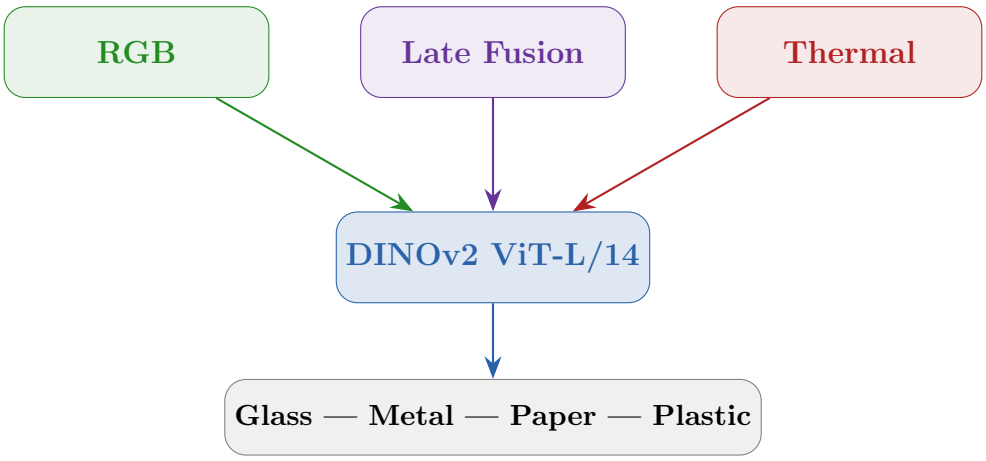
Multi-Modal Material Classification on Conveyor Belt Video Using Frozen DINOv2 Features

RGB, Thermal, and Late Fusion Pipelines
with Learnable Attention Pooling



Technical Report

February 15, 2026

Contents

1 Introduction

Automated waste sorting on high-throughput conveyor systems demands material discrimination that goes beyond what visual appearance alone can provide. Materials such as clear glass and clear plastic are nearly indistinguishable under visible light yet require completely different recycling processes. Thermal imaging in the long-wave infrared (LWIR, 7.5–14 μm) band captures material-specific thermal conductivity, specific heat capacity, and emissivity signatures that are invisible to RGB cameras.

This report presents a unified classification framework that leverages **DINOv2 ViT-L/14**—a self-supervised vision foundation model—as a frozen feature backbone for RGB, thermal, and late fusion modalities. By keeping DINOv2 frozen and training only lightweight attention pooling and MLP classification layers (∼528K parameters per single-modal pipeline, ∼1.05M for late fusion), we achieve strong classification performance while requiring minimal training data and compute.

Key contributions:

- A multi-view tracklet-level classification pipeline achieving **95.1% macro F1** on RGB, **90.6% macro F1** on thermal, and **96.0% macro F1** via late fusion—all using the same frozen DINOv2 backbone.
- Demonstration that frozen self-supervised features transfer effectively across spectral modalities (visible → LWIR) through simple grayscale channel replication.
- A spatial registration pipeline using homography warping that enables cross-modal mask transfer from RGB detection space to thermal image space.
- A late fusion architecture with modality-specific attention pooling that reduces classification errors by 26% relative to RGB alone, demonstrating complementary information in RGB and thermal modalities.

2 System Architecture

2.1 Physical Setup

The experimental system comprises a conveyor belt (160 cm × 40 cm, speed 1.8 cm/s), a 2500 W electric heater positioned 60 cm above the belt creating an active heating zone, and two cameras: an RGB camera (1280×720, 30 fps) and a FLIR T420 thermal camera (320×240 LWIR, 30 fps). Objects traverse the heating zone, acquiring material-specific thermal signatures as they heat and subsequently cool.

2.2 End-to-End Pipeline

Both the RGB and thermal classification pipelines share a common architectural pattern: detection and tracking in RGB space, feature extraction via frozen DINOv2, and classification through learnable attention pooling and an MLP head. The key difference lies in the input domain and the preprocessing required to bridge modalities.

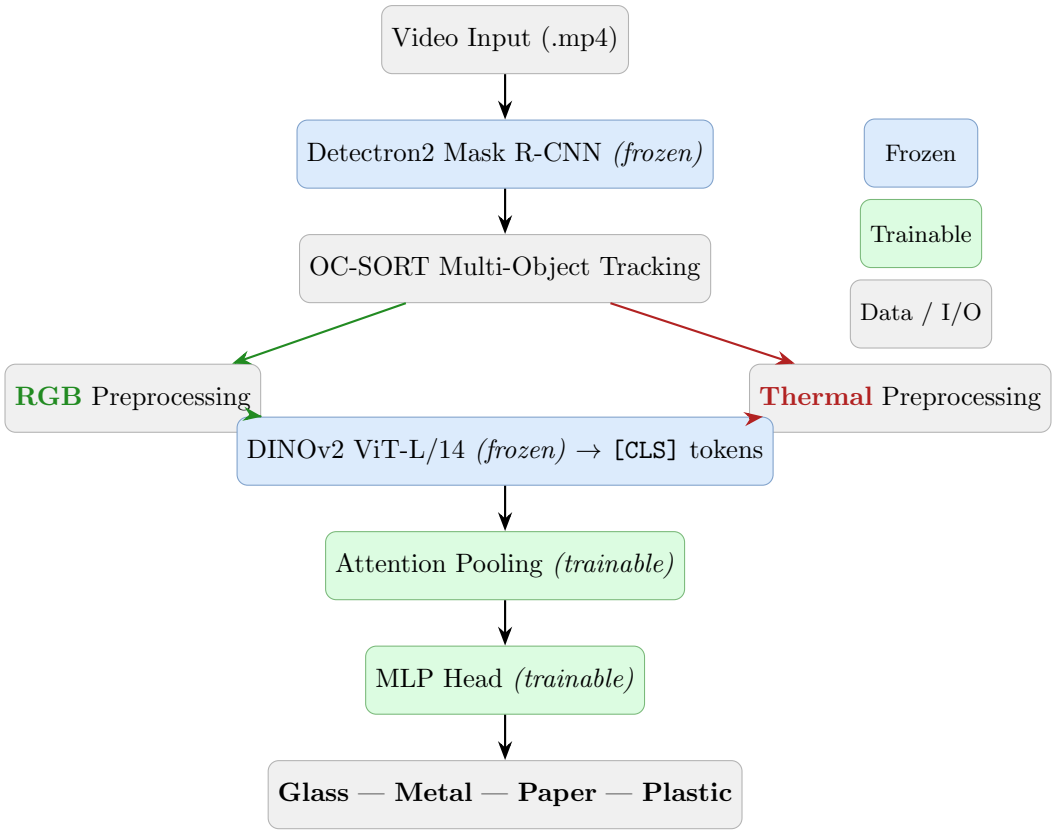


Figure 1: Unified pipeline architecture. Both RGB and thermal modalities share detection, tracking, DINOv2 feature extraction, and the trainable classification layers. Only preprocessing differs between modalities.

