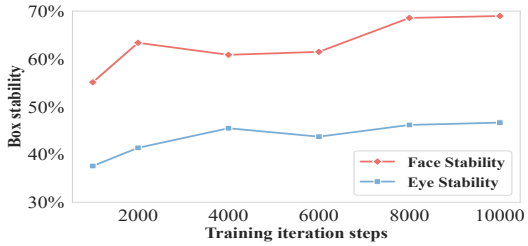

Figure 3: The variation in face and eye box stability of InstBlink during joint training.

2023; Lv et al. 2024) employ query-based designs to eliminate manually designed anchor boxes and non-maximum suppression. Thanks to end-to-end learning, this query concept extends to downstream tasks like action recognition (Liu et al. 2022; Gritsenko et al. 2023), instance segmentation (Cheng et al. 2022; Wu et al. 2022; Tan et al. 2024), and human pose estimation (Shi et al. 2022). The current SOTA unified multi-person eyeblink detector InstBlink follows this paradigm, using shared queries interacting with video features to simultaneously perform face localization and eyeblink detection. While effectiveness, this unified design has key limitations. We argue that eyeblinks and faces have inherently different characteristics and should be modeled separately rather than in the same feature space.

3 Rethinking the Unified Model Design

In this section, we rethink the SOTA unified multi-person eyeblink detection method (i.e., InstBlink (Zeng et al. 2023b)), uncovering its key limitations that contribute to its unsatisfactory performance, which motivate our solution.

3.1 Is unified feature modeling optimal?

InstBlink uses a unified feature space for both facial and eyeblink modeling, with the goal of enabling efficient feature sharing and allowing for multi-task training. However, we argue that face and eyeblink have distinct characteristics: facial modeling (localization/tracking) is relatively coarse-grained and does not require fine-grained high-resolution features, while eyeblink modeling needs dense spatio-temporal modeling to capture subtle eye region changes. This creates a representation dilemma: face modeling may have redundant features (harming efficiency), and insufficient eyeblink modeling reduces detection accuracy.

To verify this, we decomposed the feature space in a toy experiment: for face modeling, we removed high-resolution fine-grained features from the lowest encoder scale; for eyeblink modeling, we used the same face feature space (without fine-grained scale) but added a lightweight spatio-temporal module. As shown in Fig. 2: (1) Fine-grained features barely affect facial modeling (Inst-AP drops by 1.34%). (2) Separate spatio-temporal modeling for eyeblink significantly improves performance by 39.47%. (3) Overall efficiency increases by 14.28% (due to reduced redundant computation). This validates that decomposed feature modeling benefits both effectiveness and efficiency.

3.2 Is multi-task joint training optimal?

Beyond the unified feature space, InstBlink employs joint training to optimize facial and eyeblink representations simultaneously, aiming to achieve multi-task synergy for performance gains. However, we argue this paradigm also remains sub-optimal due to unstable face-eye feature learning: immature facial representations in early training stages lead to unstable eye region features, hindering effective optimization of the eyeblink detection.

To quantify this instability, we designed metrics using bounding box (bbox): for faces, we measured IoU overlap of predicted bboxes across training iterations (lower overlap indicates less stability); similarly, we introduced an eye head to generate eye region bboxes and used IoU overlap of these bboxes. This bbox stability analysis serves as a technical means to assess the underlying feature instability.

Results in Fig. 3 (calculated as IoU between current iteration proposals and those 1000 iterations prior) confirm significant instability in both face and eye features—particularly in eye features critical for eyeblink modeling. Notably, even when the model reaches a relatively stable training stage, the eye feature stability remains below 50%This indicates that joint optimization inherently causes unstable face-eye feature learning that persists beyond the initial training phases. Based on this, we propose our asynchronous training solution, where asynchronous training enables eyeblink detection to use stable face-eye features during training, significantly accelerating convergence and improving accuracy, as illustrated in Sec. 5.3.

4 Method

Building on the above analysis, we propose DeFB, a method that achieves superior effectiveness and efficiency through two core designs: learning facial and eyeblink features in decomposed sub-spaces, and adopting an asynchronous training strategy—while retaining the efficient feature sharing of existing unified models (Zeng et al. 2023b).

