Preprints.org

Article

Not peer-reviewed version

YOLOv8-Based Fire and Smoke Detection Using CLIP-Guided Dual-Teacher Knowledge Distillation

Md. Jahidul Islam *, Maruf Al Hossain Prince *, Mahamudul Hasan Masum, Sabrina Jahan, Abu Saleh Musa Miah, Muhammad Nazrul Islam, Fahmid Al Farid *, Hezerul Abdul Karim *

Posted Date: 21 October 2026

Keywords: fire detection; smoke detection; YOLOv8; knowledge distillation; CLIP; GhostConv; efficient channel attention; edge deployment; cross-dataset generalisation

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

YOLOv8-Based Fire and Smoke Detection Using CLIP-Guided Dual-Teacher Knowledge Distillation

Md. Jahidul Islam [1,*], Maruf Al Hossain Prince [2,3,*], Mahamudul Hasan Masum [4], Sabrina Jahan [2], Abu Saleh Musa Miah [5], Muhammad Nazrul Islam [6], Fahmid Al Farid [7,*] and Hezerul Abdul Karim [8,*]

[1] Department of Computer Science and Engineering, Chandpur Science and Technology University, Chandpur, Bangladesh
[2] Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology, Nilphamari, Bangladesh
[3] Department of Computer Science and Engineering, Islamic University of Technology, Gazipur, Bangladesh
[4] AI & Cybersecurity, University of Gloucestershire, Birmingham, UK
[5] Department of Computer Science and Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh
[6] Department of Computer Science and Engineering, Military Institute of Science and Technology, Dhaka, Bangladesh
[7] Faculty of Computer Science and Informatics, Berlin School of Business and Innovation, Berlin, Germany
[8] Centre for Image and Vision Computing (CIVC), COE for Artificial Intelligence, Faculty of Artificial Intelligence and Engineering (FAIE), Multimedia University, Cyberjaya 63100, Malaysia
* Correspondence: jahid@cse.cstu.ac.bd (M.J.I.); marufhossain@iut-dhaka.edu (M.A.H.P.); fahmid.alfarid@berlinsbi.com (F.A.F.); hezerul@mmu.edu.my (H.A.K.)

Highlights

What are the main findings?

- CLIP-guided dual-teacher distillation yields a 2.3M-parameter fire and smoke detector that keeps 91.6% of its larger teacher's accuracy and runs at 375.6 frames per second.
- Every detector tested, including uncompressed baselines, falls from 71–79% to at most 3.3% mAP@50 on independent fire datasets, and CLIP confounder prompts do not reduce false alarms from steam, fog, or sunsets.

What are the implications of the main findings?

- High accuracy on a single fire dataset does not predict field performance, so camera-based fire detectors should be validated on independent datasets before deployment.
- Compact detectors already meet real-time edge requirements; the remaining barrier to reliable early warning is robustness, which calls for multi-source training data and region-level vision-language supervision.

Abstract

Deep learning fire detectors are usually optimised for in-domain accuracy, while their cross-dataset generalisation, which is vital for real-world deployment, remains poorly understood. We propose a dual-teacher knowledge distillation framework that compresses a YOLOv8m teacher (25.84M parameters, 78.7 GFLOPs) into a GhostConv-enhanced YOLOv8n student with efficient channel attention (2.298M parameters, 6.2 GFLOPs) through spatial feature mimicry and CLIP ViT-B/32 semantic guidance. The distilled student achieves 72.0% mAP@50 on the D-Fire validation set (four-seed mean), retaining 91.6% of teacher performance with 11.2× fewer parameters, and 71.5% mAP@50 on the held-out test set at 375.6 FPS with TensorRT on an RTX 5080, exceeding the closest distillation baseline by 9.3 percentage points. Ablations show that CLIP semantic distillation alone reaches 72.6% mAP@50, surpassing the combined loss and revealing the importance of the transfer mechanism under architectural mismatch. Across four external benchmarks, however, all eight tested models fall from 71–79% mAP@50 on D-Fire to $\leq$3.3%, and CLIP-based confounder prompts do not reduce false positives relative to detection-only training. This first multi-benchmark generalisation analysis of a sub-5M-parameter fire detector shows that distillation delivers edge-deployment gains but leaves cross-domain robustness unresolved.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

Keywords: fire detection; smoke detection; YOLOv8; knowledge distillation; CLIP; GhostConv; efficient channel attention; edge deployment; cross-dataset generalisation

1. Introduction

Fire and smoke pose severe threats to life, property, and ecological systems. Although deep learning object detectors, particularly YOLO-family models [1,2], achieve high accuracy on fire detection benchmarks [3–5], real-world deployment imposes two competing constraints: (i) real-time operation on edge hardware, and (ii) reliable generalisation to unseen environments. These objectives are usually pursued independently: compressed models are evaluated only on training-set benchmarks, while generalisation work ignores parameter budgets.

Existing models report 68–86% mAP@50 on D-Fire, yet cross-dataset evaluation is rare. Of the fire detection papers surveyed, only two report results on more than one benchmark. Boroujeni et al. [6] document catastrophic cross-dataset collapse across all surveyed models (e.g., FLAME1 drops from 98.49% to 23.95% on BoWFire), confirming that within-dataset accuracy is not diagnostic of deployment robustness.

Knowledge distillation (KD) [7] is a natural compression strategy, but only two prior fire detection papers apply it to ground-level benchmarks: Pruned-KD [8] achieves 62.7% mAP@50 on D-Fire (as re-evaluated in [3]) via single-teacher YOLO KD, and FFYOLO [9] applies GhostConv with single-teacher KD for forest fire only. Further KD work exists for aerial fire detection [10,11], but none evaluates on D-Fire or uses a vision-language model (VLM) as a semantic teacher. Cross-dataset evaluation is absent from all prior fire KD papers.

CLIP [12], specifically the `laion2b_s34b_b79k` checkpoint pretrained on 2.3 billion image-text pairs (LAION-2B [13]), provides semantically rich class-boundary knowledge unavailable from detection-domain data. We hypothesised that CLIP's confounder-aware soft labels would (i) improve within-dataset precision and (ii) improve cross-dataset robustness. Both hypotheses are empirically tested in Sections 3 and 4.

This paper makes four contributions. First, the first VLM-guided dual-teacher KD framework for fire detection is presented. Second, a GhostConv-YOLOv8n+ECA student is described that achieves 11.2× compression at 91.6% mAP@50 retention on D-Fire validation (four-seed mean; held-out test: 71.5%, 90.9% retention). Third, a six-configuration ablation isolates the contributions of spatial KD, CLIP semantic KD, GhostConv, and ECA individually, with multi-seed confirmation and an α-sweep for the two primary configurations. Fourth, the first systematic multi-benchmark cross-dataset generalisation evaluation for a sub-5M-parameter fire detector is reported, quantifying collapse severity and CLIP confounder failure as open problems.

The methodological novelty of this work is not any single component—GhostConv, ECA, and CLIP are each established independently—but the decomposable, cross-modal dual-teacher protocol built from them and the diagnostic use to which it is put. Prior dual-teacher KD systems in fire and wildfire detection pair two same-modality vision teachers: FireNet-KD [10] distils a Swin Transformer and a CNN jointly into one student, and Keerthinathan et al. [11] pair a Grounding DINO detector with a YOLO Nano student for aerial plume detection; both fix the two teachers as a single joint objective and report only aggregate accuracy, without a mechanism to isolate either teacher's marginal contribution. The present framework instead pairs a spatial CNN teacher with a vision-language teacher operating in a fundamentally different representational space (image-text similarity rather than pixel-level feature maps), and implements the two supervisory signals as independently switchable loss terms (Table 3) rather than a fixed combination. This design is what makes the α-sweep in Section 4.2 possible: to our knowledge, no prior fire-detection KD paper quantifies the marginal and interactive contribution of each teacher signal, as opposed to reporting only the accuracy of the already-combined system. The methodological contribution is therefore the ablatable dual-teacher

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

protocol itself, and the architecture-mismatch diagnosis it enables (Section 4.2), rather than the union of pre-existing modules.

1.1. Related Work

1.1.1. Fire Detection

Lightweight fire detection work splits along two largely independent axes: architectural capacity and supervision signal. On the capacity axis, larger models buy accuracy through richer feature aggregation—YOLOGX [3] reaches 80.9% mAP@50 on D-Fire (6.2M params) via asymptotic feature pyramids and cross-level dual attention, and TFNet [5] reaches 81.2% at ∼30M through a Transformer-CNN hybrid neck—whereas the sub-5M class relies on cheaper compression modules: CCi-YOLOv8n [4] (78.5%, 3.8M) uses CARAFE up-sampling and context-guided feature fusion, and LD-YOLO [14] applies GhostConv with a spatial context awareness module to forest fire detection at 3–4M params. Notably, YOLOv11n [15] breaks the expected size-accuracy trade-off, leading the sub-5M class at 85.9% with only 2.6M params—evidence that architectural design choice, not raw compression ratio, is the dominant factor. On the orthogonal supervision axis, CLIP prompt tuning improves fire alarm classification by $\geq$12.45% [16] and transfer learning achieves competitive accuracy at reduced parameter budgets in resource-constrained settings [17], but both operate on whole-image classification rather than detection, and neither is paired with architectural compression. LD-YOLO is the only prior work to combine GhostConv with an attention-like module, yet it targets a single forest-fire domain and uses no semantic supervision. No existing lightweight detector jointly combines GhostConv-based compression with vision-language semantic supervision—the gap this work addresses.

Cross-dataset generalisation is rarely evaluated. Han et al. [18] identified critical inconsistencies in existing public fire and smoke detection datasets—stemming from heterogeneous labelling conventions and limited scene diversity—and introduced a standardised Multi-Scene Fire and Smoke Detection Benchmark (MS-FSDB; 12,586 images, 2731 scenes) with cross-dataset evaluation protocols to expose this problem. Boroujeni et al. [6] surveyed 20 years of fire and smoke datasets and documented catastrophic cross-dataset collapse across all surveyed models, confirming that within-dataset accuracy is not diagnostic of deployment robustness. Yan et al. [19] showed that unsupervised domain adaptation [20] yields only limited generalisation improvement, confirming the severity of the distribution mismatch.

For edge and aerial deployment, Keerthinathan et al. [11] distil a Grounding DINO teacher into YOLO Nano students for real-time aerial wildfire detection on drones, achieving substantial throughput gains at modest accuracy cost. DSS-YOLO [21] targets real-time mobile deployment, reporting 89.5% mAP@50 on mytest-hrswj after in-distribution training. The persistent challenge is that lightweight architectures trained on narrow data distributions generalise poorly to novel environments—a gap that neither larger model capacity nor stronger supervision has resolved in the existing literature.

1.1.2. Knowledge Distillation

KD methods are broadly categorised as response-based (mimicking final predictions), feature-based (mimicking intermediate representations), and relation-based (preserving inter-sample relationships). Hinton et al. [7] showed that temperature-scaled soft labels alone "often suffice" as a response-based signal. FitNets [22] introduced feature-level mimicry, effective when teacher and student architectures are structurally compatible. CRD [23] demonstrated empirically that contrastive, relation-based objectives outperform feature-level mimicry across heterogeneous architecture pairs; ReviewKD [24] identifies architectural alignment as a prerequisite for effective spatial distillation—both findings are directly relevant to the C2f→C3Ghost mismatch in this work.

For object detection specifically, FGD [25] and PKD [26] refine feature distillation through foreground-selective attention weighting and Pearson-normalised magnitude matching, respectively. CrossKD [27] advances detection KD by routing student head features through the teacher detection head before computing mimicry losses, thereby decoupling annotation supervision from distillation

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

supervision; on MS-COCO, this raises GFL ResNet-50 mAP from 40.2 to 43.7 and generalises to heterogeneous-backbone pairs. Hybrid CNN-Transformer KD has demonstrated effective lightweight compression in related classification tasks [28].

Among fire-specific KD papers, FireNet-KD [10] uses dual vision-based teachers (Swin Transformer + CNN) on the FASDD dataset with no VLM and no cross-dataset evaluation; Keerthinathan et al. [11] distil a Grounding DINO teacher into YOLO Nano students for aerial wildfire detection, introducing a third *plume* class but evaluating only on domain-specific aerial data; and a hierarchical multi-task KD classifier [29] includes confounder supervision in a classification-only setting. None of these works employs a vision-language model as a distillation teacher. Both FireNet-KD [10] and Keerthinathan et al. [11] are, like this work, dual-teacher systems, but each pairs two teachers from the same visual modality (Swin+CNN; Grounding DINO+YOLO) trained as one fixed joint objective; neither reports an ablation isolating each teacher's individual contribution. The cross-modal (vision + language) pairing used here, combined with independently switchable spatial and semantic loss terms, is what distinguishes the present dual-teacher design from these prior systems and is what enables the teacher-interaction analysis in Section 4.2.

1.1.3. CLIP for Dense Prediction

ViLD [30], Ma et al. [31], and VL2Lite [32] established the paradigm of distilling VLM knowledge into lightweight detection students for open-vocabulary recognition. Beyond open-vocabulary detection, WinCLIP [33] demonstrated that window-based CLIP feature aggregation—composing state words and prompt templates over sliding windows—achieves 91.8% AUROC zero-shot anomaly classification on MVTec-AD, establishing CLIP as a viable teacher for domain-specific, safety-critical visual inspection without task-specific training data. This precedent motivates using CLIP soft labels as semantic supervision for fire/confounder discrimination in our framework. Critically, DenseCLIP [34] and CORA [35] showed that global-image CLIP representations cannot provide reliable region-level supervision for dense prediction without explicit region prompting—directly explaining the confounder failure reported in Section 4.