3 DINOv2 Feature Backbone

3.1 Architecture

DINOv2 (Oquab et al., 2024) is a family of self-supervised Vision Transformers trained via self-distillation on a curated dataset of 142M images (LVD-142M). We use the **ViT-L/14** variant:

Table 1: DINOv2 ViT-L/14 architecture specification.

Property	**Value**
Patch size	14×14 pixels
Input resolution	518×518 (37×37 patches)
Embedding dimension	1024
Transformer layers	24
Attention heads	16
Total parameters	∼304M
Output used	`[CLS]` token (1024-dim)
Training status	**Completely frozen**

3.2 Rationale for Freezing

Freezing the DINOv2 backbone yields several advantages:

1. **Data efficiency:** With only 550 labeled tracklets, fine-tuning 304M parameters risks severe overfitting. Training only ∼528K parameters (0.17% of backbone size) provides strong regularization.
2. **Feature caching:** Since the backbone is deterministic given an input, `[CLS]` tokens can be *precomputed once and cached to disk*. Subsequent training operates on cached 1024-dim vectors, eliminating GPU memory and compute costs of the backbone during training.
3. **Cross-modal transfer:** The same frozen features serve both RGB and thermal inputs, enabling controlled comparison of modality-specific information content.

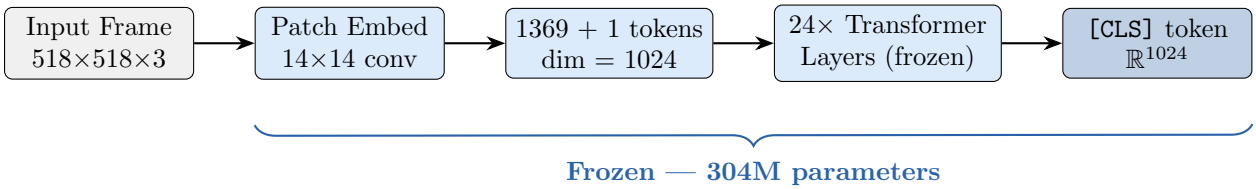


Figure 2: DINOv2 ViT-L/14 forward pass. The 518×518 input is split into 37×37 = 1369 patches plus one `[CLS]` token. All 304M parameters remain frozen; only the output `[CLS]` embedding is used.

4 RGB Classification Pipeline

4.1 Detection and Tracking

Object detection uses a frozen Detectron2 Mask R-CNN (ResNet-50-FPN backbone) trained on 4 waste material classes, producing per-frame bounding boxes and instance segmentation masks. Multi-object tracking is performed by OC-SORT with three deduplication mechanisms:

- **Observation-Centric Recovery (OCR):** a second association pass using last-observed position instead of Kalman prediction, recovering tracks where prediction has drifted.
- **Duplicate detection suppression:** discards unmatched detections overlapping with already-matched tracks.
- **Track merging:** merges tracks with consistent spatial overlap (IoU ≥ 0.7) for 3+ consecutive frames.

4.2 Per-Frame Preprocessing

For each tracked object, the pipeline extracts masked, cropped frames suitable for DINOv2 input:

Figure 3: RGB preprocessing pipeline. The instance mask isolates the object; background pixels are filled with neutral gray (128) to avoid DINOv2 normalization artifacts. BGR→RGB conversion occurs before normalization.

Critical design choice: Background pixels are filled with gray value 128, *not* black (0). After ImageNet normalization (mean= $[0.485, 0.456, 0.406]$, std= $[0.229, 0.224, 0.225]$), black pixels map to large negative values (≈ -2.1), contaminating the `[CLS]` representation. Gray ($128/255 \approx 0.502$) normalizes to values near zero, acting as a neutral filler.

4.3 Temporal Sampling and Multi-View Aggregation

Each tracklet spans hundreds to thousands of frames, representing *multiple views* of the same object as it moves along the conveyor. We uniformly sample $T = 8$ frames per tracklet:

$$\text{indices} = \lfloor \text{linspace}(0,\, N_{\text{frames}} - 1,\, 8) \rfloor$$

This provides temporal coverage across the tracklet's lifetime while keeping compute bounded. For tracklets shorter than 8 frames, all available frames are used (padded in the collation step).

4.4 Learnable Attention Pooling

Given T frame-level `[CLS]` embeddings $\{\mathbf{f}_t\}_{t=1}^{T} \in \mathbb{R}^{1024}$, we compute a single tracklet-level representation via learnable attention:

$$s_t = \frac{1}{\tau}\, \mathbf{w}_2^{\top} \tanh\big(\mathbf{W}_1 \mathbf{f}_t + \mathbf{b}_1\big) + b_2 \tag{1}$$

$$\alpha_t = \frac{\exp(s_t)}{\sum_{t'=1}^{T} \exp(s_{t'})} \tag{2}$$

$$\mathbf{g} = \sum_{t=1}^{T} \alpha_t\, \mathbf{f}_t \tag{3}$$

where $\mathbf{W}_1 \in \mathbb{R}^{256\times 1024}$, $\mathbf{w}_2 \in \mathbb{R}^{256}$, and temperature $\tau = 1.0$. The attention weights α_t are interpretable: they reveal which frames the model considers most informative for classification.