DeFB's pipeline (Fig. 4) takes videos as input to simultaneously perform face detection, tracking, and instance-level eyeblink detection, consisting of below key modules: (1) The facial modeling module for overall facial representation and instance-aware face and eye localization across the whole video; (2) The eyeblink modeling module for local dense spatio-temporal eyeblink representation and is responsible for subsequent eyeblink detection; (3) These two modules are trained asynchronously but are connected through a

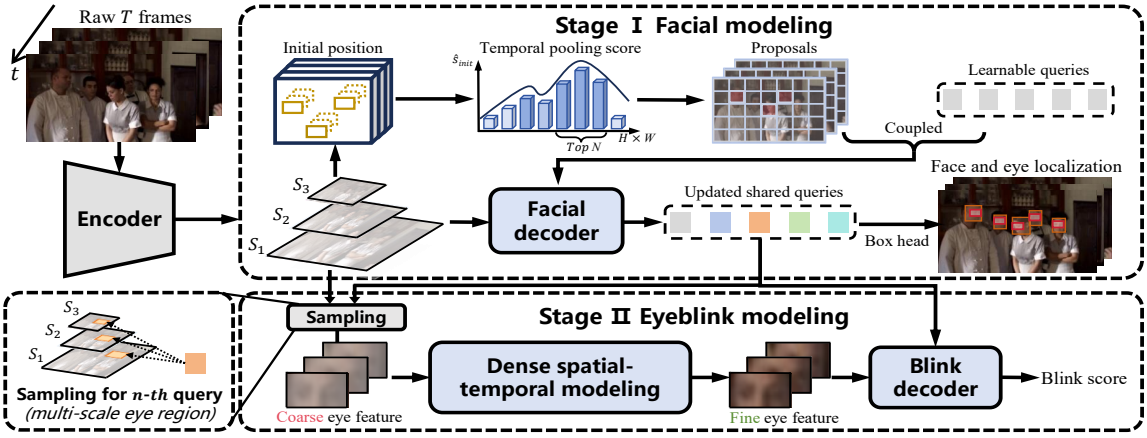

Figure 4: Overview of the proposed DeFB method.

set of shared instance-aware queries that make efficient feature sharing and information exchange while preserving the benefits of decomposed modeling.

4.1 Facial modeling

Here we begin with the illustration of the proposed facial modeling module, which is used for instance-level face-related localization and tracking while also serving as the basis for subsequent fine-grained eyeblink modeling.

Coarse-grained video feature representation. We first propose an efficient approach to represent facial features from the video. Guided by the observations in Sec. 3.1, which suggest that a relatively coarse-grained feature space is sufficient for facial modeling (while boost efficiency), we utilize a lightweight encoder, which extracts coarse-grained frame-level features from the video $I \in \mathbb{R}^{T\times 3\times H\times W}$ (where T is the video length and $H \times W$ is the resolution) and conducts spatial modeling at the highest scale of these frame-level features for enhancement. In contrast to InstBlink, we only employ the last three scales of the encoded features $\{S_1, S_2, S_3\}$ (disregarding the fine-grained high-resolution scale) for subsequent modeling. This design reduces the total representation volume by 75% compared to InstBlink, substantially boosting inference efficiency.

Dynamic proposal selection. Following the query-based paradigm, we represent human instances via spatial-temporal instance queries $\{q^i\}_{i=1}^N \in \mathbb{R}^{N\times T\times C}$, where N denotes the number of queries and C denotes the feature dimension. Unlike prior methods that rely on input-independent proposal initialization (Zeng et al. 2023b), we consider inherent spatio-temporal prior and introduce an efficient dynamic proposal selection strategy leveraging global scene features $\{S_1, S_2, S_3\}$: first, we initialize learnable queries $\{q^i\}_{i=1}^N$, then generate refined initial proposals by concatenating encoder outputs $\{S_1, S_2, S_3\}$ across scales to form $content \in \mathbb{R}^{T\times(\hat{H}\times\hat{W})\times C}$, predicting pixel-wise scores $s_{\text{init}} \in \mathbb{R}^{T\times(\hat{H}\times\hat{W})}$ and initial positions $p_{\text{init}} \in \mathbb{R}^{T\times(\hat{H}\times\hat{W})\times 4}$. Then scores will be temporally averaged to obtain $\hat{s}_{\text{init}} \in \mathbb{R}^{(\hat{H}\times\hat{W})}$, and selecting top-$N$ pixels as initial proposals $p \in \mathbb{R}^{T\times N\times 4}$. This approach produces more temporally consistent proposals aligned with real-world instance dynamics, simplifying subsequent decoding.