1.1.4. Efficient CNN Design

GhostNet [36] replaces standard convolutions with cheap linear operations (intrinsic features + cheap ghost features), achieving ∼50% MAC reduction at comparable accuracy; the C3Ghost module extends this to CSP bottleneck blocks. ECA-Net [37] delivers Squeeze-and-Excitation [38]-equivalent channel attention at negligible cost ($\leq$9 parameters), demonstrating that cross-channel interaction does not require full dimensionality reduction. Both modules have been applied independently to fire detection—GhostConv in LD-YOLO [14] and FFYOLO [9]—but were never jointly ablated in a controlled study, leaving their combined contribution and the sign of their interaction uncharacterised. This work provides the first controlled ablation that isolates GhostConv and ECA individually within an identical training protocol, directly attributing the 23.5% parameter reduction to GhostConv and the precision improvement to ECA.

2. Materials and Methods

2.1. System Overview

The framework combines three components active only during training: a frozen YOLOv8m spatial teacher [2], a frozen CLIP ViT-B/32 semantic teacher, and a trainable GhostConv-YOLOv8n student with learnable projection heads. At inference, all KD components are discarded; only the student weights are deployed.

At each training step, the spatial teacher provides feature maps at three pyramid levels (P3/P4/P5) via forward hooks; the semantic teacher computes per-image similarity scores against seven precomputed text embeddings; and the student produces detection outputs and intermediate features. Three loss terms back-propagate through the student and projectors only (Figure 1).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

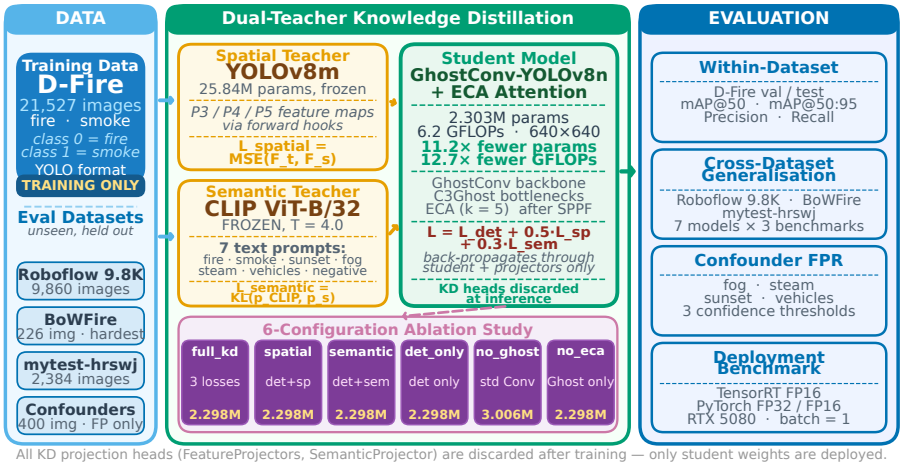


Figure 1. Overview of the FireLite-KD framework. *Left:* D-Fire (21,527 images) is the sole training set; four evaluation benchmarks (Roboflow 9.8K, BoWFire, mytest-hrswj, and confounders) are reserved exclusively for evaluation. *Centre:* dual-teacher knowledge distillation pipeline (training phase only)—frozen YOLOv8m provides P3/P4/P5 feature maps via forward hooks ($\mathcal{L}_{\text{spatial}}$, MSE); frozen CLIP ViT-B/32 processes 224×224 crops and produces 7-prompt similarity scores at T=4.0 ($\mathcal{L}_{\text{semantic}}$, KL divergence); all gradients back-propagate through the GhostConv-YOLOv8n student and projection heads only ($\mathcal{L} = \mathcal{L}_{\text{det}} + \alpha\,\mathcal{L}_{\text{spatial}} + 0.3\,\mathcal{L}_{\text{semantic}}$); all KD heads are discarded at inference. Six ablation configurations (`full_kd`, `spatial_only`, `semantic_only`, `det_only`, `no_ghost`, `no_eca`) systematically isolate each contribution. *Right:* evaluation protocol—within-dataset mAP@50/50:95 on D-Fire, cross-dataset generalisation across three external benchmarks (7 models), confounder false-positive rate at three confidence thresholds, and TensorRT deployment benchmarking on an NVIDIA RTX 5080.

2.2. Student Architecture: GhostConv-YOLOv8n with ECA

All backbone `Conv` blocks are replaced with `GhostConv` and all `C2f` modules with `C3Ghost`. GhostConv computes $C/2$ intrinsic features via standard convolution and generates the remaining $C/2$ ghost features via cheap depthwise convolution:

$$\mathbf{Y}^{\text{int}} = \text{Conv}(\mathbf{X};\,\mathbf{W}_s), \quad \mathbf{Y}^{\text{ghost}} = \text{DWConv}(\mathbf{Y}^{\text{int}};\,\mathbf{W}_d), \quad \mathbf{Y} = [\mathbf{Y}^{\text{int}} \,\|\, \mathbf{Y}^{\text{ghost}}]. \tag{1}$$

An ECA module is inserted after SPPF. For feature map $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$,

$$\mathbf{s} = \text{GAP}(\mathbf{X}), \quad \mathbf{a} = \sigma(\text{Conv1D}(\mathbf{s},\,k{=}5)), \quad \text{ECA}(\mathbf{X}) = \mathbf{a} \otimes \mathbf{X}. \tag{2}$$

ECA adds exactly 5 trainable parameters. The resulting student has **2,298,159 parameters** and **6.2 GFLOPs** at 640×640, compared to the teacher's 25,840,918 parameters and 78.7 GFLOPs—compression ratios of **11.24× (params)** and **12.69× (GFLOPs)**. Spatial KD hooks are registered at P3 (Layer 4, 64 ch), P4 (Layer 6, 128 ch), and P5 (Layer 10, 256 ch).

2.3. Spatial and Semantic Teachers

YOLOv8m is selected as the spatial teacher (trained to 78.65% mAP@50 on D-Fire). No directly comparable published YOLOv8m result on D-Fire was found; the closest reference point is the unspecified-variant YOLOv8 baseline in [3] (79.46% mAP@50, 3.0M parameters—consistent with YOLOv8n rather than our 25.84M-parameter YOLOv8m), whose own improved YOLOGX architecture reaches 80.9% (Table 7). Learnable FeatureProjectors (1×1 Conv + BN + ReLU) bridge channel-dimension mismatches between student and teacher at each pyramid level and are discarded after training; Table 1 specifies the per-scale dimensions.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

Table 1. FeatureProjector channel dimensions per FPN scale. Projectors are discarded after KD training.

Scale	Resolution	Student Layer	Student Ch	Teacher Ch
P3	80×80	4 (C3Ghost)	64	192
P4	40×40	6 (C3Ghost)	128	384
P5	20×20	10 (ECA)	256	768

CLIP ViT-B/32 [12,39] is used frozen in inference mode. Seven text prompts encode fire/smoke targets and four explicit confounders (Table 2). Embeddings are pre-computed and cached. For each training image, cosine similarities against all seven embeddings are temperature-scaled at $T = 4.0$ to produce soft target probabilities. A learnable SemanticProjector maps the student's pooled P5 map to a 7-dim distribution aligned with CLIP outputs; it is also discarded after training. CLIP image processing operates through a pipeline entirely separate from the YOLO detection pipeline: each training image is independently resized to 224×224 pixels and normalised using CLIP's training statistics (means 0.481, 0.458, 0.408; stds 0.269, 0.261, 0.276 per channel), as opposed to the 640×640 resolution and ImageNet-derived normalisation used for the YOLO pipeline. Maintaining independent preprocessing pipelines for the two teacher signals prevents cross-contamination of their respective feature spaces.

Table 2. CLIP text prompts used for semantic distillation.

Index	Prompt	Role
1	"a photograph of a real fire with visible flames"	Fire positive
2	"a photograph of thick smoke from a burning fire"	Smoke positive
3	"a photograph of a sunset or sunrise with orange sky"	Confounder
4	"a photograph of fog or mist in the atmosphere"	Confounder
5	"a photograph of industrial steam or exhaust"	Confounder
6	"a photograph of a red or orange coloured vehicle"	Confounder
7	"a normal outdoor scene without any fire or smoke"	Negative

2.4. Loss Function

The student is optimised with a weighted sum of a detection term and two distillation terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \alpha \cdot \mathcal{L}_{\text{spatial}} + \beta \cdot \mathcal{L}_{\text{semantic}}, \tag{3}$$

where $\mathcal{L}_{\text{det}}$ is the standard YOLOv8 loss (CIoU + BCE classification + DFL [40]). The spatial term $\mathcal{L}_{\text{spatial}}$ is the MSE between projected student and frozen teacher features at P3/P4/P5:

$$\mathcal{L}_{\text{spatial}} = \frac{1}{3} \sum_{l} \frac{1}{C_l H_l W_l} \left\| \hat{\mathbf{F}}_s^{(l)} - \mathbf{F}_t^{(l)} \right\|_F^2, \tag{4}$$

and the semantic term $\mathcal{L}_{\text{semantic}}$ is the KL divergence between the student projector output and the CLIP soft labels:

$$\mathcal{L}_{\text{semantic}} = \text{KL}(\mathbf{p}_{\text{CLIP}} \,\|\, \hat{\mathbf{p}}_s). \tag{5}$$

This forward-KL direction—CLIP as reference, student as approximant—minimises the expected log-ratio under the CLIP distribution, penalising the student for placing low probability mass wherever CLIP assigns high confidence, which is the intended behaviour for prompt-guided semantic supervision.

The hyperparameters are $\alpha = 0.5$, $\beta = 0.3$, and $T = 4.0$ for the primary `full_kd` ablation run. A post hoc α-sweep over $\{0.1, 0.2, 0.3, 0.5\}$ identified α=0.2 as the operating point that best recovers combined-loss performance; full results are discussed in Section 4.2. Six ablation configurations were designed to isolate the contribution of each framework component, as detailed in Table 3.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

Table 3. Six ablation configurations used to isolate each framework component.

Config	$\mathcal{L}_{\text{det}}$	$\mathcal{L}_{\text{spatial}}$	$\mathcal{L}_{\text{semantic}}$	Architecture
`full_kd` [1]	✓	✓(α=0.5)	✓(β=0.3)	GhostConv + ECA
`spatial_only`	✓	✓	—	GhostConv + ECA
`semantic_only`	✓	—	✓	GhostConv + ECA
`det_only`	✓	—	—	GhostConv + ECA
`no_ghost`	✓	✓	✓	Standard Conv + ECA
`no_eca`	✓	✓	✓	GhostConv, no ECA

[1] `full_kd` uses α=0.5 for ablation consistency across all six configurations; the α-sweep (Section 4.2, Table 12) identifies α=0.2 as the empirically optimal value for deployment.

2.5. Datasets

D-Fire [41] (training + within-dataset evaluation): 21,527 images, YOLO bounding-box annotations, two classes (fire = 0, smoke = 1), train/val/test split 14,122/3099/4306. All training is restricted to D-Fire only; external datasets are reserved exclusively for evaluation.

Roboflow Fire-Smoke 9.8K [42] (cross-dataset evaluation #1): 9860 community-contributed images covering diverse indoor/outdoor scenarios, native YOLO format.

BoWFire [43] (cross-dataset evaluation #2—hardest): 226 images from broadcast footage. Ground-truth boxes are derived via connected-components analysis of binary segmentation masks (min. area 200 px^2), yielding 586 boxes across 119 fire-positive images; the 107 non-fire images have empty labels.

mytest-hrswj [44] (cross-dataset evaluation #3): publicly available open-source dataset on Roboflow Universe (Mytyolov8, May 2024), 2384 images, YOLO bounding-box format, fire class only (smoke: N/A). DSS-YOLO [21] reports 89.5% mAP@50 after in-distribution training on this benchmark.

Confounder suite (false-positive evaluation only): 400 images, 100 per category (fog, steam, sunset, vehicles), all with empty labels, never included in any training split.

Dataset preprocessing. Four preprocessing steps were applied to ensure YOLO framework compatibility across all datasets. First, a class-identity audit confirmed that all three YOLO-format datasets share the convention class 0 = fire, class 1 = smoke; the single-class mytest-hrswj dataset was verified to use class 0 = fire exclusively, precluding silent label-assignment errors that would corrupt training without raising an exception. Second, BoWFire's pixel-level binary segmentation masks were converted to YOLO axis-aligned bounding boxes using connected-components analysis (`connectedComponentsWithStats`, 8-connectivity) with a minimum component area of 200 px^2 to suppress noise artefacts, yielding 586 bounding boxes distributed across 119 fire-positive images; the 107 non-fire images received empty label files. Third, the two Roboflow datasets were distributed with malformed `data.yaml` configurations specifying a relative validation path (`../valid/images`) with no explicit `path:` root key, causing Ultralytics to resolve paths one directory level above the dataset root; both files were corrected by adding absolute `path:` keys and removing the erroneous prefix. Fourth, 400 confounder images were organised into the standard YOLO `images/labels/` hierarchy with empty label files, since Ultralytics derives label paths by substituting `/images/` with `/labels/` in each image file path and cannot locate labels in flat directory layouts. Following all corrections, nine `data.yaml` configurations were independently verified by checking directory existence, image counts, and image-to-label correspondence; all nine passed without error.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

2.6. Implementation

All experiments were run on an NVIDIA RTX 5080 (16 GB GDDR7) with an AMD Ryzen 9 9900X CPU and 32 GB DDR5 memory, using PyTorch 2.10.0+cu130 under WSL2. Table 4 summarises the training hyperparameters used for the spatial teacher and for all student/knowledge-distillation configurations.