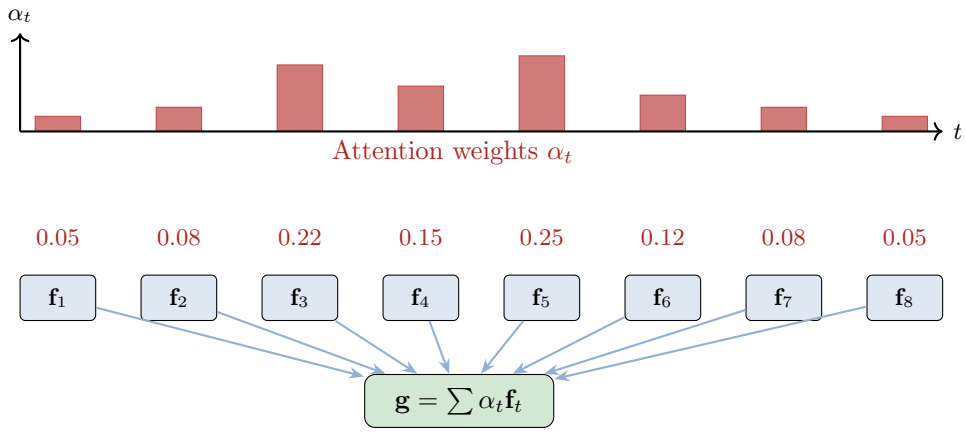


Figure 4: Attention pooling mechanism. The bar chart shows learned attention weights α_t over 8 sampled frames. Frames capturing distinctive object features (e.g., clear material boundaries) receive higher weight. The weighted sum produces a single 1024-dim tracklet representation $\mathbf{g}$.

4.5 MLP Classification Head

The pooled representation $\mathbf{g} \in \mathbb{R}^{1024}$ is classified by a lightweight MLP:

$$\hat{\mathbf{y}} = \mathbf{W}_4 \cdot \text{Dropout}_{0.3}\big(\text{GELU}\big(\mathbf{W}_3 \cdot \text{LayerNorm}(\mathbf{g}) + \mathbf{b}_3\big)\big) + \mathbf{b}_4$$

where $\mathbf{W}_3 \in \mathbb{R}^{256\times 1024}$, $\mathbf{W}_4 \in \mathbb{R}^{4\times 256}$, and $\hat{\mathbf{y}} \in \mathbb{R}^4$ are raw logits over the four material classes.

Table 2: Trainable parameter breakdown.

Component	Parameters	Details
Attention Pool	262,657	Linear($1024 \to 256$) + Linear($256 \to 1$)
MLP Head	265,476	LayerNorm + Linear($1024 \to 256$) + Linear($256 \to 4$)
Total trainable	**528,133**	**0.17%** of DINOv2 backbone

5 Thermal Classification Pipeline

The thermal pipeline reuses the *same* detection, tracking, attention pooling, and MLP head architecture as the RGB pipeline. The critical differences lie in (1) how input frames are obtained and (2) how RGB-space masks are transferred to thermal image coordinates.

5.1 Spatial Registration via Homography

Since the RGB and thermal cameras observe the scene from different viewpoints and at different resolutions (1280×720 vs. 320×240), spatial correspondence is established via a 3×3 homography matrix $\mathbf{H}$ computed per experiment using SuperPoint–SuperGlue feature matching with RANSAC outlier rejection. The registration achieves a mean reprojection error of 2.27 ± 0.48 pixels at thermal resolution with a 99.1% inlier rate.

5.2 Temporal Frame Matching

The two cameras lack hardware synchronization. Temporal alignment is established through an adaptive search procedure, producing a mapping:

$$\mathcal{M} : \text{RGB frame index} \to \text{thermal frame index}$$

stored as a CSV per experiment. Not all RGB frames have thermal correspondences—the mapping is sparse (e.g., 26,233 matched pairs out of 26,447 RGB frames for experiment 0798).

5.3 Cross-Modal Mask Warping

For each tracked object, the binary instance mask $M_{\text{RGB}} \in \{0,1\}^{H_r \times W_r}$ from Detectron2 (in RGB space) is warped to thermal space:

$$M_{\text{thermal}} = \text{WarpPerspective}\big(M_{\text{RGB}}, \mathbf{H}, (W_t, H_t)\big)$$

using nearest-neighbor interpolation to preserve the binary nature of the mask. This warped mask then serves the same role as the original RGB mask: isolating the object from the background.

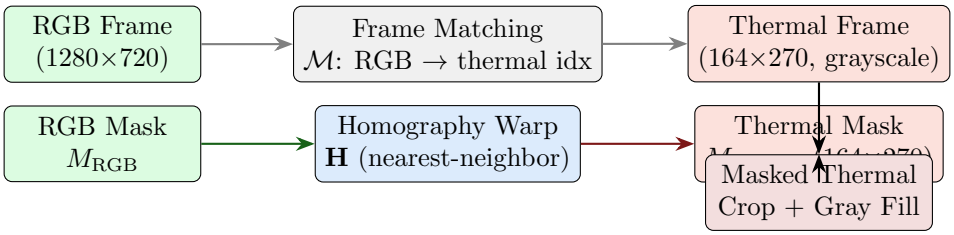


Figure 5: Cross-modal preprocessing. The RGB binary mask is warped to thermal coordinates via homography $\mathbf{H}$. The frame matching map $\mathcal{M}$ identifies the corresponding thermal frame for each RGB frame index.

5.4 Thermal Preprocessing Pipeline

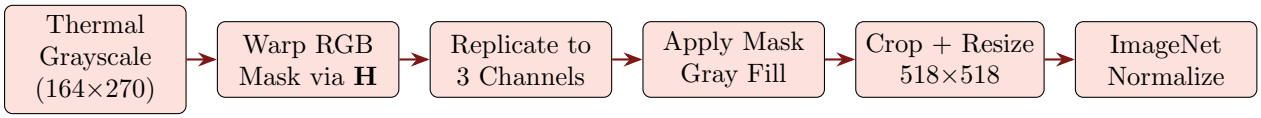


Figure 6: Thermal preprocessing pipeline. Two additional steps compared to RGB: (1) homography-based mask warping and (2) grayscale-to-3-channel replication for DINOv2 compatibility.

Grayscale → 3-channel conversion. DINOv2 expects 3-channel RGB input. The single-channel thermal intensity image $I \in \mathbb{R}^{H \times W}$ is converted by simple channel replication:

$$I_{3\text{ch}} = [I;\, I;\, I] \in \mathbb{R}^{H \times W \times 3}$$

This preserves the raw thermal intensity distribution without introducing false color artifacts. An alternative approach applies a perceptual colormap (e.g., Inferno) before replication, though the default configuration uses direct replication (`colormap: null`).