Facial decoder. The Facial decoder decodes queries for instance-level facial localization across the video. First, multi-head self-attention (Vaswani et al. 2017) is applied to queries $\{q^i\}_{i=1}^N$ along temporal and spatial dimensions to capture long-range spatio-temporal dependencies. For each frame, encoder features $\{S_1, S_2, S_3\}_t$ and frame-specific queries $q_t \in \mathbb{R}^{N\times C}$ are updated independently using a multi-scale DETR decoder architecture (Zhu et al. 2021; Zhao et al. 2023). The updated queries $\{q^i\}_{i=1}^N$ are fed to a box head, outputting:

$$b_{\text{face}}, b_{\text{eye}}, s = \text{Boxhead}(q), \tag{1}$$

where $b_{\text{face}} \in \mathbb{R}^{N\times T\times 4}$ (face positions), $b_{\text{eye}} \in \mathbb{R}^{N\times T\times 4}$ (eye positions), and $s \in \mathbb{R}^{N\times T}$ (confidence scores). After M iterations, the final results from the last iteration are used for facial localization.

4.2 Eyeblink modeling

As analyzed in Sec. 3.1, separate modeling of fine-grained eyeblink features yields notable benefits. We thus introduce a lightweight eyeblink modeling module, which treats eyeblink feature learning as a refinement process based on global coarse-grained features and incorporates shared queries as a bridge between stages, combined with an asynchronous training strategy (detailed in Sec. 4.3). This design offers two key advantages: (1) Eyeblink feature learning avoids interference from unstable face-eye features in joint optimization; (2) Building eye features on facial features with shared queries retains efficient feature sharing, maintaining high inference efficiency as a unified model. Below, we detail the fine-grained eyeblink modeling module.

Fine-grained eyeblink feature extraction. For the n-th instance identified by the facial modeling module, we leverage a shared instance query q^n (originally used for face/eye localization) and two key feature sources derived from it: the updated query features $q^n \in \mathbb{R}^{T\times C}$ from the facial decoder, and the coarse eye features $F^n \in \mathbb{R}^{T\times S\times C}$—extracted via

multi-scale ROI align (He et al. 2017) based on the eye region predicted by the paired query. Specifically, the ROI align operation samples S points from global coarse-grained features based on the final eye positions b_{eye} (decoded by q^n), generating spatially localized eye features across all T frames. To capture subtle spatiotemporal dynamics critical for eyeblink modeling, we apply multi-layer self-attention along temporal and spatial dimension to F^n, refining it into fine-grained local eye features $\bar{F}^n \in \mathbb{R}^{T\times S\times C}$.

Eyeblink detection. The refined eye features $\bar{F}^n$ are fed into the Blink decoder, where the shared query q^n (reused from facial modeling) interacts with $\bar{F}^n$ through a two-stage attention mechanism to predict eyeblink scores. First, multi-head cross-attention (MHCA) (Vaswani et al. 2017) enables q^n to aggregate discriminative information from $\bar{F}^n$; then, multi-head self-attention (MHSA) along temporal dimension further refines the query by modeling internal dependencies. This process is formulated as:

$$\begin{gathered} q^n = \text{MHCA}(q^n, \bar{F}^n), \quad q^n = \text{MHSA}(q^n), \\ \text{blink}_{score} = \text{MLP}(q^n), \end{gathered} \tag{2}$$

where $\text{blink}_{score} \in \mathbb{R}^T$ denotes the predicted eyeblink probability for each frame. During inference, we determine eyeblink intervals by applying a threshold to blink_{score}.

4.3 Asynchronous training

As analyzed in Sec. 3.2, eyeblink feature learning is severely affected by unstable face-eye feature dynamics during joint training. To address this, we propose an asynchronous training strategy (illustrated in Fig. 4) that first trains the facial modeling module independently, then uses the stabilized global facial features to guide eyeblink module training. Below, we detail the training process for each module.

For facial modeling. The facial module is trained separately to obtain accurate instance localization, avoiding the adverse effects of unstable facial representations on eyeblink modeling. Its training involves two loss components: *(1) Dynamic proposal selection loss*: Initial positions p_{init} and scores s_{init} are matched with ground-truth faces via frame-level Hungarian matching (Kuhn 1955), with the loss computed as:

$$\mathcal{L}_{init\_pos} = \mathcal{L}_{box}(p_{init}, \overline{b}_{face}) + \mathcal{L}_{cls}(s_{init}, \overline{s}), \tag{3}$$

where $\overline{b}_{\text{face}}$ and $\overline{s}$ denote ground-truth face locations and categories, respectively. *(2) Facial decoder loss*: Each decoder layer's predictions (face/eye positions and scores) are matched with ground truths via instance-level Hungarian matching. The loss is formulated as:

$$\begin{gathered} \mathcal{L}_{cls} = \mathcal{L}_{cls}(s, \overline{s}), \\ \mathcal{L}_{track\_pos} = \mathcal{L}_{box}(b_{face}, \overline{b}_{face}) + \mathcal{L}_{box}(b_{eye}, \overline{b}_{eye}), \\ \mathcal{L}_{decoder} = \mathcal{L}_{track\_pos} + \mathcal{L}_{cls}, \end{gathered} \tag{4}$$

where b_{face}, b_{eye}, and s are the decoder's predicted face/eye locations and scores; $\overline{b}_{\text{eye}}$ is the ground-truth eye location.