Table 4. Training hyperparameters for the spatial teacher and all student/KD configurations.

Hyperparameter	Teacher	Student/KD
Epochs	100 (best @ ep. 76)	100
Optimiser	AdamW [45]	AdamW
Learning rate	1×10^{-3}	1×10^{-3}
LR schedule	Cosine + 3 warmup epochs	CosineAnnealingLR
Weight decay	5×10^{-4}	5×10^{-4}
Momentum β_1	0.937 [1]	0.9
Batch size	16	24
Image size	640	640
Augmentation	Ultralytics default	Albumentations (train only)

[1] Teacher β_1=0.937 follows the Ultralytics YOLOv8 default; student β_1=0.9 follows the PyTorch AdamW default.

Student training applies the following Albumentations [46] transforms to training images only: random horizontal flip (p=0.5), HSV jitter (hue ±0.015, saturation ±0.7, value ±0.4), random scale (±0.5), and mosaic augmentation (p=1.0). No augmentation is applied at evaluation.

All evaluations use raw unaugmented images at imgsz=640. TensorRT (TRT) engines are exported at FP16; FPS is measured over 1000 iterations (500 warm-up) at batch size 1.

3. Results

3.1. Within-Dataset Performance (D-Fire Validation)

Four-seed training of `full_kd` and `semantic_only` (original run plus three independent seeds) confirms stable convergence across random initialisations (Figure 2). Total training loss trajectories are equally consistent across seeds (Figure 3), confirming that the mAP@50 gap is not attributable to optimisation instability or differential loss convergence. The `semantic_only` > `full_kd` ordering is statistically robust: the mean difference of Δ=0.0052 mAP@50 exceeds the sum of one standard deviation for both models (0.0015 + 0.0031 = 0.0046), establishing non-overlapping 1σ intervals. All student models are precision-heavy ($P > R$) relative to the teacher—a conservative detection profile desirable for fire detection, where false alarms trigger costly evacuations (Table 5, Figure 4).

Table 5. D-Fire validation results (imgsz=640, all-class mean). Best student value per column in bold.

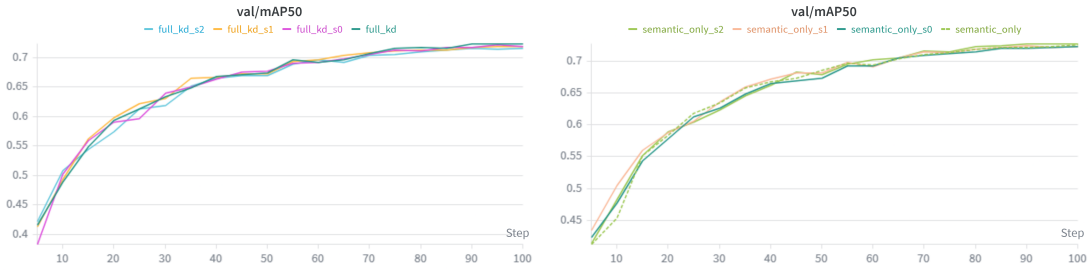
Model	Params	GFLOPs	mAP@50	mAP@50:95	P	R	F1
YOLOv8m Teacher	25.84M	78.7	0.7865	0.4720	0.7650	0.7370	0.7508
`no_ghost`	3.006M	8.1	**0.7361**	**0.4256**	**0.7480**	**0.6584**	**0.7003**
`semantic_only` [1]	2.298M	6.2	0.7255±0.0015	0.4116±0.0010	0.7434±0.0079	0.6567±0.0045	0.6963±0.0038
`spatial_only`	2.298M	6.2	0.7248	0.4102	0.7356	0.6530	0.6918
`det_only`	2.298M	6.2	0.7205	0.4122	0.7378	0.6499	0.6911
`full_kd` [1]	2.298M	6.2	0.7203±0.0031	0.4094±0.0018	0.7395±0.0057	0.6510±0.0045	0.6926±0.0035
`no_eca`	2.298M	6.2	0.7156	0.4107	0.7182	0.6534	0.6842

[1] Four-seed mean ± population s.d. (original run + three independent seeds); all other rows are single-seed.

Per-class decomposition (Table 6). Smoke recall consistently lags fire recall (∼0.58 vs. ∼0.72), reflecting smoke's greater morphological complexity. Backbone structural differences across the three architecture variants are illustrated in Figure 5. Training dynamics for all six ablation configurations—both validation mAP@50 trajectories and total loss curves—are shown in Figure 6; all configurations converge smoothly without instability, confirming that the mAP@50 spread in Table 5 reflects genuine architectural and loss differences rather than optimisation variance.

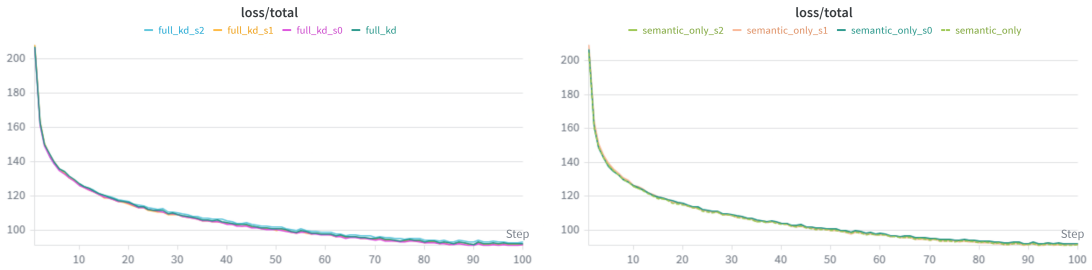
© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1



(**a**) (**b**)

Figure 2. Validation mAP@50 convergence over 100 training epochs for three independent seeds. (**a**) `full_kd` seeds s0/s1/s2 alongside the original single-seed run (dark green); all four curves are tightly clustered (s.d. $\leq$ 0.0024), confirming reproducible convergence to 0.719±0.002. (**b**) `semantic_only` seeds s0/s1/s2; even tighter clustering (s.d. $\leq$ 0.0021) converging to 0.725±0.002. The consistent ∼0.006 gap between `semantic_only` and `full_kd` across seeds demonstrates that the CLIP-only configuration reliably outperforms the combined loss under C2f→C3Ghost architectural mismatch.



(**a**) (**b**)

Figure 3. Total training loss over 100 epochs for three independent seeds. (**a**) `full_kd` seeds s0/s1/s2 alongside the original single-seed run; all four curves follow an identical exponential decay profile with near-perfect overlap throughout training. (**b**) `semantic_only` seeds s0/s1/s2; the clustering is even tighter than for `full_kd`, consistent with the simpler two-term loss surface ($\mathcal{L}_{\text{det}} + \beta\mathcal{L}_{\text{semantic}}$ versus the three-term combined loss). The loss stability across seeds rules out gradient instability as a cause of the mAP@50 gap between the two configurations.

Table 6. Per-class D-Fire validation results.

Model	Fire mAP@50	Fire Recall	Smoke mAP@50	Smoke Recall
Teacher	0.845	0.797	0.731	0.676
YOLOv8n$_{\text{base}}$	0.834	0.781	0.711	0.632
`no_ghost`	0.799	0.741	0.673	0.576
`semantic_only`	0.783	0.724	0.670	0.584
`full_kd`	0.783	0.733	0.665	0.574
`spatial_only`	0.777	0.716	0.672	0.590
`det_only`	0.781	0.724	0.660	0.575
`no_eca`	0.771	0.715	0.660	0.591

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

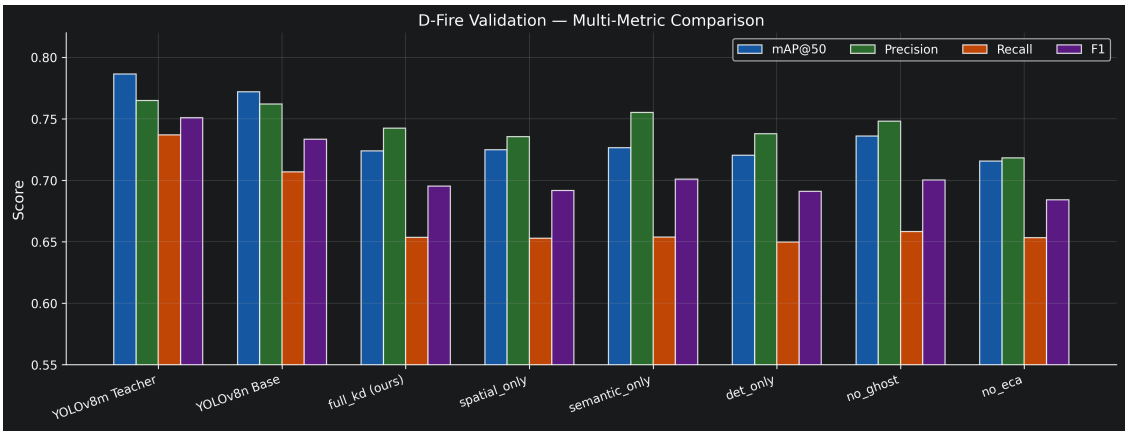


Figure 4. D-Fire validation multi-metric grouped bar chart (mAP@50, Precision, Recall, F1) for all seven models. `no_ghost` leads on all metrics; `semantic_only` achieves the highest mAP@50 among GhostConv+ECA students (0.726, four-seed mean).

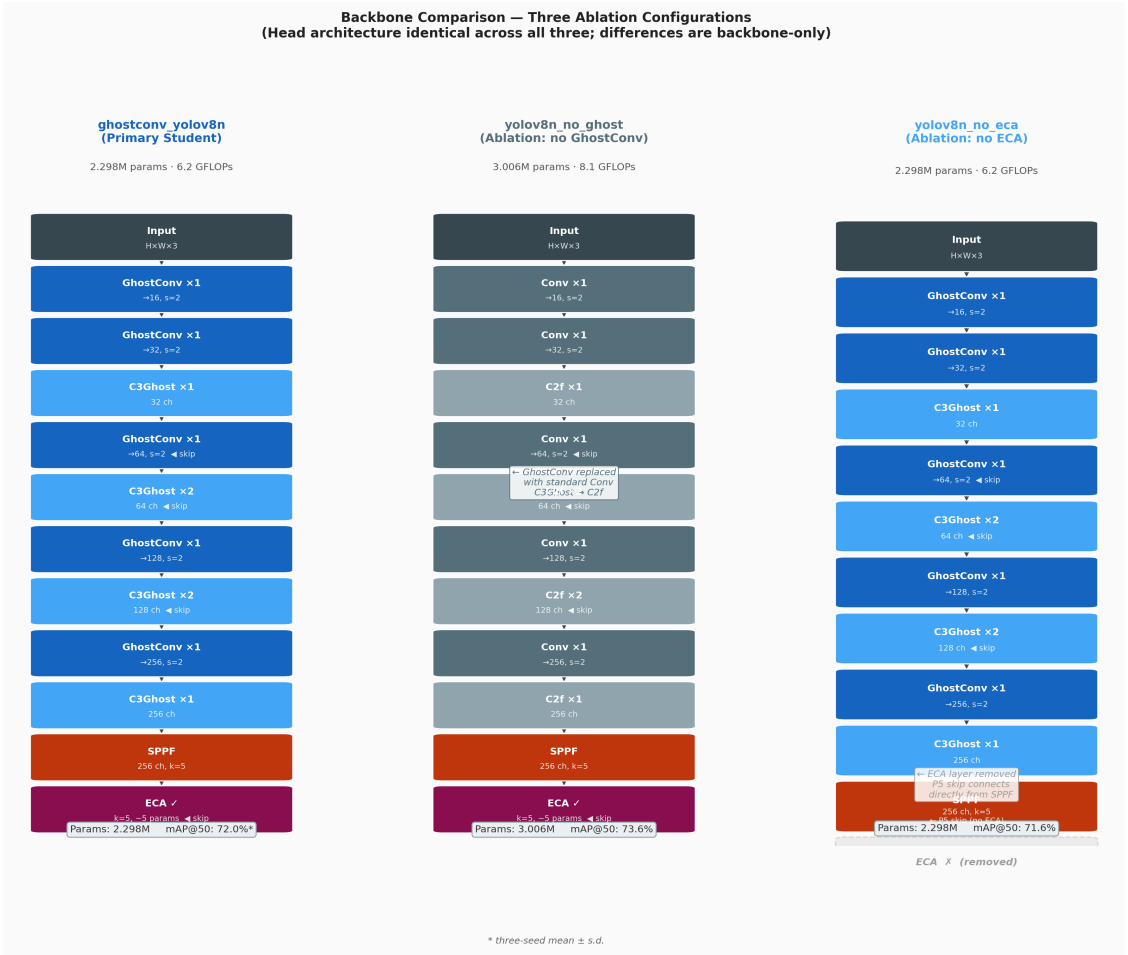
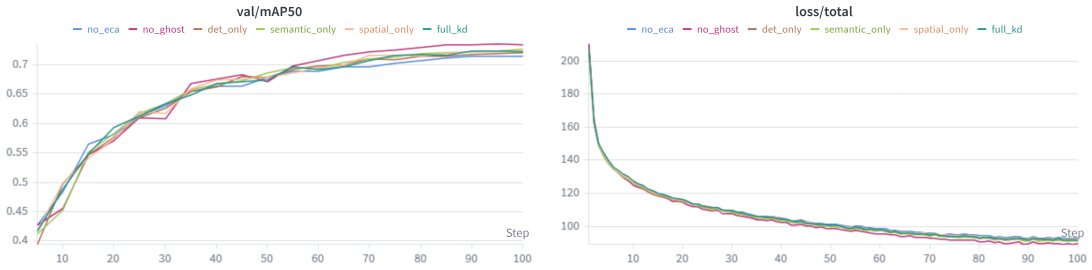


Figure 5. Backbone comparison across the three structural ablation variants. `ghostconv_yolov8n` (primary student, 2.298M params, 6.2 GFLOPs) replaces all Conv layers with GhostConv and all C2f modules with C3Ghost, with ECA (k=5) after SPPF. `yolov8n_no_ghost` (3.006M, 8.1 GFLOPs) substitutes standard Conv and C2f at every layer. `yolov8n_no_eca` (2.298M, 6.2 GFLOPs) is identical to the primary student with the ECA module removed. Quantitative mAP@50 results for all six configurations are in Table 5.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1



(**a**) (**b**)

Figure 6. Training dynamics for all six ablation configurations over 100 epochs. (**a**) Validation mAP@50 trajectories. `no_ghost` (standard Conv+ECA) consistently leads throughout training, converging to 0.736. The four GhostConv+ECA configurations (`full_kd`, `semantic_only`, `spatial_only`, `det_only`) cluster tightly in the 0.716–0.726 band, with `semantic_only` emerging as the best GhostConv student from epoch ∼60 onward. `no_eca` trails marginally, consistent with the −0.99 pp mAP@50 gap identified in Table 5 (`semantic_only`: 0.726 vs. `no_eca`: 0.716). (**b**) Total training loss curves. All six configurations follow near-identical exponential decay profiles with no instability, confirming that the performance differences are attributable to architecture and loss formulation rather than optimisation variance.