5.5 RGB vs. Thermal: Side-by-Side Comparison

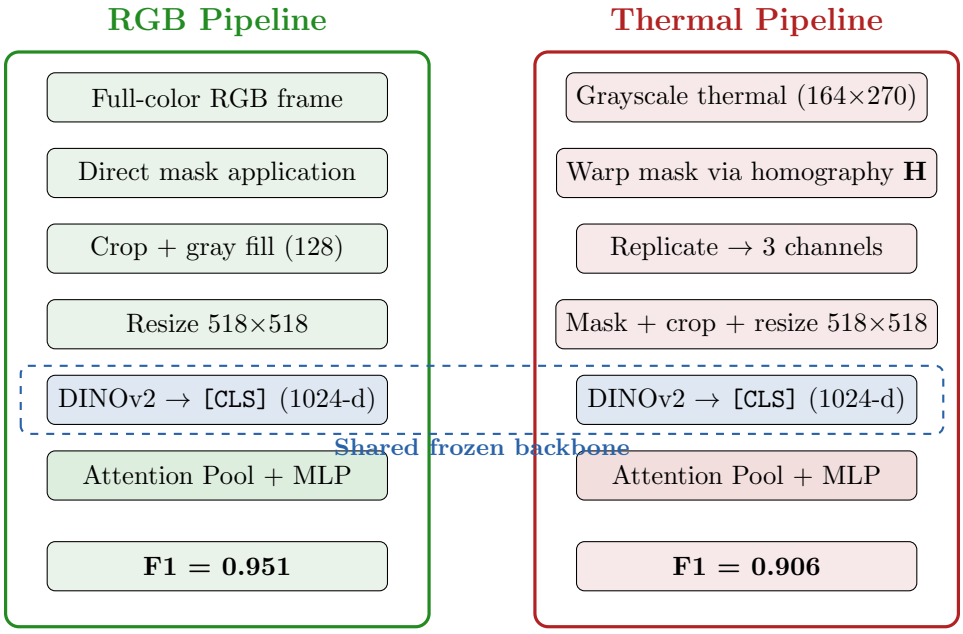


Figure 7: Side-by-side comparison of RGB and thermal pipelines. Both share the same frozen DINOv2 backbone and trainable classification layers. Key differences are highlighted: thermal requires homography warping and grayscale-to-3-channel conversion.

6 Late Fusion Pipeline

The late fusion pipeline combines RGB and thermal representations to exploit complementary information from both modalities. Rather than fusing raw inputs or intermediate features, we fuse at the *tracklet representation level*—after each modality has been independently pooled via its own attention mechanism.

6.1 Architecture

Each modality maintains an independent attention pool (Eq. 1–3) that learns modality-specific temporal attention weights. The pooled representations are concatenated and classified by a shared MLP head with doubled input dimension:

$$\mathbf{g}_{\text{rgb}} = \text{AttentionPool}_{\text{rgb}}\big(\{\mathbf{f}_t^{\text{rgb}}\}_{t=1}^{T}\big) \in \mathbb{R}^{1024} \tag{4}$$

$$\mathbf{g}_{\text{th}} = \text{AttentionPool}_{\text{th}}\big(\{\mathbf{f}_t^{\text{th}}\}_{t=1}^{T}\big) \in \mathbb{R}^{1024} \tag{5}$$

$$\mathbf{g}_{\text{fused}} = [\mathbf{g}_{\text{rgb}};\, \mathbf{g}_{\text{th}}] \in \mathbb{R}^{2048} \tag{6}$$

$$\hat{\mathbf{y}} = \text{MLPHead}(\mathbf{g}_{\text{fused}}) \in \mathbb{R}^{4} \tag{7}$$

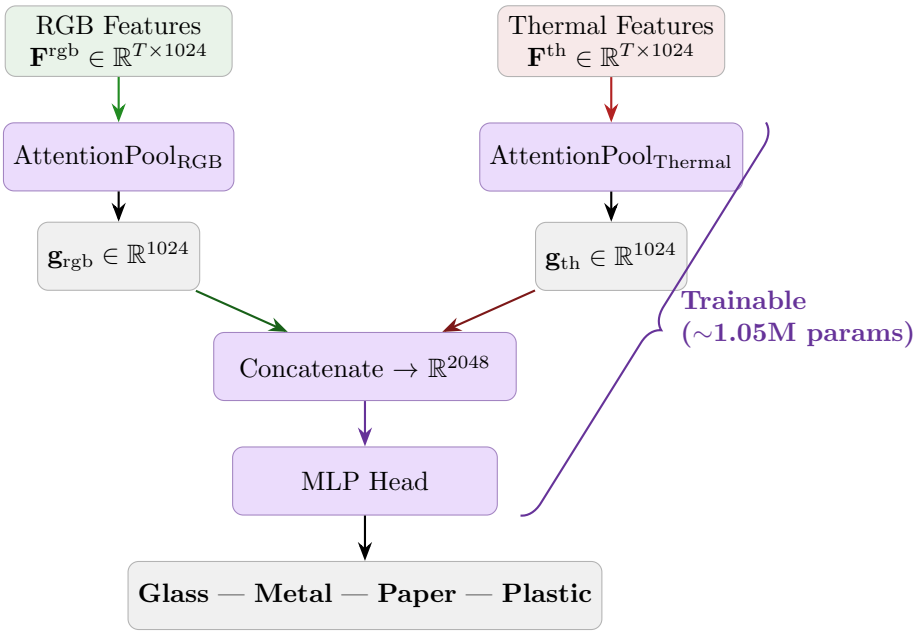


Figure 8: Late fusion architecture. RGB and thermal features are pooled independently via modality-specific attention mechanisms, concatenated into a 2048-dim vector, and classified by a shared MLP head. All components shown are trainable; DINOv2 feature extraction (not shown) remains frozen.

6.2 Parameter Count

The late fusion model roughly doubles the trainable parameter count compared to single-modal pipelines, due to the second attention pool and the wider MLP input layer:

Table 3: Parameter comparison: single-modal vs. late fusion.

Component	**Single-Modal**	**Late Fusion**
Attention Pool (RGB)	262,657	262,657
Attention Pool (Thermal)	—	262,657
MLP Head	265,476	527,620
Total trainable	**528,133**	**1,052,934**
% of DINOv2	0.17%	0.35%

The MLP head input dimension doubles from 1024 to 2048 for the fused model, accounting for the increase from 265K to 528K parameters. Even at ∼1.05M parameters, the fused model trains only 0.35% of the backbone size.