For eyeblink modeling. The eyeblink module is trained solely for eyeblink detection optimization. The loss is defined as:

$$\mathcal{L}_{\text{blink}} = \mathcal{L}_{CrossEntropy}(\text{blink}_{\text{score}}, \text{blink}_{\text{gt}}) \tag{5}$$

where $\text{blink}_{\text{score}}$ is the predicted eyeblink probability sequence, and blink_{gt} is the corresponding ground-truth.

5 Experiment

5.1 Dataset and evalutaion metrics

The MPEblink dataset (Zeng et al. 2023b) is the only existed multi-person unconstrained eyeblink dataset, which serve as our main evaluation benchmark. Each frame is annotated with instance-level bounding boxes and blink labels, totaling 8,748 blink actions. Previous multi-person eyeblink detection method have been evaluated on this dataset using two metrics: Instance-AP and Blink-AP. Instance-AP: 3D IoU between predicted trajectories and ground truth (face localization). Blink-AP: Temporal IoU between predicted/-ground truth blink intervals for true positive instances with IoU $\geq 50\%$ in Instance-AP.

The HUST-LEBW dataset (Hu et al. 2020) is the only other unconstrained eyeblink dataset besides MPEblink. Although this dataset does not belong to the multi-person eyeblink detection task, to demonstrate the generalization ability of our method, we also conduct experiments on this dataset and compare it with other methods, where F1 score is reported following existing methods.

Implementation details. The face localization module initializes the encoder with RT-DETRv2-M weights (Lv et al. 2024) (pre-trained on COCO (Lin et al. 2014)), with input resolution 640×360, $N = 100$ queries, 6 decoder layers, AdamW optimizer (1×10^{-5} backbone lr, 1×10^{-4} other lr), 2000 warm-up steps, trained on 2×3090 GPUs (batch size 8). The eyeblink module uses 5×4 ROI sampling points, 6 layers for both spatial-temporal transformer encoder and blink decoder, trained with AdamW (1×10^{-4} backbone lr) for 20 epochs. Inference uses clip length 42, with joint prediction of tracking and blink results via IoU-based fusion.

5.2 Benchmark results on MPEblink dataset

Baselines. Multi-stage methods follow the protocols established in prior work (Zeng et al. 2023b): facial landmark detection and tracking $\rightarrow$ eye region /landmark extraction $\rightarrow$ eyeblink detection. For the unified method, we directly adopt InstBlink using its open-source code, following standard practices.

Main results. The overall metrics are shown in Tab. 1, and the comparison of speed is shown in Tab. 2:

• In terms of the eyeblink metric, Blink-AP, our method outperforms multi-stage methods (Blink-AP: 24.65% vs. 4.69%) and InstBlink (24.65% vs. 10.11%) on Blink-AP. Improvements are more pronounced with higher action IoU thresholds: Blink-$\text{AP}_{0.75}$ (24.62% vs. 7.16%) and Blink-$\text{AP}_{0.95}$ (4.40% vs. 0.62%). This is attributed to finer-grained features in our blink module and mitigation of unstable face-eye feature learning in joint training.

• In terms of the instance-awareness metric, Instance-AP, our method outperforms tracking-by-detection frameworks via long-sequence modeling. Compared to InstBlink, lighter facial features enable stacking more decoder layers (4 vs. 6) with lower hardware cost. The dynamic proposal selection eases decoder refinement, further boosting face localization.