3.2. Comparison with Published State of the Art on D-Fire

Table 7 places our students alongside published D-Fire results.

Table 7. Comparison with published D-Fire results. Our models (bold) are evaluated on the D-Fire validation split under a fixed training protocol (100 epochs, D-Fire train split only, α=0.5, β=0.3, T=4.0); held-out test results are in Table 8. Literature entries use independently optimised training configurations (epochs, augmentation, and hyperparameters may differ); they establish the competitive landscape but are not equivalent-protocol comparisons.

Model	Params	mAP@50	mAP@50:95	Year
Published methods (independently optimised training configurations):				
YOLOv11n [15]	2.6M	85.9%	55.8%	2025
TFNet [5]	∼30M	81.2%	46.8%	2025
YOLOGX [3]	6.2M	80.9%	N/R	2024
CCi-YOLOv8n [4]	3.8M	78.5%	46.6%	2024
YOLOv11-CHBG [47]	∼1.8M	78.4%	—	2025
This work—D-Fire validation split, fixed training protocol:				
Ours: `no_ghost`	**3.006M**	**73.6%**	**42.6%**	2026
Ours: `semantic_only` [1]	**2.298M**	**72.6%±0.2%**	**41.2%±0.1%**	2026
Ours: `full_kd` [1]	**2.298M**	**72.0%±0.3%**	**40.9%±0.2%**	2026
Published KD and architecture baselines (independently optimised):				
YOLOv8n baseline [4]	3.2M	68.4%	38.2%	2024
Pruned-KD [8] [2]	2.64M	62.7%	48.8%	2021

[1] Four-seed mean ± s.d. [2] D-Fire result as re-evaluated in [3]; the original paper evaluated on a proprietary UAV dataset.

The `full_kd` student achieves +3.6 pp over the published YOLOv8n baseline (68.4%) at 28% fewer parameters and +9.3 pp over Pruned-KD at comparable parameter count (Figure 7); both margins are computed from the four-seed mean (72.0% ± 0.3%, Table 7). The Pruned-KD margin warrants qualification: Wang et al. [8] originally trained on a private 1844-image custom forest-fire dataset, not D-Fire; the 62.7% figure is a third-party re-evaluation on D-Fire reported in [3], making the +9.3 pp gap a cross-benchmark margin rather than a direct same-dataset head-to-head. Published state-of-the-art entries—YOLOv11n (85.9%), CCi-YOLOv8n (78.5%), TFNet (81.2%)—used independently optimised training configurations; their figures establish the competitive landscape but are not equivalent-protocol comparisons with our results. The primary contribution of this work is not state-of-the-art

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

accuracy on D-Fire but the compression ratio, the KD analysis, and the multi-benchmark evaluation: no sub-5M-parameter model in the literature has been evaluated simultaneously on D-Fire, Roboflow, BoWFire, and the mytest dataset.

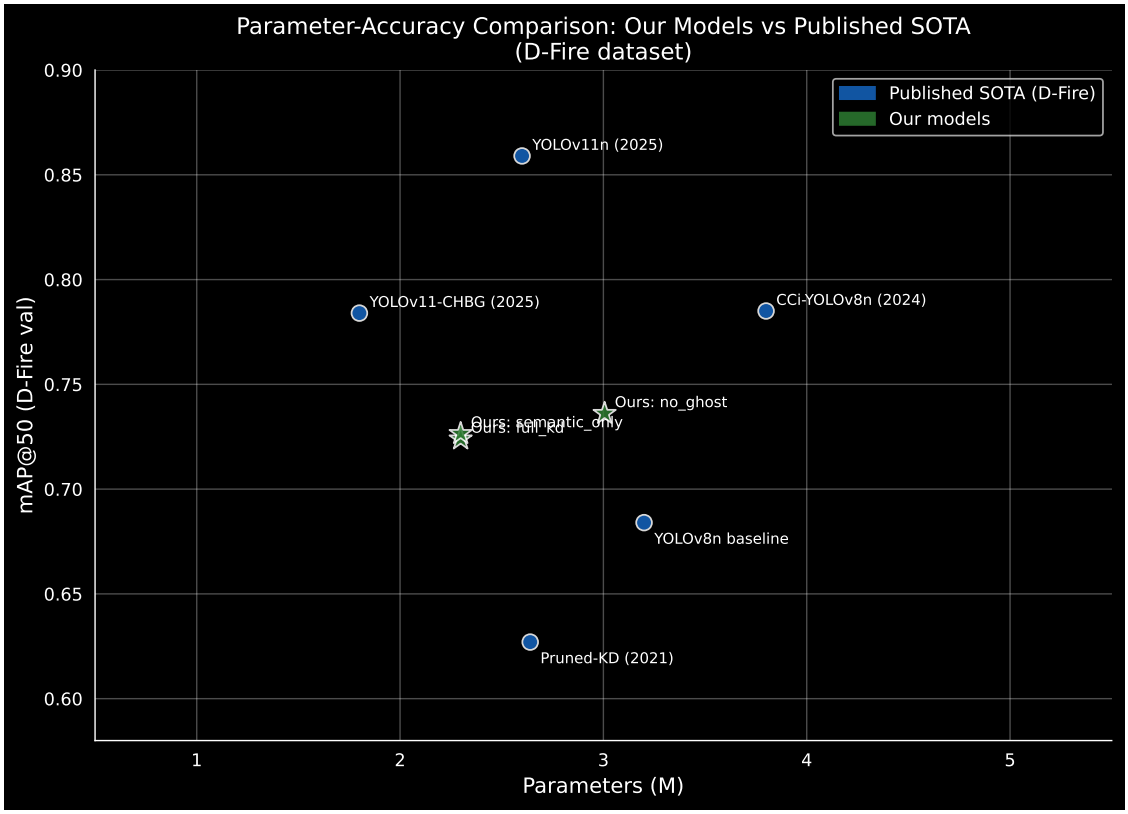


Figure 7. Parameter–accuracy trade-off: our configurations (starred markers, D-Fire validation split, fixed training protocol) versus published D-Fire methods (filled circles, independently optimised training configurations). Our 2.3M-parameter students (`semantic_only` 72.6%, `full_kd` 72.0%) exceed the published YOLOv8n baseline (68.4%, 3.2M) and Pruned-KD (62.7%, 2.64M), occupying a distinct efficiency region. YOLOv11n (2.6M, 85.9%) and CCi-YOLOv8n (3.8M, 78.5%) are published figures from separately optimised configurations; they indicate the current state-of-the-art landscape but do not constitute equivalent-protocol comparisons with our results.

3.3. Cross-Dataset Generalisation

Table 8 reports mAP@50 on the held-out D-Fire test split and the three external benchmarks.

Table 8. Cross-dataset mAP@50 (imgsz=640). Avg. ext. excludes D-Fire test. $\text{YOLOv8n}_{\text{base}}$ = vanilla YOLOv8n trained with identical protocol (no KD, standard Conv).

Model	D-Fire Test	Roboflow	BoWFire	mytest	Avg. ext.
Teacher	0.7868	0.0255	0.0333	0.0238	0.0275
$\text{YOLOv8n}_{\text{base}}$	0.7797	0.0155	0.0275	0.0283	0.0238
`no_ghost`	0.7324	0.0140	0.0247	0.0177	0.0188
`full_kd`	0.7154	0.0158	0.0296	0.0194	0.0216
`spatial_only`	0.7142	0.0166	0.0219	0.0262	0.0216
`semantic_only`	0.7224	0.0136	0.0199	0.0138	0.0158
`det_only`	0.7196	0.0111	0.0217	0.0130	0.0153
`no_eca`	0.7087	0.0143	0.0250	0.0180	0.0191

All eight models—including the vanilla YOLOv8n baseline—collapse to mAP@50 $\leq$ 0.034 on every external benchmark (Figure 8). Crucially, $\text{YOLOv8n}_{\text{base}}$ achieves 0.7797 mAP@50 on D-Fire test (nearly matching the 25.8M-parameter teacher) yet collapses identically on external datasets (avg. ext. $=$ 0.024). This result confirms that the generalisation failure is caused by the **single-source D-Fire**

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

training distribution, not by KD, GhostConv compression, or any other architectural factor. Table 9 confirms that this is driven by **precision collapse** ($P = 0.013$–0.054), not recall collapse ($R = 0.107$–0.524): models continue to detect genuine fire regions at reasonable recall but are overwhelmed by false positives on out-of-distribution backgrounds.

Table 9. Cross-dataset precision/recall decomposition on external benchmarks.

Model	Roboflow P	Roboflow R	BoWFire P	BoWFire R	mytest P	mytest R
Teacher	0.031	0.434	0.054	0.203	0.034	0.245
`no_ghost`	0.017	0.475	0.039	0.217	0.026	0.327
`full_kd`	0.019	0.499	0.044	0.251	0.025	0.335
`spatial_only`	0.017	0.524	0.033	0.217	0.034	0.426
`semantic_only`	0.015	0.473	0.029	0.213	0.020	0.278
`det_only`	0.013	0.455	0.033	0.227	0.016	0.223
`no_eca`	0.020	0.107 [1]	0.037	0.275	0.025	0.385

[1] The `no_eca` Roboflow recall anomaly (0.107 vs. 0.434–0.524 for all others) suggests that ECA provides disproportionate benefit for fire recall on high-diversity OOD distributions.

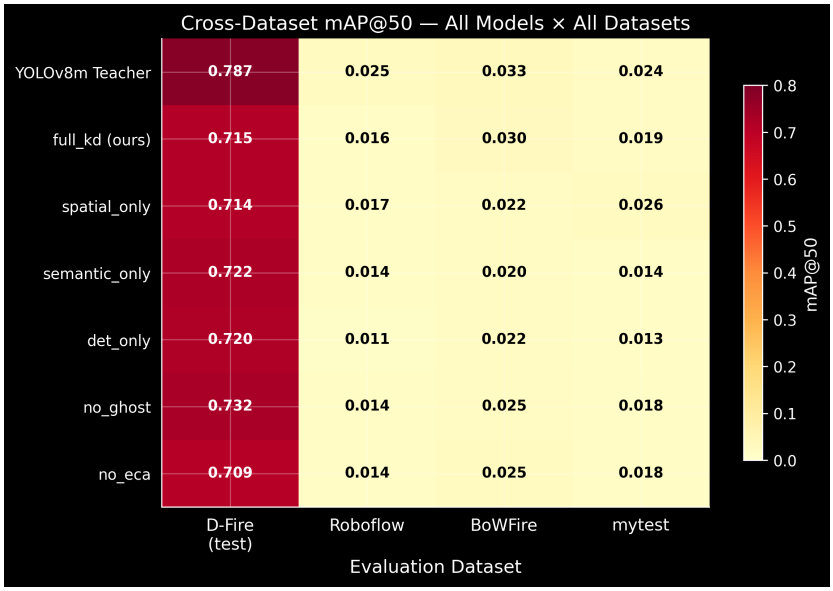


Figure 8. 7×4 cross-dataset mAP@50 heatmap for the six KD ablations and the teacher. The dramatic colour transition from the D-Fire column (0.71–0.79) to external columns ($\leq$0.034) visualises catastrophic generalisation collapse affecting all models equally. Model scale and distillation strategy make no meaningful difference to the gap. YOLOv8n$_{\text{base}}$ (Table 8) is omitted from this figure for space; its near-identical collapse pattern (avg. ext. = 0.024) is the primary evidence ruling out KD and architecture as causal factors.