6.3 Design Rationale: Late vs. Early Fusion

We adopt late fusion (decision-level) over early fusion (feature-level or input-level) for three reasons:

1. **Modality-specific temporal attention.** RGB and thermal frames may carry discriminative information at different temporal points in the tracklet (e.g., thermal signatures are most distinctive during heating/cooling transitions). Independent attention pools allow each modality to learn its own temporal weighting.
2. **Graceful degradation.** Late fusion architectures can naturally handle missing modalities at inference time by zero-padding the absent branch, whereas early fusion requires both modalities at all times.
3. **Controlled ablation.** By sharing the DINOv2 backbone and varying only the fusion strategy, we can directly compare single-modal vs. multi-modal performance with minimal confounding factors.

7 Training Methodology

7.1 Feature Caching Strategy

Since DINOv2 is frozen, features are precomputed once per tracklet and cached to disk as PyTorch tensors:

$$\mathbf{F}_i = [\mathbf{f}_1, \ldots, \mathbf{f}_T]_i \in \mathbb{R}^{T \times 1024} \quad \text{saved as } \texttt{features/\{video\}\_track\_\{id\}.pt}$$

Training then operates purely on cached features without loading DINOv2 (∼304M parameters are never on GPU during training). The model used during training, `CachedMaterialClassifier`, contains only the attention pool and MLP head.

7.2 Optimization

Table 4: Training hyperparameters (identical for all three approaches: RGB, thermal, and fused).

Hyperparameter	**Value**
Optimizer	AdamW
Learning rate	1×10^{-3}
Weight decay	1×10^{-4}
Batch size	8
Epochs	20
Loss function	CrossEntropyLoss
Label smoothing	0.1
Warmup schedule	LinearLR, 5 epochs ($10^{-8} \rightarrow 1.0$)
Main schedule	CosineAnnealingLR, 15 epochs
Seed	42

7.3 Learning Rate Schedule

The two-phase schedule combines linear warmup with cosine annealing:

$$\eta(e) = \begin{cases} \eta_0 \cdot \left(10^{-8} + (1 - 10^{-8}) \cdot e/5\right) & \text{if } e < 5 \\ \frac{\eta_0}{2}\left(1 + \cos\left(\pi \cdot (e-5)/15\right)\right) & \text{if } e \geq 5 \end{cases}$$

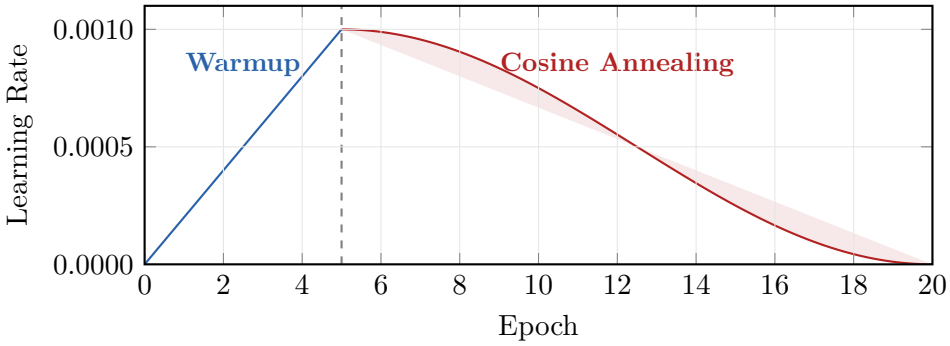


Figure 9: Learning rate schedule. Linear warmup from near-zero to 10^{-3} over 5 epochs, followed by cosine decay over the remaining 15 epochs.

7.4 Data Augmentation

Augmentation is applied *per-frame after masking and cropping*, with consistent random state across all T frames in a tracklet to maintain spatial coherence:

Table 5: Data augmentation configuration. Saturation and hue jitter are disabled for thermal (grayscale input).

Augmentation	**RGB**	**Thermal**
Random horizontal flip	$p = 0.5$	$p = 0.5$
Brightness jitter	± 0.2	± 0.2
Contrast jitter	± 0.2	± 0.2
Saturation jitter	± 0.1	*disabled* (0.0)
Hue jitter	± 0.05	*disabled* (0.0)
Random erasing	$p = 0.1$	$p = 0.1$

Fused training. The late fusion model trains on pre-cached features from both modalities simultaneously. No additional augmentation is applied beyond what was used during feature caching. The same optimizer, schedule, and hyperparameters (Table 4) are used for all three pipelines to ensure a fair comparison.

8 Experimental Results

8.1 Dataset

The dataset comprises **19 experiment videos** yielding **550 labeled tracklets** (filtered to $\geq$1000 frames per tracklet for quality):

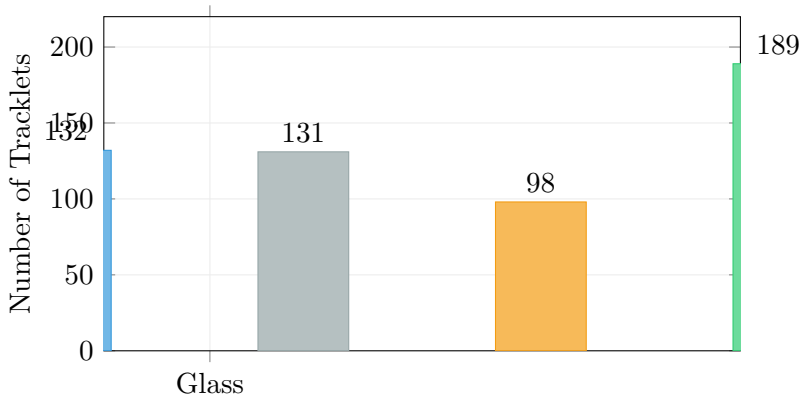


Figure 10: Class distribution across 550 labeled tracklets. Plastic is the most frequent class (34.4%), paper the least (17.8%).

8.2 RGB Results: 5-Fold Stratified Cross-Validation

Table 6: RGB pipeline: 5-fold stratified CV results (550 tracklets, 19 videos).

Metric	Value
Mean Accuracy	0.9509 ± 0.0093
Mean Macro F1	0.9507 ± 0.0110

Table 7: RGB pipeline: per-class metrics pooled across all 5 folds.