• In terms of speed, multi-stage methods are slower than unified methods due to cross-stage feature redundancy. Our method, compared to InstBlink, uses coarser global features

Type	Method	Blink-AP$_{0.5:0.95}$	Blink-AP$_{0.5}$	Blink-AP$_{0.75}$	Blink-AP$_{0.95}$	Inst-AP
Multi-stage	Hu et al. (Hu et al. 2020)	0.57	2.68	0.04	0.00	56.70
	Daza et al. (Daza et al. 2021)	1.12	4.00	0.63	0.02	
	Blink detection+ (Phuong et al. 2022)	0.15	0.58	0.06	0.00	
	BlinkFormer (Liu, Xu, and Lu 2023)	4.69	19.95	0.54	0.00	
	Zeng et al. (Zeng et al. 2023a)	4.16	16.68	1.04	0.00	
Unified	InstBlink (Zeng et al. 2023b)	10.11	27.19	7.16	0.62	67.89
	DeFB (ours)	**24.65**	**44.17**	**24.62**	**4.40**	**76.07**

Table 1: The main results on the MPEblink dataset.

Method	Time/per image (ms)
Hu et al. (Hu et al. 2020)	T (=9.3) + 5.7 × #faces
Daza et al. (Daza et al. 2021)	T (=9.3) + 9.1 × #faces
Blink detection+ (Phuong et al. 2022)	T (=9.3) + 5.4 × #faces
BlinkFormer (Liu, Xu, and Lu 2023)	T (=9.3) + 10.8 × #faces
Zeng et al. (Zeng et al. 2023a)	T (=9.3) + 13.8 × #faces
InstBlink (Zeng et al. 2023b)	8.9 + D (=2.6)
DeFB (ours)	**6.1 + D (=2.6)**

Table 2: Inference speed is evaluated on MPEblink using a single NVIDIA 3090 GPU. T denotes InsightFace's (Guo et al. 2022) single-frame inference time (including preprocessing); #faces is the number of faces in videos; D represents unified methods' preprocessing time.

Decomposed feature space	Async training	BlinkAP$_{0.5:0.95}$
✓	✓	**24.65**
✓	×	14.47
×	✓	12.93
×	×	7.96

Table 3: Component effect of the proposed DeFB.

for face modeling and fine-grained features only for eye regions, achieving higher speed (164 FPS).

5.3 Ablation study

Decomposed feature and asynchronous training. To validate our two core improvements for unified eyeblink detection, we conduct ablation studies on MPEblink, focusing on decomposed feature space between face and eyeblink, and asynchronous training. The results in Tab. 3 show a significant accuracy boost from our method.

Specifically, feature decomposition enhances eyeblink detection by enabling fine-grained spatio-temporal modeling of eye features, capturing subtle eyelid movements while improving efficiency—this significantly raises Blink-AP from 12.93% to 24.65%. As showed in Fig. 5, this decomposition greatly improves inter-class separability between eyeblink and non-eyeblink features, highlighting the need for dedicated eye modeling due to granularity incompatibility with unified face representations.

The asynchronous training strategy further boosts performance (Blink-AP from 14.47% to 24.65%) by shielding eyeblink features from unstable face tracking during joint training. As shown in Fig. 6, it accelerates convergence with faster loss reduction and accuracy gains, while our stability metric confirms this dynamics: joint training keeps eye feature stability below 50%, whereas asynchronous learning achieves 100% stability via fixed face features, yielding smoother curves. These results validate our improvements, with the stability metric proving effective for analyzing training dynamics—consistent with final performance.

Dynamic proposal selection	Inst-AP	Inst-AP$_{0.5}$	Inst-AP$_{0.75}$
✓	**76.07**	**88.87**	**83.72**
×	74.63	88.03	81.36

Table 4: Effect of the dynamic proposal selection.

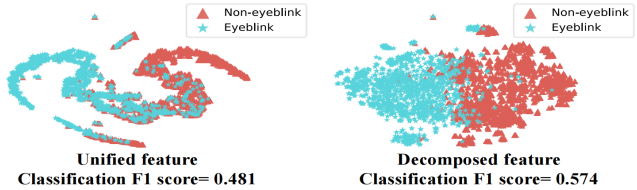

Figure 5: feature separability analysis.

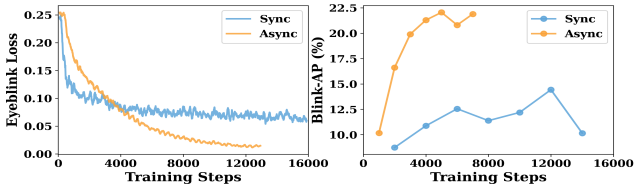

Figure 6: Convergence analysis.

Dynamic proposal selection. Ablation experiments validate the efficacy of our dynamic proposal selection. As shown in Tab. 4, the method generates initial positions via video feature-guided selection, whereas the baseline uses rasterized uniform sampling. Results confirm that our strategy enhances tracking performance.