Feature-space analysis confirms the domain shift responsible for the collapse. Figure 9 shows t-SNE projections of the `full_kd` student's P5 layer features (Layer 10, global average pooled) extracted from 1726 images across all four datasets. D-Fire test images cluster distinctly from all external benchmarks, with Roboflow, BoWFire, and mytest forming separate, non-overlapping regions in the 2D embedding. This distributional disjointness at the deepest feature level explains why the student's detection head—trained entirely on D-Fire feature representations—generates uncontrolled false positives when presented with structurally different image distributions at inference.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

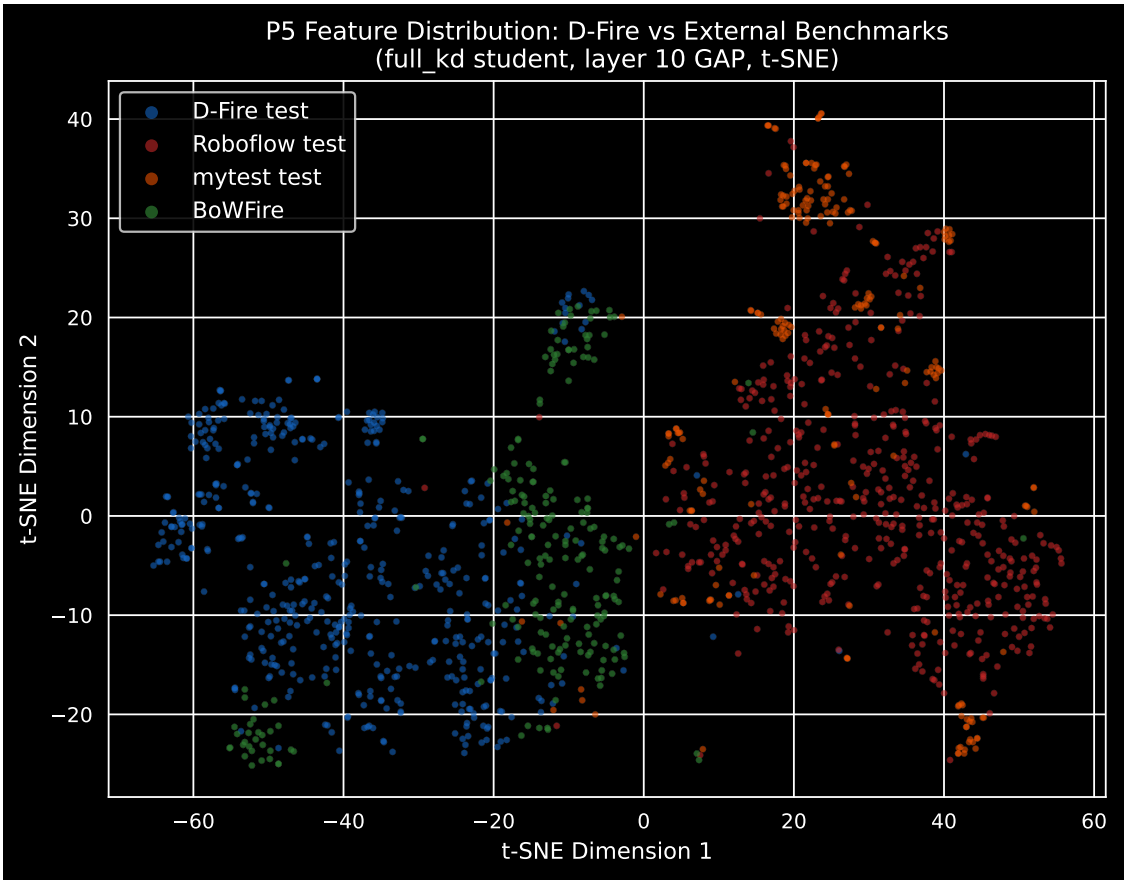


Figure 9. t-SNE (perplexity = 30) of `full_kd` student P5 features (Layer 10, global average pooled, 256-d) for 500 D-Fire test images (blue), 500 Roboflow images (red), 500 mytest images (orange), and 226 BoWFire images (green). D-Fire occupies a distinct region; external benchmarks form separate, non-overlapping clusters. This distributional disjointness at the deepest feature level is the mechanistic cause of the precision collapse documented in Tables 8 and 9.

Qualitative inspection of false-positive detections on out-of-distribution images further illustrates the failure modes (Figure 10). On Roboflow negatives (empty-label images with no fire) and BoWFire non-fire images, the student fires on dark backgrounds, textured walls, and illuminated objects that share local colour or texture statistics with D-Fire training images. On confounder categories, steam patches produce the most frequent false positives due to their visual similarity to smoke.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

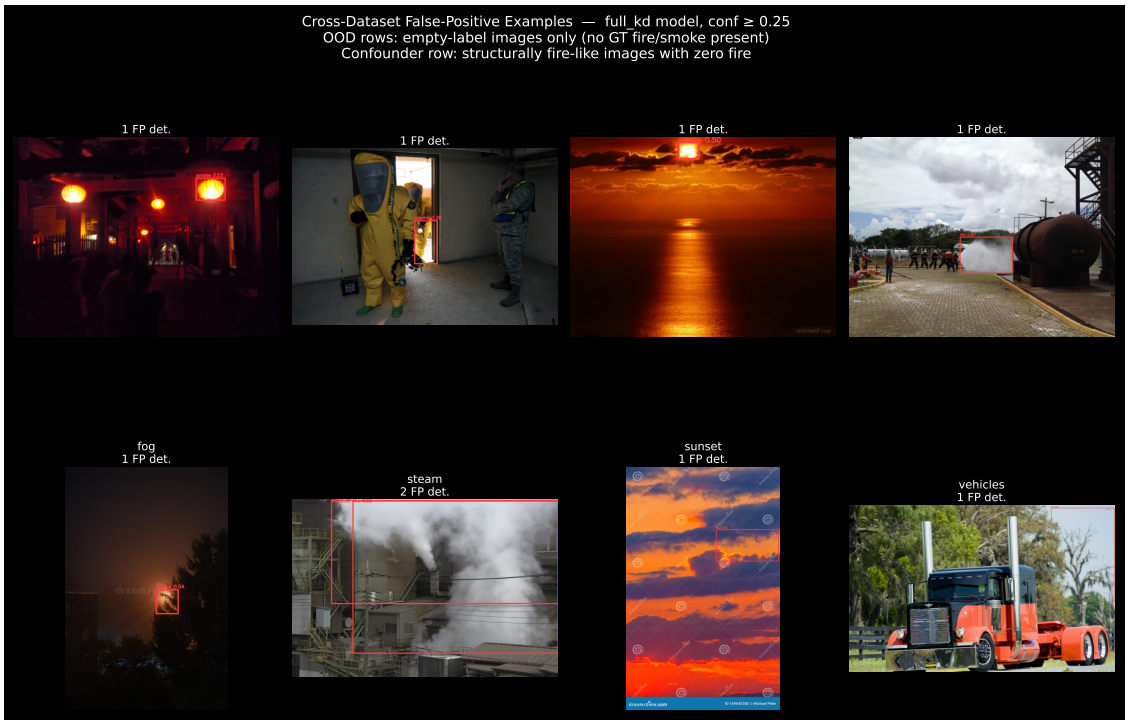


Figure 10. Representative false-positive detections (conf $\geq$ 0.25) on out-of-distribution images. Each cell shows the predicted bounding box with class label (fire/smoke) and confidence score. Row A: Roboflow empty-label negatives. Row B: mytest empty-label negatives. Row C: BoWFire non-fire images. Row D: confounder categories (fog, steam, sunset, vehicles). All detections shown are incorrect; steam (Row D, col. 2) consistently produces the highest-confidence false positives across all models, consistent with its visual similarity to smoke in the training distribution.

3.4. Confounder False-Positive Analysis

Table 10 reports the false-positive rate (FPR) on the 400-image confounder suite.

Table 10. Confounder false-positive rate (100 images per category, three confidence thresholds). Per-category rates are at conf $\geq$ 0.25. Best value per column in bold.

Model	Fog	Steam	Sunset	Vehicles	FPR@0.25	FPR@0.50	FPR@0.75
Teacher	10%	39%	9%	5%	15.75%	10.00%	3.50%
`no_ghost`	18%	31%	**4%**	**1%**	**13.50%**	**5.50%**	**1.00%**
`semantic_only`	30%	**29%**	13%	7%	19.75%	10.00%	1.50%
`full_kd`	22%	41%	15%	4%	20.50%	6.50%	0.50%
`det_only`	30%	33%	6%	4%	18.25%	7.75%	0.50%
`no_eca`	19%	39%	10%	6%	18.50%	6.50%	1.00%
`spatial_only`	33%	43%	8%	5%	22.25%	8.75%	0.50%

Steam produces the highest FPR (29–43%) due to visual similarity to smoke; vehicles produce the lowest (1–7%) (Figure 11). `no_ghost` achieves the best overall FPR at all thresholds, outperforming even the teacher. Critically, CLIP-distilled models (`semantic_only`, `full_kd`) do **not** outperform `det_only` at any confidence threshold.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

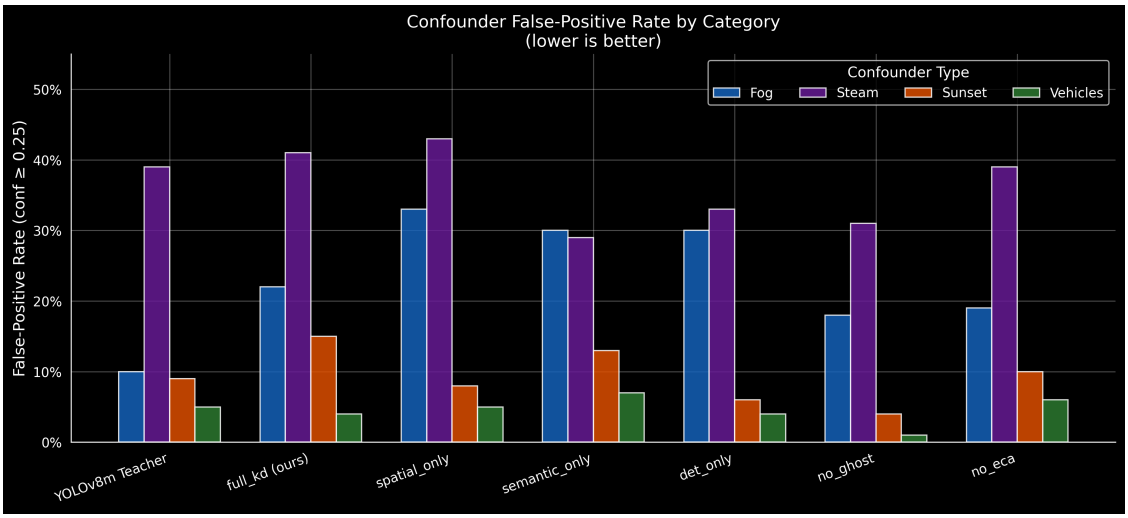


Figure 11. Per-category confounder FP rate (conf $\geq$ 0.25) for all seven models. Steam (29–43%) is consistently hardest; vehicles (1–7%) easiest. `no_ghost` achieves the best overall suppression (13.5%), outperforming the 25.8M-parameter teacher (15.75%).

3.5. Deployment Benchmark

Table 11 reports inference throughput and latency for the teacher and all six student/KD configurations under PyTorch FP32, PyTorch FP16, and TensorRT FP16 execution.

Table 11. Inference benchmark (NVIDIA RTX 5080, batch = 1, imgsz=640). PT = PyTorch; TRT = TensorRT FP16.

Model	Params	GFLOPs	PT FP32 (FPS)	PT FP16 (FPS)	TRT FP16 (FPS)	TRT (ms)
Teacher	25.84M	78.7	187.2	143.9	287.4	3.479
`full_kd`	2.298M	6.2	206.6	158.7	**375.6**	**2.663**
`spatial_only`	2.298M	6.2	225.9	155.8	343.3	2.913
`semantic_only`	2.298M	6.2	223.9	156.3	277.3 [2]	3.606
`det_only`	2.298M	6.2	237.1	156.2	341.9	2.925
`no_ghost`	3.006M	8.1	307.3 [1]	291.1 [1]	368.0	2.717
`no_eca`	2.298M	6.2	192.8	175.0	304.3	3.286

[1] `no_ghost` PyTorch FPS is paradoxically higher than that of GhostConv models despite 30% more GFLOPs: standard Conv maps directly to BLAS `im2col`+GEMM kernels with contiguous memory access, whereas GhostConv's cheap-operation branch introduces gather/scatter overhead that limits GPU utilisation at small batch sizes. [2] The `semantic_only` TRT shortfall is attributable to non-deterministic tactic selection; PyTorch FP16 shows near-identical throughput across all GhostConv+ECA models (CV $<$ 1%).