Class	Precision	Recall	F1	Count
Glass	0.978	0.992	0.985	132
Metal	0.921	0.977	0.948	131
Paper	0.957	0.908	0.932	98
Plastic	0.951	0.926	0.938	189

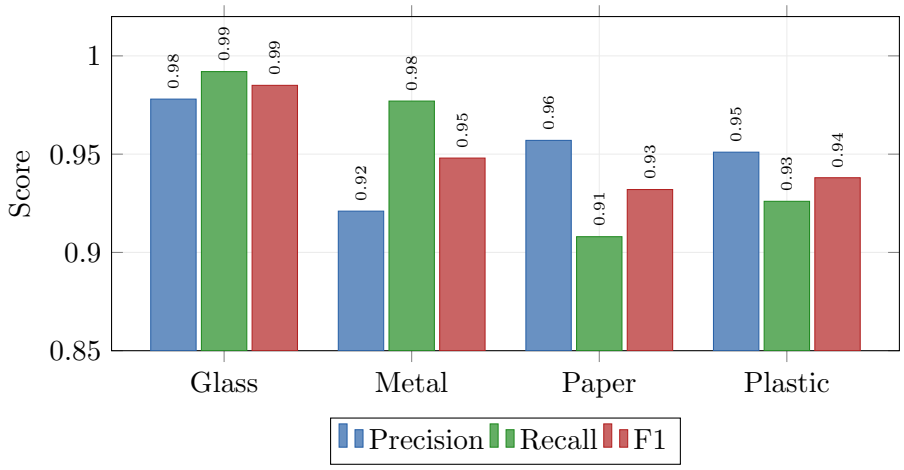


Figure 11: RGB pipeline per-class performance. Glass achieves near-perfect classification (F1=0.985). Paper and plastic show the most room for improvement.

8.3 Confusion Matrix Analysis

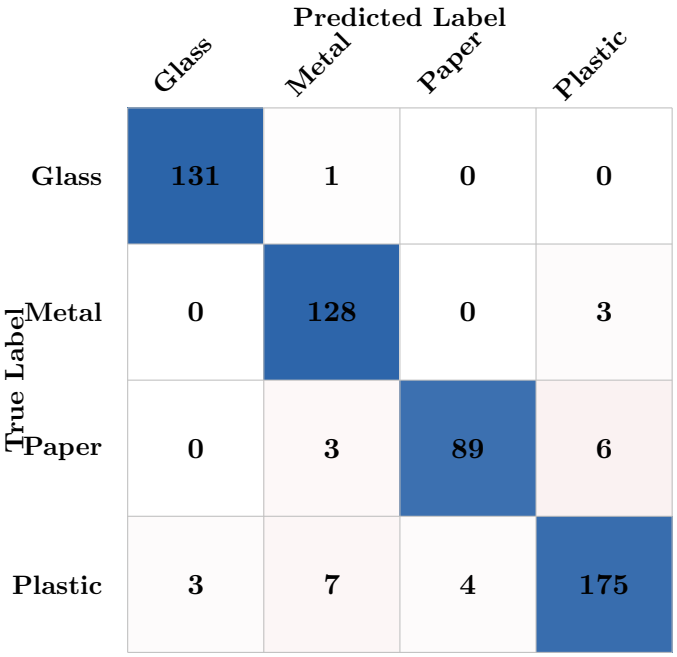


Figure 12: RGB confusion matrix (pooled across 5 folds, 550 samples). Only 27 misclassifications. Glass is near-perfect (131/132). Most confusion occurs between plastic↔metal (10 errors) and plastic↔paper (10 errors).

Error analysis: The 27 misclassifications reveal interpretable patterns:

- **Glass** (1 error): Near-perfect—distinctive visual appearance (transparency, reflectivity).
- **Metal** (3 errors): 3 samples confused with plastic, likely due to similar surface finish on certain items.
- **Paper** (9 errors): Most challenging class; confused with both metal (3) and plastic (6), suggesting visual similarity in deformed/wrinkled paper.
- **Plastic** (14 errors): Largest error count but from the largest class (189 samples). Confusions spread across metal (7), paper (4), and glass (3).

8.4 Thermal Results: 5-Fold Stratified Cross-Validation

Table 8: Thermal pipeline: 5-fold stratified CV results (550 tracklets, 19 videos).

Metric	Value
Mean Accuracy	0.9127 ± 0.0384
Mean Macro F1	0.9056 ± 0.0405

Table 9: Thermal pipeline: per-class metrics pooled across all 5 folds.

Class	Precision	Recall	F1	Count
Glass	0.956	0.985	0.970	132
Metal	0.938	0.924	0.931	131
Paper	0.792	0.816	0.804	98
Plastic	0.929	0.905	0.917	189

Error analysis: The 48 thermal misclassifications reveal distinct patterns compared to RGB:

- **Glass** (2 errors): Remains the easiest class; thermal emissivity is highly distinctive.
- **Metal** (10 errors): Confused with glass (4), paper (5), and plastic (1). Metal's variable thermal conductivity creates diverse thermal signatures.
- **Paper** (18 errors): Most challenging; 11 confused with plastic and 5 with metal. Paper and plastic share similar thermal mass, making spatial thermal features less discriminative.
- **Plastic** (18 errors): Confused primarily with paper (16) and metal (2). The paper↔plastic confusion dominates thermal errors.

8.5 Late Fusion Results: 5-Fold Stratified Cross-Validation

Table 10: Late fusion pipeline: 5-fold stratified CV results (550 tracklets, 19 videos).

Metric	Value
Mean Accuracy	0.9636 ± 0.0100
Mean Macro F1	0.9602 ± 0.0136

Table 11: Late fusion pipeline: per-class metrics pooled across all 5 folds.

Class	Precision	Recall	F1	Count
Glass	1.000	0.985	0.992	132
Metal	0.949	0.985	0.966	131
Paper	0.966	0.878	0.920	98
Plastic	0.949	0.979	0.964	189

8.6 Comparative Analysis

8.6.1 Three-Way Method Comparison

Table 12 presents all methods on a common footing, including prior thermal baselines and the three DINOv2-based approaches.

Table 12: Complete classification methods comparison. Prior baselines use handcrafted temporal features from thermal intensity time-series. DINOv2 approaches use frozen spatial features. All DINOv2 results are 5-fold stratified CV on 550 tracklets; per-class values are pooled F1 scores.