5.4 Benchmark results on HUST-LEBW dataset

To verify DeFB's generalization, we test it on the single-person HUST-LEBW dataset. Given HUST-LEBW's limited training data and absence of precise face labels, uni-

Type	Method	Eye	Recall	Precision	F1
Multi-stage	Soukupová and Cech	Left	36.07	64.71	46.32
		Right	30.16	57.58	39.58
	Hu et al.	Left	54.10	89.19	67.35
		Right	44.44	76.71	56.28
	mEBAL2	Left	89.68	80.14	84.64
		Right	87.80	78.26	82.76
	Blinkdetection+	Both	58.99	80.05	67.90
	BlinkFormer	Left	81.97	75.18	78.42
		Right	86.51	79.56	82.89
	Zeng et al.	Left	**91.80**	89.60	90.69
		Right	91.27	**92.74**	**92.00**
Unified	InstBlink	Both	**91.34**	76.82	83.45
	DeFB(Ours)	Both	88.98	**84.96**	**86.92**

Table 5: DeFB cross-dataset result on HUST-LEBW.

Method	Baseline	DeFB Plug-in	Decoder Speed
Actionformer	1.42	26.36 **(+24.94)**	2.39ms
TriDet	1.59	26.96 **(+25.37)**	1.67ms
DyFADet	1.76	27.32 **(+25.56)**	4.02ms
DiGIT	1.71	26.61 **(+24.90)**	4.59ms
DeFB (alone)	—	24.65	0.51ms

Table 6: Blink-AP and latency of DeFB as plug-in.

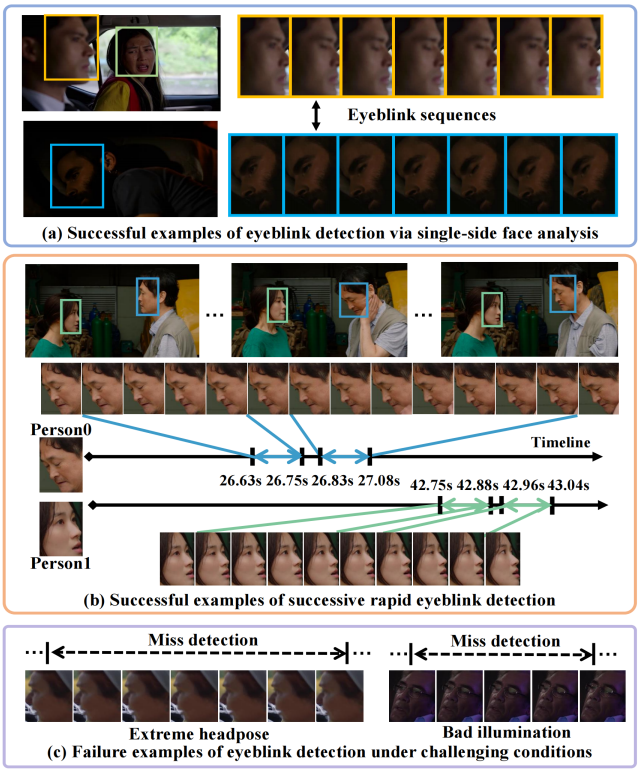

Figure 7: Qualitative analysis of successful and failed cases.

fied methods (including DeFB) use cross-dataset testing (trained on MPEblink training set, tested on HUST-LEBW). Results (Tab. 5) show DeFB maintains robust generalization, outperforming most multi-stage and unified methods. Unlike multi-stage methods requiring pre-located facial regions, DeFB operates end-to-end (face detection + tracking + eyeblink detection) without auxiliary inputs, demonstrating greater versatility and broader application potential.

5.5 Plug-and-play boost for action detectors

As shown in Tab 6, general action detectors (Zhang, Wu, and Li 2022; Shi et al. 2023; Yang et al. 2024; Kim et al. 2025) exhibit limited eyeblink detection capability even with accuray GT eye crop provided as input (Blink-AP $< 2\%$, as shown in the baseline column). Beyond its strong standalone capabilities—achieving a superior balance of performance and efficiency, DeFB can also serve as a plug-in to significantly boost the performance for SOTA action detectors. By integrating with them, it consistently lifts their accuracy, outperforming their standalone results. Notably, this dual role validates its versatility as both a high-performance standalone solution and a powerful booster for existing methods.