`full_kd` achieves 375.6 FPS—**30.7% faster** than the teacher with 11.2$\times$ fewer parameters (Figure 12). All students exceed the real-time threshold of 30 FPS by at least 8$\times$.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

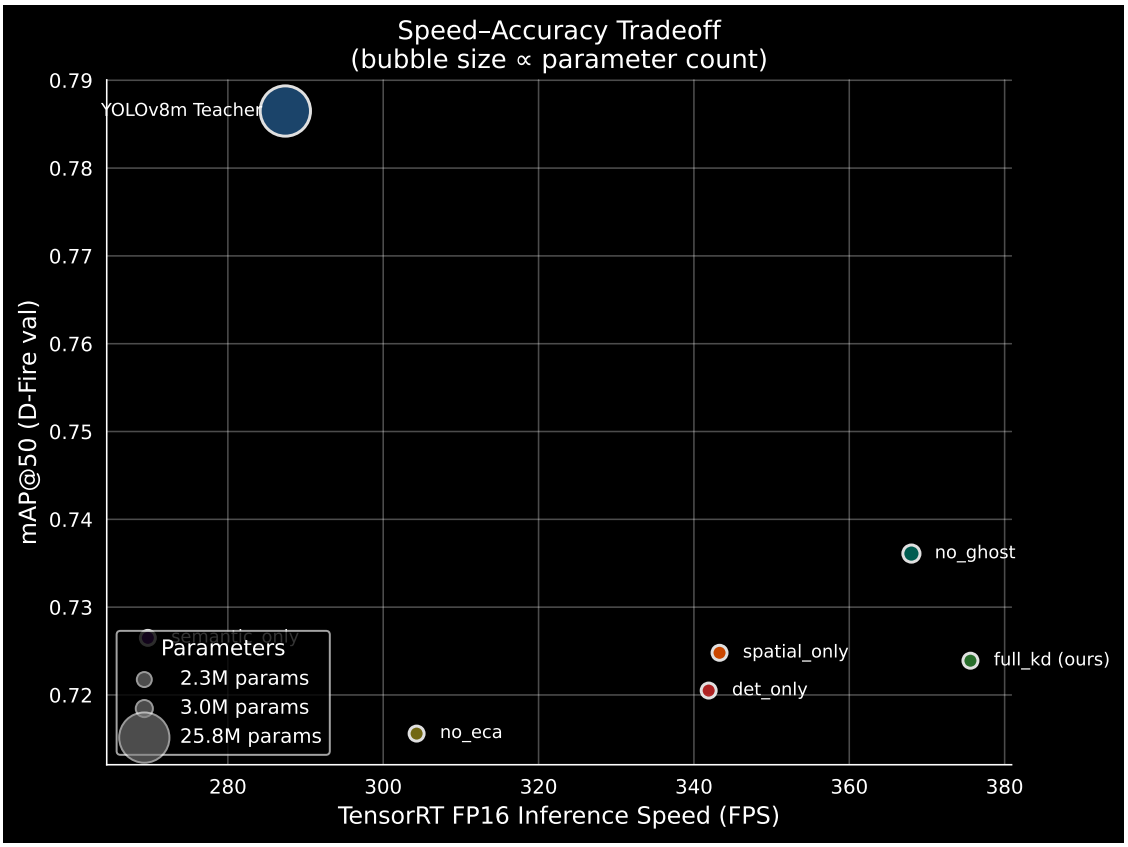


Figure 12. TensorRT FP16 inference speed vs. D-Fire validation mAP@50; bubble size is proportional to parameter count. `full_kd` (2.3M, 72.0%, 375.6 FPS) and `no_ghost` (3.0M, 73.6%, 368.0 FPS) define the student efficiency frontier. The teacher (25.8M, 78.7%, 287.4 FPS) is higher in mAP@50 but slower in throughput. `no_eca` (304.3 FPS) and `semantic_only` (277.3 FPS, TRT tactic anomaly) anchor the lower speed bound.

4. Discussion

4.1. Compression Efficiency and Architecture Trade-Offs

GhostConv delivers ∼23.5% parameter reduction versus `no_ghost` (3.006M → 2.298M) at a cost of −1.58 pp mAP@50 (73.61% → 72.03%). The mAP@50:95 gap (0.4256 → 0.4094) is proportionally larger, confirming that GhostConv's linear ghost features lose fine spatial detail needed for tight bounding-box localisation. For region-level alarm triggering in surveillance systems, this localisation trade-off is operationally acceptable; for fire boundary delineation applications, `no_ghost` is preferable.

ECA contributes +0.99 pp mAP@50 (`semantic_only` four-seed mean: 0.7255±0.0015 vs. `no_eca`: 0.7156) and, more operationally significant, +2.52 pp precision (0.7434±0.0079 vs. 0.7182) at a cost of 5 parameters. For systems where false alarms trigger costly evacuations, this precision gain is meaningful at negligible overhead. ECA also provides a disproportionate recall benefit on high-diversity OOD distributions, as evidenced by the `no_eca` Roboflow recall anomaly (0.107 vs. 0.434–0.524 for all ECA models; Table 9).

KD versus detection-only training provides +0.34–0.60 pp mAP@50 and consistently improves precision across all KD variants.

4.2. KD Loss Decomposition and α-Sweep

The multi-seed mAP@50 ordering—`semantic_only` (0.7255±0.0015) > `spatial_only` (0.7248, single-seed) > `det_only` (0.7205, single-seed) > `full_kd` (0.7203±0.0031)—is statistically confirmed for the two primary configurations: the Δ=0.0052 gap between `semantic_only` and `full_kd` exceeds the combined 1σ envelope (0.0046), establishing non-overlapping intervals across four independent seeds (original run plus three re-initialised seeds; Figure 2).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

This is not a novel finding but a domain-specific confirmation of established KD theory. CRD [23] demonstrated that under architectural mismatch, semantic distillation outperforms spatial feature mimicry because the spatial MSE gradient becomes noisy when teacher features (C2f residual) and student features (C3Ghost linear ghost) have structurally incompatible generation mechanisms. At α=0.5 > β=0.3, the noisy spatial gradient dominates the more reliable semantic signal.

To test whether the ordering is correctable by reweighting, an α-sweep was performed over $\{0.1, 0.2, 0.3, 0.5\}$ with β=0.3 and T=4.0 held fixed (Figure 13; validation trajectories in Figure 14). The results are summarised in Table 12.

Table 12. α-sweep results: `full_kd` at four spatial KD weights (100 epochs, β=0.3, T=4.0 fixed; all values single-seed). `semantic_only` (α=0, single-seed) is shown for reference; its four-seed mean is mAP@50 0.7255±0.0015, P 0.7434±0.0079, F1 0.6963±0.0038 (Table 5). All precision and F1 deltas discussed in the text use this four-seed mean, not the single-seed row below, as the `semantic_only` reference point.

α	mAP@50	mAP@50:95	P	F1
0 (`sem. only`)	**0.7265**	0.4121	0.7552	**0.7009**
0.1	0.7196	0.4099	0.7343	0.6876
0.2	**0.7265**	**0.4153**	**0.7628**	0.6971
0.3	0.7257	0.4094	0.7574	0.6964
0.5 (original)	0.7239	0.4118	0.7425	0.6953

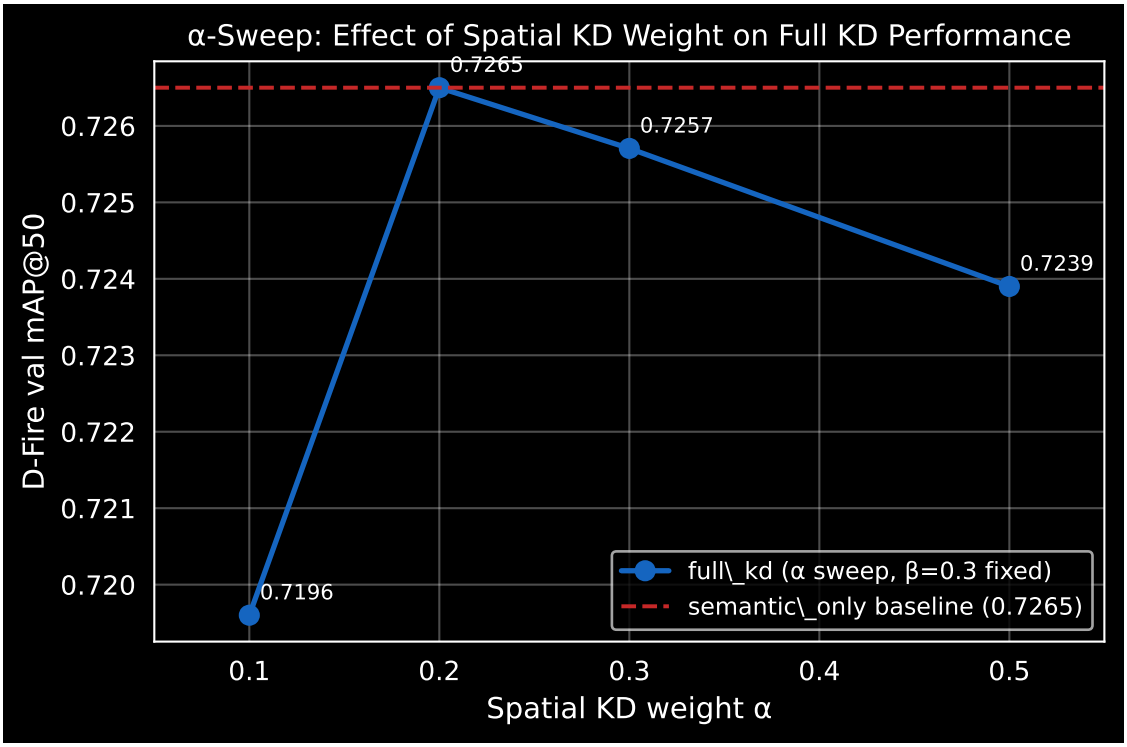


Figure 13. mAP@50 versus spatial KD weight α (with β=0.3 fixed). The horizontal dashed line indicates `semantic_only` (α=0, 0.7265). At α=0.2, `full_kd` matches the `semantic_only` mAP@50 exactly and achieves the highest precision in the sweep (0.7628 vs. 0.7552). Beyond α=0.2, increasing the spatial weight monotonically degrades performance, consistent with noisy spatial KD under C2f→C3Ghost mismatch overriding the CLIP semantic signal.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

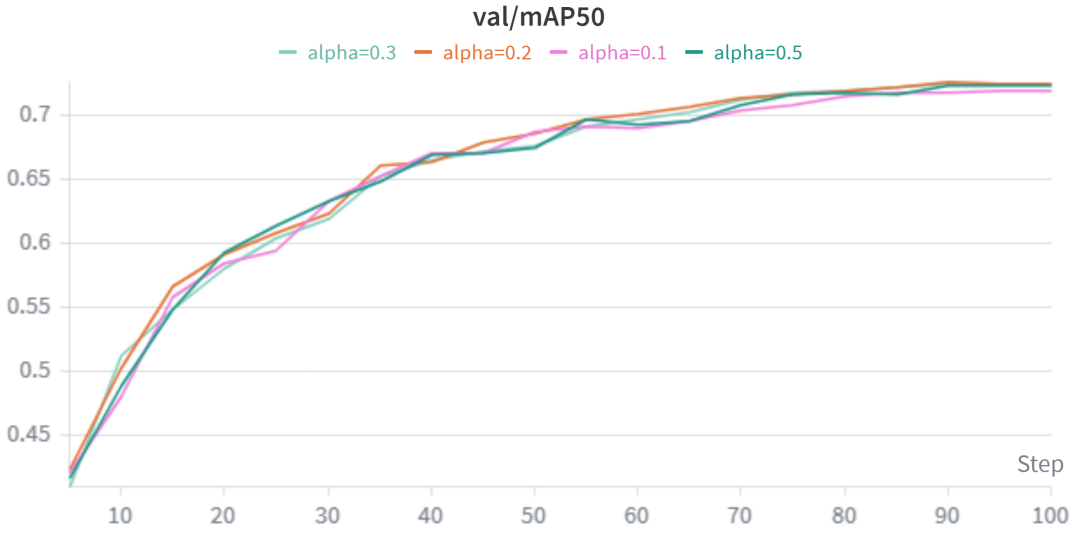


Figure 14. Validation mAP@50 trajectories over 100 epochs for all four α values in the spatial KD weight sweep ($\alpha \in \{0.1, 0.2, 0.3, 0.5\}$, β=0.3, T=4.0 fixed). All four configurations converge by epoch 80 within a narrow 0.007-unit band (0.7196–0.7265), with curves tightly clustered from epoch 40 onward; α=0.1 shows a modest dip below the group in early training before recovering. Final-epoch ordering confirms α=0.2 as the optimum (0.7265, equal to the single-seed `semantic_only` result), with performance degrading monotonically at higher weights (α=0.3: 0.7257; α=0.5: 0.7239) as the spatial MSE gradient increasingly suppresses the CLIP semantic signal, and falling below the `semantic_only` baseline at α=0.1 (0.7196). Quantitative results are in Table 12.

At α=0.2, `full_kd` matches the four-seed `semantic_only` mean in mAP@50 within noise (0.7265 vs. 0.7255±0.0015) and achieves the highest precision among all student/KD configurations (0.7628 vs. the four-seed `semantic_only` mean of 0.7434±0.0079, +1.94 pp). F1 at α=0.2 (0.6971) is statistically indistinguishable from the four-seed `semantic_only` mean (0.6963±0.0038), indicating that optimally weighted spatial KD recovers CLIP-only performance on mAP@50 and F1 while also delivering a measurable precision gain, rather than trading precision for a recall penalty as a single-seed comparison would suggest. The α=0.1 result (0.7196) undershoots the four-seed `semantic_only` mAP@50 mean, suggesting an optimal narrow window around α=0.2. Together, these results establish that the `semantic_only` > `full_kd` finding at α=0.5 is not an artefact of arbitrary weighting, but rather reflects the fundamental asymmetry between CLIP's strong semantic signal and spatial MSE's structurally noisy gradient under C2f→C3Ghost mismatch. Practitioners adopting this framework should use α=0.2 as the deployment hyperparameter, since it recovers `semantic_only`-level mAP@50 and F1 while providing a measurable precision advantage; the ablation suite in this work uses α=0.5 solely for cross-configuration consistency, not as a performance recommendation.

These results invite a direct question: if `semantic_only` matches or exceeds the combined loss on D-Fire at every α tested above 0.1, why retain spatial supervision at all rather than deploying `semantic_only` outright? Two considerations motivate keeping the dual-teacher design. First, the decomposable spatial/semantic loss structure is itself the instrument that produced this finding—without an ablatable spatial branch, the architecture-mismatch diagnosis above could not have been established, and the α-sweep would have nothing to sweep. Second, and more consequentially, spatial supervision confers a measurable cross-dataset generalisation benefit that semantic supervision alone does not, as quantified next in Section 4.3; the dual-teacher framework's value is therefore realised primarily outside the D-Fire distribution rather than within it.