Method	Features	Params	Acc.	F1	Glass	Metal	Paper	Plastic
Temporal thermal features (prior work)								
SVM (5 features)	Handcrafted	—	68.2	0.678	0.786	0.490	0.732	0.703
BiGRU	Learned	—	69.1	0.676	0.815	0.636	0.537	0.716
InceptionTime	Learned	—	68.2	0.672	0.764	0.609	0.615	0.700
MiniRocket	Kernel-based	—	65.5	0.644	0.754	0.553	0.611	0.658
Gundupalli	Peak histogram	—	55.5	0.550	0.760	0.370	0.540	0.530
DINOv2 spatial features (this work)								
RGB-only	Frozen ViT-L	528K	95.1	0.951	0.985	0.948	0.932	0.938
Thermal-only	Frozen ViT-L	528K	91.3	0.906	0.970	0.931	0.804	0.917
Late Fusion	Frozen ViT-L	1.05M	**96.4**	**0.960**	**0.992**	**0.966**	**0.920**	**0.964**

8.6.2 Per-Fold Comparison

Table 13 shows per-fold results, enabling direct comparison of fold-level consistency across methods. All three approaches use identical `StratifiedKFold(5, shuffle=True, random_state=42)` splits.

Table 13: Per-fold macro F1 scores for the three DINOv2-based approaches. Bold indicates the best method per fold.

	RGB		Thermal		Late Fusion	
Fold	Acc	F1	Acc	F1	Acc	F1
1	0.945	0.939	0.855	0.848	0.945	**0.935**
2	0.955	0.957	0.927	0.923	0.973	**0.970**
3	0.964	0.965	0.945	0.941	0.964	**0.959**
4	0.936	0.937	0.882	0.868	0.964	**0.965**
5	0.955	0.957	0.955	0.949	0.973	**0.973**
Mean	0.951	0.951	0.913	0.906	**0.964**	**0.960**
±Std	±0.009	±0.011	±0.038	±0.041	**±0.010**	**±0.014**

Late fusion achieves the highest F1 in all 5 folds and exhibits the lowest variance (±0.014 vs. ±0.011 for RGB and ±0.041 for thermal), demonstrating both superior performance and stability. Notably, fusion is most beneficial when one modality underperforms: in Fold 4, where thermal F1 drops to 0.868, fusion achieves 0.965—substantially exceeding both individual modalities.

8.6.3 Per-Class F1 Comparison

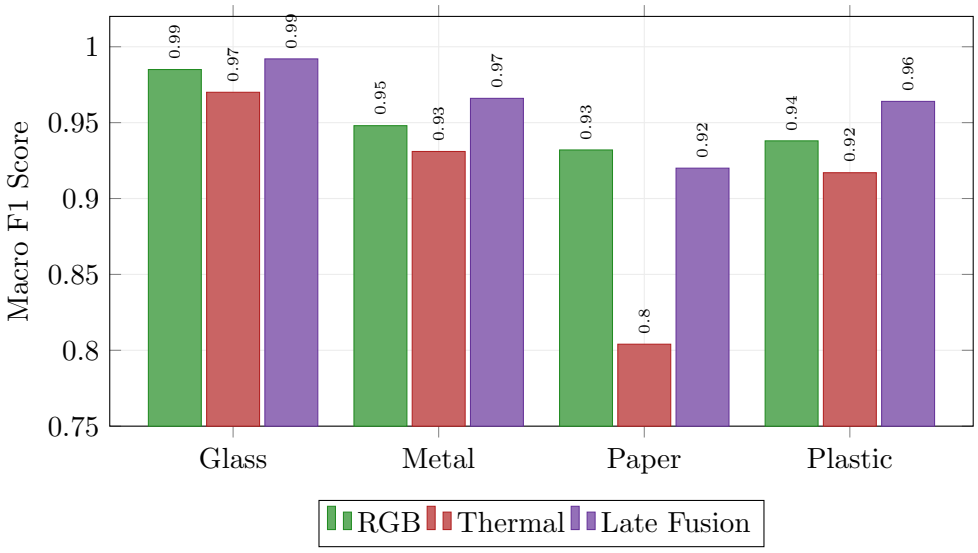
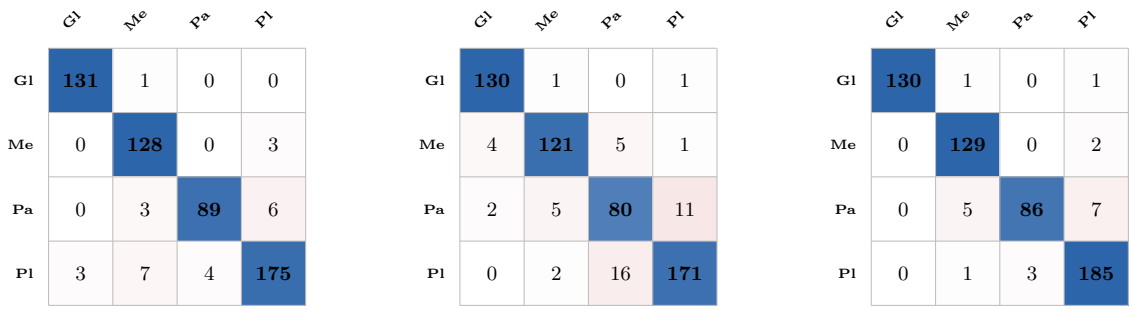


Figure 13: Per-class F1 comparison across the three DINOv2-based approaches. Fusion improves over RGB-only for glass (+0.7 pp), metal (+1.8 pp), and plastic (+2.6 pp). Paper F1 decreases slightly with fusion (0.932→0.920) due to persistent paper↔plastic confusion in the thermal branch.

8.6.4 Confusion Matrix Comparison



(a) RGB (27 errors) (b) Thermal (48 errors) (c) Late Fusion (20 errors)

Figure 14: Side-by-side confusion matrices (pooled across 5 folds, 550 samples each). Late fusion reduces total errors from 27 (RGB) and 48 (thermal) to 20. Key improvements: metal→glass errors eliminated (4→0), plastic→metal errors reduced (7→1), and plastic→paper errors reduced (4→3). The dominant remaining error pattern is paper↔plastic confusion.

8.6.5 Analysis of Error Patterns Across Modalities

The three confusion matrices (Figure 14) reveal complementary error patterns:

- **Glass:** Near-perfect in all three modalities (F1 $\geq$ 0.970). Both RGB (transparency, reflectivity) and thermal (distinctive emissivity) provide strong discriminative signals.
- **Metal:** RGB excels at metal recall (0.977) while thermal is weaker (0.924). Fusion improves metal recall to 0.985, suggesting that RGB resolves thermal ambiguities for reflective surfaces. The metal→glass confusion present in thermal (4 errors) is completely eliminated by fusion.
- **Paper:** The most challenging class across all modalities. Thermal performs poorly (F1=0.804), with 18 errors primarily from paper↔plastic confusion. RGB performs much better (F1=0.932). Fusion achieves F1=0.920, a slight decrease from RGB alone, suggesting that thermal's paper confusion partially propagates through the fusion mechanism.
- **Plastic:** Fusion provides the largest improvement (RGB F1=0.938 → fusion F1=0.964, +2.6 pp). RGB's plastic→metal confusion (7 errors) is substantially reduced to 1 in fusion, indicating that thermal information helps disambiguate plastic from metal.