5.6 Qualitative analysis

We visualize DeFB's predictions on MPEblink. As shown in Fig. 7 (a), DeFB can accurately detect eyeblinks even under extreme headpose (only single-side face is visible). The Fig. 7 (b) highlights a fast consecutive eyeblink scenario where InstBlink struggles, but our method DeFB effectively identifies the rapid eyeblink sequence, demonstrating superior robustness. The Fig. 7 (c) reveals DeFB's limitations, which cannot accurately detect extreme difficult eyeblink samples under extreme headpose and bad illumination.

6 Conclusion and Limitation

In this work, we rethink the SOTA unified model design for multi-person eyeblink detection, uncovering two key limitations: (1) A unified feature space for face localization and eyeblink detection causes a representation dilemma—redundancy hinders efficiency for face modeling, while insufficient eyeblink modeling degrades accuracy; (2) Joint optimization leads to unstable face-eye feature learning due to immature facial representations, impairing eyeblink detection. To address these, we propose DeFB, a decomposed feature learning paradigm with asynchronous training. DeFB achieves significant improvements in accuracy and speed, pushing eyeblink detection to a new state (Blink-AP_{50} $\approx 45\%$). Additionally, as a plug-in, it compensates for general action detectors' lack of fine-grained spatio-temporal modeling, substantially boosting their eyeblink accuracy—validating its value as both a standalone solution and a performance booster. However, it remains less robust in extreme poses and poor lighting in unconstrained scenarios. Future work will explore specialized designs to tackle these challenges.

Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China under Grant No. 62271221, the National Social Science Foundation of China under Grant No. 25BXW041, and the Taihu Lake Innovation Fund for Future Technology, Huazhong University of Science and Technology (HUST) under Grant 2023-B-8.

References

Bergasa, L. M.; Nuevo, J.; Sotelo, M. A.; Barea, R.; and Lopez, M. E. 2006. Real-time system for monitoring driver vigilance. *IEEE Transactions on Intelligent Transportation Systems*, 7(1): 63–77.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proc. European Conference on Computer Vision (ECCV)*, 213–229.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1290–1299.

Dang, D. M. N.; Duong, V. H.; Wang, J. C.; and Duc, N. B. 2024. YOWOv3: An Efficient and Generalized Framework for Human Action Detection and Recognition. arXiv:2408.02623.

Daza, R.; DeAlcala, D.; Morales, A.; Tolosana, R.; Cobos, R.; and Fierrez, J. 2021. ALEBk: Feasibility study of attention level estimation via blink detection applied to e-learning. *arXiv preprint arXiv:2112.09165*.

Daza, R.; Morales, A.; Fierrez, J.; and Tolosana, R. 2020. MEBAL: A multimodal database for eye blink detection and attention level estimation. In *Proc. International Conference on Multimodal Interaction (ICMI)*, 32–36.

Daza, R.; Morales, A.; Fierrez, J.; Tolosana, R.; and Vera-Rodriguez, R. 2024. mEBAL2 database and benchmark: Image-based multispectral eyeblink detection. *Pattern Recognition Letters*, 182: 83–89.

Drutarovsky, T.; and Fogelton, A. 2014. Eye blink detection using variance of motion vectors. In *Proc. European Conference on Computer Vision (ECCV)*, 436–448.

Fogelton, A.; and Benesova, W. 2016. Eye blink detection based on motion vectors analysis. *Computer Vision and Image Understanding*, 148: 23–33.

Gritsenko, A.; Xiong, X.; Djolonga, J.; Dehghani, M.; Sun, C.; Lučić, M.; Schmid, C.; and Arnab, A. 2023. End-to-End Spatio-Temporal Action Localisation with Video Transformers. *arXiv e-prints*, arXiv:2304.12160.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6047–6056.

Guo, J.; Deng, J.; An, X.; Yu, J.; and Gecer, B. 2022. InsightFace: 2D and 3D Face Analysis Project. https://github.com/deepinsight/insightface.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2961–2969.

Hu, G.; Xiao, Y.; Cao, Z.; Meng, L.; Fang, Z.; Zhou, J. T.; and Yuan, J. 2020. Towards Real-Time Eyeblink Detection in the Wild: Dataset, Theory and Practices. *IEEE Transactions on Information Forensics and Security*, 15: 2194–2208.

Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; and Black, M. J. 2013. Towards understanding action recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 3192–3199.

Kim, H.-J.; Lee, Y.; Hong, J.-H.; and Lee, S.-W. 2025. DiGIT: Multi-Dilated Gated Encoder and Central-Adjacent Region Integrated Decoder for Temporal Action Detection Transformer. *arXiv e-prints*, arXiv:2505.05711.