4.3. Cross-Dataset Generalisation: A Universal Open Problem

All eight models—spanning 11.2× in parameter count and including the vanilla YOLOv8n baseline—fall within an absolute range of 0.012 mAP@50 on external datasets (0.0153–0.0275), com-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

pared to a within-dataset spread of 0.078. The generalisation gap (∼0.76 mAP@50 units) is approximately ten times larger than the architectural or training effect.

The vanilla YOLOv8n$_{\text{base}}$—identical training protocol, standard Conv, no KD—achieves 0.7797 mAP@50 on D-Fire test (nearly matching the full teacher) yet collapses to avg. ext. = 0.024, the same order of magnitude as all KD variants. This conclusively rules out distillation artefacts or GhostConv compression as causal factors. The t-SNE feature visualisation (Figure 9) and qualitative failure examples (Figure 10) confirm that the P5 feature space is partitioned by dataset of origin, not by semantic content: external images occupy non-overlapping regions of the feature manifold, causing the detection head to generate uncontrolled false positives.

`det_only`—no KD, detection loss only—collapses equally to all KD variants, ruling out distillation artefacts. The teacher—standard C2f at 25.8M parameters—collapses equally to all students, ruling out architectural modifications. Together they establish that the collapse is independent of both KD strategy and architecture type, implicating **single-source D-Fire training** as the primary cause. CLIP-distilled models generalise no better than `det_only`, directly refuting the hypothesis that CLIP's domain-diverse pretraining improves cross-dataset transfer in this formulation.

A finer-grained reading of Table 8, however, distinguishes the two CLIP-supervised configurations from one another and provides the empirical justification for retaining the dual-teacher design despite the D-Fire-only `semantic_only` advantage reported in Section 4.2. `full_kd` and `spatial_only`—the two configurations that retain spatial feature mimicry against the YOLOv8m teacher—achieve the highest average external mAP@50 among all GhostConv+ECA students (0.0216 each), exceeding `semantic_only` (0.0158) and `det_only` (0.0153) by 37–41% relative. Table 9 localises this advantage to recall: `full_kd` and `spatial_only` sustain higher recall on Roboflow (0.499, 0.524) and mytest (0.335, 0.426) than `semantic_only` (0.473, 0.278), indicating that spatial feature mimicry against the teacher preserves more of the teacher's general-purpose objectness cues under domain shift than semantic supervision alone provides—even though this same spatial term is shown in Section 4.2 to be sub-optimally weighted for in-domain mAP@50 at α=0.5. Semantic distillation alone therefore yields the best within-distribution accuracy, but spatial distillation contributes an external-generalisation benefit that semantic supervision does not replicate; the two signals are complementary along different evaluation axes rather than redundant, which is the direct answer to why the dual-teacher architecture is retained rather than collapsed to `semantic_only`.

4.4. CLIP Confounder Prompts: A Constructive Negative Result

CLIP-distilled models do not reduce FPR relative to `det_only` (Table 10). The failure has two reinforcing causes.

First, **scale mismatch**: CLIP operates on 224×224 whole-image crops, producing a global embedding. Detection false positives occur at the region level—a 15% steam patch activates local convolution filters regardless of the global CLIP label learned from training (Figure 15). Second, **coverage gap**: CLIP soft labels are computed only on D-Fire training images; the student receives no supervision signal from actual confounder scenes encountered at deployment. DenseCLIP [34] and CORA [35] confirm that region-conditioned CLIP features are required for dense prediction—precisely the modification needed to address this failure.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

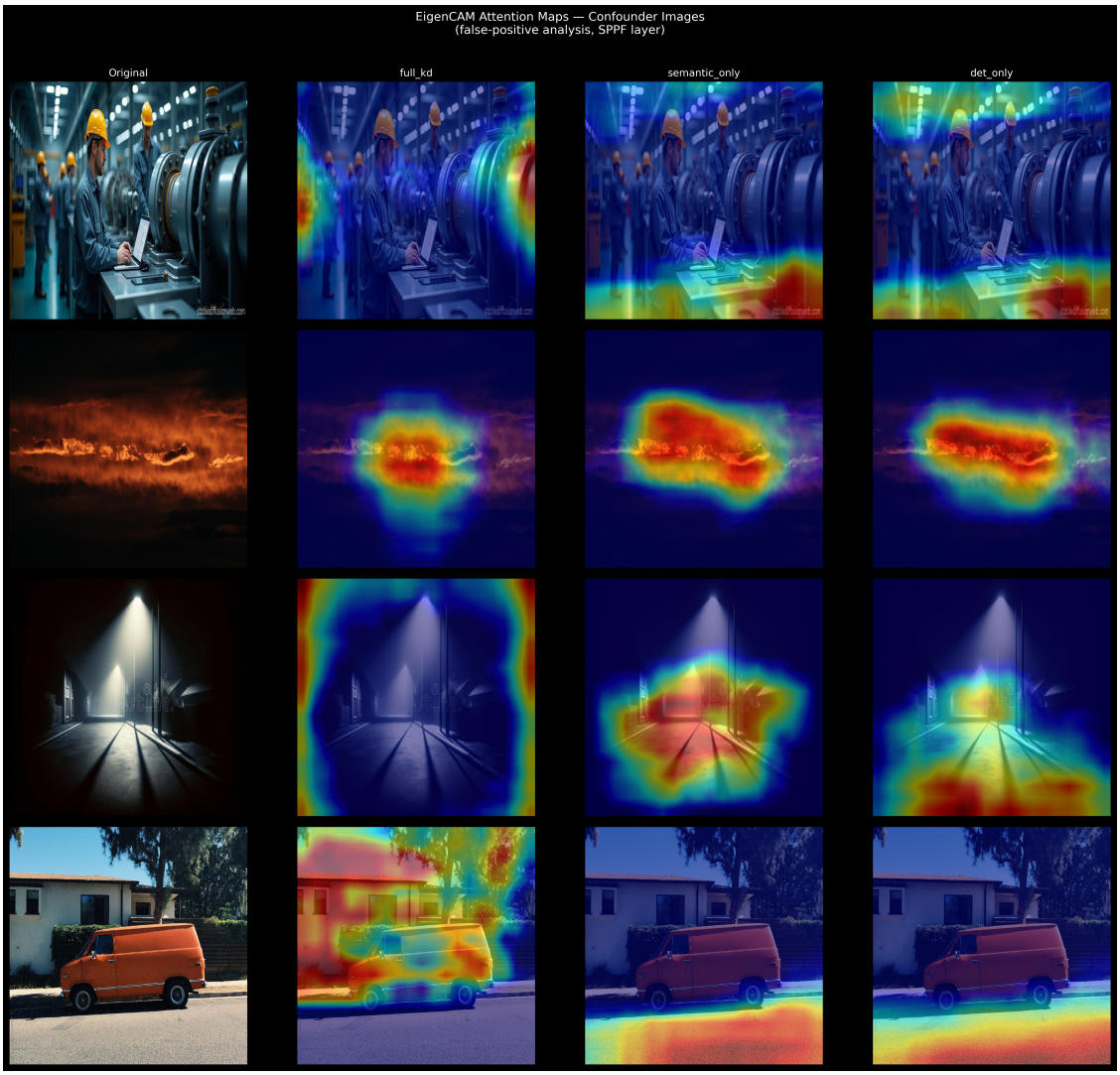


Figure 15. EigenCAM [48] activation maps on confounder images (steam, sunset, fog, vehicles) for `full_kd`, `semantic_only`, and `det_only`. All three models activate on steam and fog patches regardless of CLIP distillation strategy, demonstrating that false-positive activations are driven by local texture patterns that global-image CLIP supervision cannot suppress.

4.5. Limitations

BoWFire bounding boxes are algorithmically derived; mAP@50 is internally consistent across all models but should not be interpreted as absolute localisation accuracy. Multi-seed confirmation is provided for `full_kd` and `semantic_only`; the remaining four configurations (`spatial_only`, `det_only`, `no_ghost`, `no_eca`) are single-seed, so differences < 0.003 mAP@50 among these should be treated as indicative. The α-sweep identifies α=0.2 as optimal but tests each value as a single run; a multi-seed α-sweep would fully quantify the optimum. Extended training beyond 100 epochs would likely improve mAP@50:95. All benchmarks use an RTX 5080; absolute FPS on edge hardware (e.g., Jetson Orin NX) would differ, though relative compression gains are architecture-portable.

5. Conclusions

A dual-teacher KD framework has been presented—the first to use a VLM as a semantic distillation signal for fire detection—compressing a YOLOv8m teacher into a GhostConv-YOLOv8n+ECA student at 11.2× parameter compression and 12.7× FLOP reduction with 91.6% mAP@50 retention on D-Fire validation (four-seed mean: 72.0% $\pm$ 0.3%; held-out test: 71.5%, 90.9% retention). The student surpasses Pruned-KD [8] by +9.3 pp on D-Fire validation (cross-benchmark context: Pruned-KD was originally trained on a different custom forest-fire dataset; 62.7% is a third-party D-Fire re-evaluation [3]) and the

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

published YOLOv8n baseline by +3.6 pp with fewer parameters, reaching 375.6 FPS under TensorRT. ECA delivers +0.99 pp mAP@50 and disproportionate recall preservation on OOD distributions at a cost of 5 parameters. Multi-seed experiments confirm that CLIP semantic distillation statistically outperforms combined spatial + semantic KD across all four seeds (0.7255±0.0015 vs. 0.7203±0.0031), consistent with CRD theory under C2f→C3Ghost architectural mismatch. An α-sweep identifies α=0.2 as the operating point where the combined loss recovers to the four-seed `semantic_only` mean (mAP@50 = 0.7265 vs. 0.7255±0.0015; F1 = 0.6971 vs. 0.6963±0.0038) while achieving a +1.94 pp precision gain over that mean (precision = 0.7628 vs. 0.7434±0.0079). Although semantic distillation alone attains the best in-domain accuracy, the dual-teacher design is retained because spatial supervision measurably improves cross-dataset generalisation—`full_kd` and `spatial_only` exceed `semantic_only`'s average external mAP@50 by 37–41% relative (Section 4.3)—demonstrating that the two teacher signals are complementary across evaluation axes rather than one simply subsuming the other.

The cross-dataset evaluation—the first systematic multi-benchmark study for a sub-5M-parameter fire detector, controlled by the inclusion of a vanilla YOLOv8n baseline—establishes that all architectures, all KD strategies, and even uncompressed baselines collapse from 71–79% to ≤3.3% mAP@50 on external benchmarks. The t-SNE feature analysis confirms that P5 features are partitioned by training domain, not semantic content. CLIP confounder prompts provide no false-positive reduction, explained by the incompatibility between global-image CLIP supervision and region-level detection. These results are reported as quantified open problems: multi-source training and region-conditioned CLIP distillation (DenseCLIP, CORA) are the direct paths forward.

Author Contributions: Conceptualization, M.J.I., M.A.H.P. and A.S.M.M.; methodology, M.A.H.P. and S.J.; software, M.A.H.P.; validation, M.A.H.P., M.H.M. and S.J.; formal analysis, M.A.H.P., M.H.M., F.A.F. and H.A.K.; investigation, M.A.H.P. and M.H.M.; resources, A.S.M.M.; data curation, M.A.H.P. and S.J.; writing—original draft preparation, M.A.H.P.; writing—review and editing, M.J.I., M.H.M., S.J., A.S.M.M., M.N.I., F.A.F. and H.A.K.; visualization, M.A.H.P. and A.S.M.M.; supervision, M.J.I., M.N.I., F.A.F. and H.A.K.; project administration, M.J.I. and M.N.I.; funding acquisition, M.J.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Chandpur Science and Technology University, Bangladesh, under Research Project Code RP-C1/01/25-26.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The D-Fire dataset is publicly available at https://github.com/gaiasd/DFireDataset. The BoWFire dataset is available at http://www.ic.unicamp.br/~rocha/pub/downloads/2014-bowfire/. The Roboflow Fire-Smoke 9.8K and mytest-hrswj datasets are available on Roboflow Universe. The confounder dataset (400 images), trained model weights (`kd_student_*.pt`), and evaluation scripts will be deposited on Zenodo upon acceptance.