Overall, late fusion reduces total errors by 26% relative to RGB alone (27→20) and by 58% relative to thermal alone (48→20). The improvement is largest for classes where the two modalities have complementary error patterns (metal, plastic), and smallest where both modalities share the same confusion (paper↔plastic).

9 Discussion

9.1 Why DINOv2 Features Transfer Across Modalities

DINOv2's self-supervised pretraining learns visual representations that capture structural and textural properties rather than color-specific features. This makes the `[CLS]` token inherently robust to domain shift:

- **Texture sensitivity:** DINOv2 encodes surface texture, edges, and shape—properties visible in both RGB and thermal modalities (thermal images reveal surface roughness and material boundaries through emissivity gradients).
- **Color invariance:** The self-distillation training objective encourages representations that are invariant to augmentations including color jitter, making the model less dependent on RGB-specific color information.

- **Grayscale compatibility:** By replicating the single thermal channel to 3 channels, DINOv2 processes thermal images as "grayscale photographs"—a domain well-represented in its 142M training images.

9.2 Spatial vs. Temporal Thermal Features

The prior thermal baselines (Table 12) exploit *temporal* thermal dynamics—how material temperature changes over time as objects heat and cool. The DINOv2-based thermal pipeline instead uses *spatial* features from individual thermal frames, achieving a mean macro F1 of 0.906—a 23 percentage point improvement over the best temporal baseline (BiGRU, F1=0.676). These approaches capture complementary information:

Table 14: Spatial vs. temporal thermal feature paradigms.

	Temporal Features	**Spatial Features (DINOv2)**
Input	Thermal intensity time-series $I(t)$	Individual thermal frames $I(x, y)$
Captures	Heating/cooling dynamics, thermal conductivity, specific heat	Surface texture, shape, emissivity patterns, spatial temperature distribution
Robustness	Requires calibrated heating zone	Works with any thermal frame
Architecture	SVM, RNNs, 1D-CNNs	Frozen ViT + attention pooling

9.3 Complementarity of RGB and Thermal Modalities

The per-fold comparison (Table 13) and confusion matrices (Figure 14) reveal that RGB and thermal errors are partially complementary:

- **RGB strengths:** Superior paper classification (F1=0.932 vs. 0.804 thermal) and overall higher accuracy. RGB captures color and fine texture cues that distinguish paper from plastic.
- **Thermal strengths:** Thermal helps disambiguate plastic from metal—fusion reduces plastic→metal errors from 7 to 1 compared to RGB alone. Thermal emissivity signatures provide material composition cues invisible to RGB.
- **Shared weakness:** Both modalities struggle with paper↔plastic, which remains the dominant error pattern even after fusion (12 of 20 fused errors). This suggests that paper and plastic share both visual and thermal similarities in our dataset.

The fusion improvement (+1.0 pp mean F1 over RGB, with 26% fewer errors) demonstrates that the modalities provide complementary rather than redundant information, despite the large performance gap between them individually.

9.4 Parameter Efficiency

All three DINOv2-based approaches are remarkably parameter-efficient:

- **Single-modal:** 528K trainable parameters (0.17% of the 304M backbone). Trains in ∼2 minutes on cached features.
- **Late fusion:** 1.05M trainable parameters (0.35% of backbone). Despite doubling the parameter count, the fused model remains lightweight. The additional cost is one extra attention pool (263K) and a wider MLP input layer (262K extra).
- **Feature caching eliminates runtime overhead:** Since both RGB and thermal features are precomputed, fusion training is no more expensive than single-modal training—the model never loads the 304M-parameter DINOv2 backbone.

9.5 Practical Implications

1. **Minimal training footprint:** With only ∼528K–1.05M trainable parameters and feature caching, training completes in minutes on a single GPU. No fine-tuning of the 304M-parameter backbone is required.
2. **Modality-agnostic architecture:** The same attention pool + MLP head design works for RGB, thermal, and fused inputs, differing only in the preprocessing stage and feature concatenation.
3. **Interpretable attention:** The learned attention weights α_t (Eq. 2) reveal which temporal views are most discriminative, potentially guiding camera placement or exposure timing in deployment.
4. **Multi-modal deployment trade-off:** Late fusion provides a 1.0 pp F1 improvement over RGB-only at the cost of requiring a synchronized thermal camera and spatial calibration. For applications where near-perfect accuracy is critical (e.g., regulatory compliance), the additional hardware is justified; for cost-sensitive deployments, RGB-only at 95.1% F1 may suffice.

10 Conclusion

This report presented a three-way material classification system for conveyor belt waste sorting, comparing RGB-only, thermal-only, and late fusion pipelines—all built on frozen DINOv2 ViT-L/14 features. The key findings are:

1. **Frozen foundation model features are highly effective** for material classification. RGB achieves **95.1% macro F1** with 528K trainable parameters (0.17% of backbone). Thermal achieves **90.6% macro F1** with the same architecture, a 23 pp improvement over the best prior temporal baseline.
2. **Late fusion improves over both individual modalities,** achieving **96.0% macro F1** with 1.05M trainable parameters (0.35% of backbone). Fusion reduces classification errors by 26% relative to RGB alone, with the largest gains for metal and plastic.
3. **RGB and thermal provide complementary error patterns.** RGB excels at paper discrimination while thermal helps disambiguate plastic from metal. Fusion exploits this complementarity, though the shared paper↔plastic confusion persists.
4. **Attention pooling provides an interpretable aggregation mechanism** for multi-view tracklet classification, and the modality-specific pools in the fusion model enable each modality to learn its own temporal weighting.
5. **Feature caching enables rapid experimentation** by decoupling the expensive backbone forward pass from the lightweight training loop. All three pipelines train in minutes on cached features.

The late fusion pipeline achieves near-perfect glass classification (F1=0.992), strong metal and plastic performance (F1≥0.964), and reduces total misclassifications to just 20 out of 550 tracklets. The remaining errors are concentrated in paper↔plastic confusion, suggesting that future improvements should focus on features that capture the physical differences between these materials—potentially combining spatial DINOv2 features with temporal thermal dynamics.