Köpüklü, O.; Wei, X.; and Rigoll, G. 2019. You Only Watch Once: A Unified CNN Architecture for Real-Time Spatiotemporal Action Localization.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2022. Dn-detr: Accelerate detr training by introducing query denoising. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13619–13627.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2014. Microsoft COCO: Common Objects in Context. *arXiv e-prints*, arXiv:1405.0312.

Liu, B.; Xu, Y.; and Lu, F. 2023. SynBlink and BlinkFormer: A Synthetic Dataset and Transformer-Based Method for Video Blink Detection. In *Proc. British Machine Vision Conference (BMVC)*, 127–134.

Liu, X.; Wang, Q.; Hu, Y.; Tang, X.; Zhang, S.; Bai, S.; and Bai, X. 2022. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31: 5427–5441.

Lv, W.; Zhao, Y.; Chang, Q.; Huang, K.; Wang, G.; and Liu, Y. 2024. RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer. arXiv:2407.17140.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional detr for fast training convergence. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 3651–3660.

Pan, G.; Sun, L.; Wu, Z.; and Lao, S. 2007. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 1–8.

Phuong, T. T.; Hien, L. T.; Vinh, N. D.; et al. 2022. An Eye Blink detection technique in video surveillance based on Eye Aspect Ratio. In *Proc. 24th International Conference on Advanced Communication Technology (ICACT)*, 534–538.

Radlak, K.; Bozek, M.; and Smolka, B. 2015. Silesian deception database: Presentation and analysis. In *Proc. ACM Multimodal Deception Detection Workshop (MDDW)*, 29–35.

Rosenfield, M. 2011. Computer vision syndrome: a review of ocular causes and potential treatments. *Ophthalmic and Physiological Optics*, 31(5): 502–515.

Shi, D.; Wei, X.; Li, L.; Ren, Y.; and Tan, W. 2022. End-to-End Multi-Person Pose Estimation With Transformers. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11069–11078.

Shi, D.; Zhong, Y.; Cao, Q.; Ma, L.; Li, J.; and Tao, D. 2023. TriDet: Temporal Action Detection with Relative Boundary Modeling. *arXiv e-prints*, arXiv:2303.07347.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Soukupová, T.; and Cech, J. 2016. Real-Time Eye Blink Detection using Facial Landmarks. In *Proc. Computer Vision Winter Workshop (CVWW)*, 1–8.

Tan, M.; Zhuang, Z.; Chen, S.; Li, R.; Jia, K.; Wang, Q.; and Li, Y. 2024. EPMF: Efficient Perception-Aware Multi-Sensor Fusion for 3D Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 8258–8273.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Wu, J.; Yarram, S.; Liang, H.; Lan, T.; Yuan, J.; Eledath, J.; and Medioni, G. 2022. Efficient video instance segmentation via tracklet query and proposal. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 959–968.

Yang, J.; and Kun, D. 2023. YOWOv2: A Stronger yet Efficient Multi-level Detection Framework for Real-time Spatio-temporal Action Detection. *arXiv preprint arXiv:2302.06848*.

Yang, L.; Zheng, Z.; Han, Y.; Cheng, H.; Song, S.; Huang, G.; and Li, F. 2024. DyFADet: Dynamic Feature Aggregation for Temporal Action Detection. *arXiv e-prints*, arXiv:2407.03197.

Zeng, W.; Xiao, Y.; Hu, G.; Cao, Z.; Wei, S.; Fang, Z.; Zhou, J. T.; and Yuan, J. 2023a. Eyelid's Intrinsic Motion-Aware Feature Learning for Real-Time Eyeblink Detection in the Wild. *IEEE Transactions on Information Forensics and Security*, 18: 5109–5121.

Zeng, W.; Xiao, Y.; Wei, S.; Gan, J.; Zhang, X.; Cao, Z.; Fang, Z.; and Zhou, J. T. 2023b. Real-time Multi-person Eyeblink Detection in the Wild for Untrimmed Video. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13854–13863.

Zhang, C.; Wu, J.; and Li, Y. 2022. ActionFormer: Localizing Moments of Actions with Transformers. *arXiv e-prints*, arXiv:2202.07925.

Zhao, H.; Wang, Y.; Lu, W.; Yi, Z.; Liu, J.; and Gong, M. 2025. Real-time dual-eye collaborative eyeblink detection with contrastive learning. *Pattern Recognition*, 162: 111440.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; and Chen, J. 2023. DETRs Beat YOLOs on Real-time Object Detection. arXiv:2304.08069.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proc. International Conference on Learning Representations (ICLR)*.