Acknowledgments: During the preparation of this manuscript, the authors used the Scientific Agent Skills toolkit (K-Dense Inc., 2026) [49], an AI-agent-based software framework, for the purposes of formatting the manuscript to the target journal's LaTeX template, managing and validating the bibliography and in-text citation style, and checking compliance with the journal's submission guidelines. This tool was not used to generate, analyse, or interpret any research data, results, or scientific conclusions, nor to draft the scientific content of any section. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

The following abbreviations are used in this manuscript:

CLIP	Contrastive Language–Image Pre-training
CRD	Contrastive Representation Distillation
ECA	Efficient Channel Attention
FPN	Feature Pyramid Network
FPR	False-Positive Rate
FPS	Frames Per Second
GAP	Global Average Pooling
GFLOPs	Giga Floating-Point Operations
KD	Knowledge Distillation
KL	Kullback–Leibler
mAP	Mean Average Precision
MSE	Mean Squared Error
OOD	Out-of-Distribution
SPPF	Spatial Pyramid Pooling—Fast
TRT	TensorRT
t-SNE	t-Distributed Stochastic Neighbour Embedding
ViT	Vision Transformer
VLM	Vision-Language Model
YOLO	You Only Look Once

References

1. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91, https://doi.org/10.1109/CVPR.2016.91.
2. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. GitHub repository, 2023. Version 8.4.24, AGPL-3.0 license. https://github.com/ultralytics/ultralytics.
3. Li, C.; Du, Y.; Zhang, X.; Wu, P. YOLOGX: an improved forest fire detection algorithm based on YOLOv8. *Frontiers in Environmental Science* **2025**, *12*, 1486212. https://doi.org/10.3389/fenvs.2024.1486212.
4. Lv, K.; Wu, R.; Chen, S.; Lan, P. CCi-YOLOv8n: Enhanced fire detection with CARAFE and context-guided modules. In Proceedings of the International Conference on Intelligent Computing. International Conference on Intelligent Computing (ICIC), Springer, 2025, pp. 128–140. DOI: 10.1007/978-981-96-9901-8_11, https://doi.org/10.1007/978-981-96-9901-8_11.
5. Liu, H.; Zhang, F.; Xu, Y.; Wang, J.; Lu, H.; Wei, W.; Zhu, J. TFNet: Transformer-Based Multi-Scale Feature Fusion Forest Fire Image Detection Network. *Fire* **2025**, *8*, 59. https://doi.org/10.3390/fire8020059.
6. Boroujeni, S.P.H.; Mehrabi, N.; Afghah, F.; McGrath, C.P.; Bhatkar, D.; Biradar, M.A.; Razi, A. Eyes on the Environment: AI-Driven Analysis for Fire and Smoke Classification, Segmentation, and Detection. arXiv preprint, 2025. arXiv:2503.14552v2 (originally titled "Fire and Smoke Datasets in 20 Years: An In-depth Review"). DOI: 10.48550/arXiv.2503.14552, https://doi.org/10.48550/arXiv.2503.14552.
7. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. arXiv preprint, 2015, [arXiv:stat.ML/1503.02531]. arXiv:1503.02531. NIPS 2014 Deep Learning and Representation Learning Workshop (non-archival), https://doi.org/10.48550/arXiv.1503.02531.
8. Wang, S.; Zhao, J.; Ta, N.; Zhao, X.; Xiao, M.; Wei, H. A real-time deep learning forest fire monitoring algorithm based on an improved Pruned + KD model. *Journal of Real-Time Image Processing* **2021**, *18*, 2319–2329. https://doi.org/10.1007/s11554-021-01124-9.
9. Yun, B.; Zheng, Y.; Lin, Z.; Li, T. FFYOLO: A lightweight forest fire detection model based on YOLOv8. *Fire* **2024**, *7*, 93. https://doi.org/10.3390/fire7030093.
10. Ahmad, N.; Akbar, M.; Alkhammash, E.H.; Jamjoom, M.M. FireNet-KD: Swin Transformer-Based Wildfire Detection with Multi-Source Knowledge Distillation. *Fire* **2025**, *8*, 295. https://doi.org/10.3390/fire8080295.
11. Keerthinathan, P.; Sandino, J.; Mahendren, S.; Uthayasooriyan, A.; Galvez, J.; Hamilton, G.; Gonzalez, F. Advancing Real-Time Aerial Wildfire Detection Through Plume Recognition and Knowledge Distillation. *Drones* **2025**, *9*, 827. https://doi.org/10.3390/drones9120827.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

12. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning (ICML). 38th International Conference on Machine Learning (ICML), PMLR, 2021, Vol. 139, pp. 8748–8763, [2103.00020]. arXiv:2103.00020. https://proceedings.mlr.press/v139/radford21a.html, https://doi.org/10.48550/arXiv.2103.00020.
13. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS). Advances in Neural Information Processing Systems 35 (NeurIPS), 2022, pp. 25278–25294. DOI: 10.52202/068431-1833, https://doi.org/10.52202/068431-1833.
14. Lin, Z.; Yun, B.; Zheng, Y. LD-YOLO: A Lightweight Dynamic Forest Fire and Smoke Detection Model with Dysample and Spatial Context Awareness Module. *Forests* **2024**, *15*, 1630. https://doi.org/10.3390/f15091630.
15. Alkhammash, E.H. A comparative analysis of YOLOv9, YOLOv10, YOLOv11 for smoke and fire detection. *Fire* **2025**, *8*, 26. https://doi.org/10.3390/fire8010026.
16. Wu, S.; Qiao, Y.; He, S.; Zhou, J.; Wang, Z.; Li, X.; Wang, F. FireCLIP: Enhancing forest fire detection with multimodal prompt tuning and vision-language understanding. *Fire* **2025**, *8*, 237. https://doi.org/10.3390/fire8060237.
17. Hasan, M.; Prince, M.M.A.H.; Ansari, M.S.; Jahan, S.; Miah, A.S.M.; Shin, J. Firelite: Leveraging Transfer Learning for Efficient Fire Detection in Resource-Constrained Environments. In Proceedings of the 2024 27th International Conference on Computer and Information Technology (ICCIT). 27th International Conference on Computer and Information Technology (ICCIT), IEEE, 2024, pp. 545–550. DOI: 10.1109/ICCIT64611.2024.11022050, https://doi.org/10.1109/ICCIT64611.2024.11022050.
18. Han, X.; Pu, N.; Feng, Z.; Bei, Y.; Zhang, Q.; Cheng, L.; Xue, L. Benchmarking Multi-Scene Fire and Smoke Detection. In Proceedings of the Proc. Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Chinese Conference on Pattern Recognition and Computer Vision (PRCV), 2024, [2410.16631]. arXiv:2410.16631, https://doi.org/10.48550/arXiv.2410.16631.
19. Yan, Z.; Wang, L.; Qin, K.; Zhou, F.; Ouyang, J.; Wang, T.; Hou, X.; Bu, L. Unsupervised Domain Adaptation for Forest Fire Recognition Using Transferable Knowledge from Public Datasets. *Forests* **2023**, *14*, 52. https://doi.org/10.3390/f14010052.
20. Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of the Proc. ICML. International Conference on Machine Learning (ICML), PMLR, 2015, pp. 1180–1189. No DOI assigned by PMLR. https://proceedings.mlr.press/v37/ganin15.html.
21. Wang, H.; Fu, X.; Yu, Z.; Zeng, Z. DSS-YOLO: An improved lightweight real-time fire detection model based on YOLOv8. *Scientific Reports* **2025**, *15*, 8963. https://doi.org/10.1038/s41598-025-93278-w.
22. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. FitNets: Hints for Thin Deep Nets. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 2015; [1412.6550]. arXiv:1412.6550, https://doi.org/10.48550/arXiv.1412.6550.
23. Tian, Y.; Krishnan, D.; Isola, P. Contrastive Representation Distillation. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020), 8th International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, 2020; [1910.10699]. arXiv:1910.10699. https://openreview.net/forum?id=SkgpBJrtvS, https://doi.org/10.48550/arXiv.1910.10699.
24. Chen, P.; Liu, S.; Zhao, H.; Jia, J. Distilling Knowledge via Knowledge Review. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5008–5017, [2104.09044]. DOI: 10.1109/CVPR46437.2021.00497, https://doi.org/10.1109/CVPR46437.2021.00497.
25. Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; Yuan, C. Focal and global knowledge distillation for detectors. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4633–4642. DOI: 10.1109/CVPR52688.2022.00460, https://doi.org/10.1109/CVPR52688.2022.00460.
26. Cao, W.; Zhang, Y.; Gao, J.; Cheng, A.; Cheng, K.; Cheng, J. PKD: General distillation framework for object detectors via Pearson correlation coefficient. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS). Advances in Neural Information Processing Systems 35 (NeurIPS), 2022, pp. 15394–15406. DOI: 10.52202/068431-1120, https://doi.org/10.52202/068431-1120.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

27. Wang, J.; Chen, Y.; Zheng, Z.; Li, X.; Cheng, M.M.; Hou, Q. CrossKD: Cross-Head Knowledge Distillation for Object Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16520–16530. DOI: 10.1109/CVPR52733.2024.01563, https://doi.org/10.1109/CVPR52733.2024.01563.
28. Prince, M.M.A.H.; Islam, M.J.; Jahan, S.; Miah, A.S.M.; Chowdhury, M.N.H. A Hybrid CNN-Transformer Knowledge Transfer Technique for Lightweight Skin Lesion Classification using Dermoscopic Image. In Proceedings of the 2025 IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS). IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS), 2025, pp. 1–6. DOI: 10.1109/COMPAS67506.2025.11381851, https://doi.org/10.1109/COMPAS67506.2025.11381851.
29. El-Madafri, I.; Peña, M.; Olmedo-Torre, N. Real-Time Forest Fire Detection with Lightweight CNN Using Hierarchical Multi-Task Knowledge Distillation. *Fire* **2024**, *7*, 392. https://doi.org/10.3390/fire7110392.
30. Gu, X.; Lin, T.Y.; Kuo, W.; Cui, Y. Open-Vocabulary Object Detection via Vision and Language Knowledge Distillation. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022). 10th International Conference on Learning Representations (ICLR 2022), 2022, [2104.13921]. arXiv:2104.13921. https://openreview.net/forum?id=lL3lnMbR4WU, https://doi.org/10.48550/arXiv.2104.13921.
31. Ma, Z.; Luo, G.; Gao, J.; Li, L.; Chen, Y.; Wang, S.; Zhang, C.; Hu, W. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14054–14063. DOI: 10.1109/CVPR52688.2022.01368, https://doi.org/10.1109/CVPR52688.2022.01368.
32. Jang, J.; Ma, C.; Lee, B. VL2Lite: Task-specific knowledge distillation from large vision-language models to lightweight networks. In Proceedings of the Proc. IEEE/CVF CVPR. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025. DOI: 10.1109/CVPR52734.2025.02799, https://doi.org/10.1109/CVPR52734.2025.02799.
33. Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; Dabeer, O. WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19606–19616. DOI: 10.1109/CVPR52729.2023.01878, https://doi.org/10.1109/CVPR52729.2023.01878.
34. Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; Lu, J. DenseCLIP: Language-guided dense prediction with context-aware prompting. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18061–18070. DOI: 10.1109/CVPR52688.2022.01755, https://doi.org/10.1109/CVPR52688.2022.01755.
35. Wu, X.; Zhu, F.; Zhao, R.; Li, H. CORA: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7031–7040. DOI: 10.1109/CVPR52729.2023.00679, https://doi.org/10.1109/CVPR52729.2023.00679.
36. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More features from cheap operations. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1577–1586. DOI: 10.1109/CVPR42600.2020.00165, https://doi.org/10.1109/CVPR42600.2020.00165.
37. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11531–11539. DOI: 10.1109/CVPR42600.2020.01155, https://doi.org/10.1109/CVPR42600.2020.01155.
38. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proc. IEEE/CVF CVPR. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018. DOI: 10.1109/CVPR.2018.00745, https://doi.org/10.1109/CVPR.2018.00745.
39. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 21 October 2026 https://doi.org/10.20944/preprints202610.2054.v1

Scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021). 9th International Conference on Learning Representations (ICLR 2021), 2021, [2010.11929]. arXiv:2010.11929, https://doi.org/10.48550/arXiv.2010.11929.

40. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). Advances in Neural Information Processing Systems 33 (NeurIPS), 2020, Vol. 33, pp. 21002–21012, [2006.04388]. arXiv:2006.04388, https://doi.org/10.48550/arXiv.2006.04388.
41. de Venâncio, P.V.A.B.; Rezende, T.M.; Lisboa, A.C.; Barbosa, A.V. Fire Detection based on a Two-Dimensional Convolutional Neural Network and Temporal Analysis. In Proceedings of the Proc. IEEE Latin American Conf. on Computational Intelligence (LA-CCI). IEEE Latin American Conference on Computational Intelligence (LA-CCI), 2021. DOI: 10.1109/la-cci48322.2021.9769824, https://doi.org/10.1109/la-cci48322.2021.9769824.
42. Roboflow. Fire and Smoke Detection Dataset. https://universe.roboflow.com/, 2024. Roboflow Universe. Accessed: June 2026.
43. Chino, D.Y.T.; Avalhais, L.P.S.; Rodrigues, J.F.; Traina, A.J.M. BoWFire: Detection of fire in still images by integrating pixel color and texture analysis. In Proceedings of the 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). 28th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2015, pp. 95–102. DOI: 10.1109/SIBGRAPI.2015.19, https://doi.org/10.1109/SIBGRAPI.2015.19.
44. Mytyolov8. Mytest Dataset. https://universe.roboflow.com/mytyolov8/mytest-hrswj, 2024. Open Source Dataset. Roboflow Universe. Visited 2026-06-17.
45. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the Proc. ICLR. International Conference on Learning Representations (ICLR), 2019, [1711.05101]. arXiv:1711.05101, https://doi.org/10.48550/arXiv.1711.05101.
46. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and Flexible Image Augmentations. *Information* **2020**, *11*, 125. https://doi.org/10.3390/info11020125.
47. Jiang, Y.; Liu, P.; Han, Y.; Xiao, B. YOLOv11-CHBG: A Lightweight Fire Detection Model. *Fire* **2025**, *8*, 338. https://doi.org/10.3390/fire8090338.
48. Gildenblat, J.; contributors. pytorch-grad-cam: Advanced AI Explainability for PyTorch. GitHub repository, 2021. https://github.com/jacobgil/pytorch-grad-cam.
49. K-Dense Inc.. Scientific Agent Skills: A Comprehensive Collection of Scientific Tools for AI Agents, 2026. 161 skills covering databases, packages, integrations, and analysis tools.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.