DECOUPLING TEMPORAL ACTION OWNERSHIP AND ACTOR GEOMETRY FOR SPATIO-TEMPORAL ACTION LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatio-temporal action localization must identify what action occurs, when it starts and ends, and where each actor is in every frame. We introduce YOLO-ST, a detector that separates temporal action ownership from spatial actor geometry. A Kinetics-pretrained VideoMAE-L pathway supplies semantics, a frozen $(2+1)$D pathway that sees every frame supplies motion near likely actors, and a frozen COCO-pretrained YOLO11-L pyramid supplies frame appearance. Dense per-frame heads own geometry, tube queries own actor identity and temporal extent, and at inference a training-free snap moves each linked tube box toward the dense detection of the same class. Evaluating on UCF101-24, we find that published frame mAP is not one metric: on identical predictions, the conventions in use span 6.6 points, from 88.47 under the all-frame evaluator that ROAD recommends to 95.10 under our re-implementation of the one released with YOWOFormer. Under the latter our checkpoint exceeds YOWOFormer-L's reported 93.35; under ROAD's two all-frame evaluators, whose protocol BMViT cites, it scores 88.47 and 89.30, below BMViT's 90.7, because it fires on frames outside the annotated actions, a failure that gating by the tubes' own temporal extent does not repair. Video mAP is 89.82 / 73.39 / 34.97 at spatio-temporal overlap 0.2, 0.5, and 0.5:0.95, level with STAR-L on the first two and 0.83 below it on the third. Substantial gaps to the strongest results remain on JHMDB-21 (91.5 frame), AVA v2.2 (33.23) and MultiSports (42.11 frame).

1 INTRODUCTION

Spatio-temporal action localization (STAL) must answer three coupled questions: *what* action occurs, *when* it occurs, and *where* the responsible actor is in every frame. This is harder than video classification because a detector must retain fine spatial evidence while reasoning over long and variable temporal extents. It is also harder than frame detection because independent boxes must be assigned to persistent actors and assembled into action tubes. UCF101-24 (Soomro et al., 2012; Singh et al., 2017), JHMDB-21 (Jhuang et al., 2013), MultiSports (Li et al., 2021), and AVA (Gu et al., 2018) expose different parts of this problem: untrimmed videos with complete tubes, short trimmed clips, dense multi-person sports motion, and sparse multi-label keyframes.

Real-time systems in the YOWO family combine a 2D keyframe branch with a 3D clip branch (Köpüklü et al., 2019; Yang & Dai, 2023; Nguyen et al., 2026). YOWOFormer pairs VideoMAE with a YOLO11 frame detector through a spatial query adapter and a learned fusion module (Thakaew et al., 2026). These models are strong frame localizers, but the prediction remains centered on one keyframe. Conversely, MOC (Li et al., 2020c), TubeR (Zhao et al., 2022), and STAR (Gritsenko et al., 2024) explicitly represent trajectories or tubelets. Their temporal coherence is useful at video level, yet a tube owner can still produce less accurate frame-local boxes than a dense detector.

We study this mismatch as an *ownership–geometry decomposition*. Temporal ownership determines which actor performs an action and for how long; actor geometry determines the box at each instant. YOLO-ST keeps the two on separate heads over one pyramid built from three pathways: Kinetics-pretrained VideoMAE-L tokens for semantics, a frozen $(2+1)$D motion trunk that processes all

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

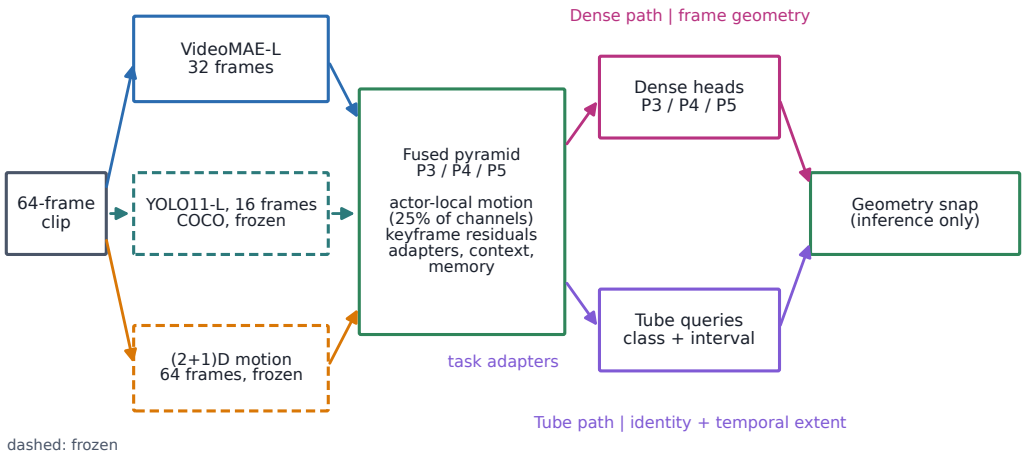

Figure 1: YOLO-ST at a glance. A Kinetics-pretrained VideoMAE-L pathway, a frozen COCO-pretrained YOLO11-L frame pyramid, and a frozen $(2+1)$D motion pathway (dashed) are fused into one P3–P5 pyramid, with motion added actor-locally to a quarter of the channels. Dense heads own frame geometry; tube queries, reached through task adapters, own actor identity and temporal extent. At inference a training-free geometry snap moves each linked tube box toward the dense head's same-class detection; it changes no dense prediction.

64 frames of a clip, and a frozen COCO-pretrained YOLO11-L pyramid for frame appearance. Dense heads own frame geometry, tube queries own identity and temporal extent, and at inference a training-free geometry snap lets each tube borrow the dense head's box.

Measuring this model exposed a problem with the benchmark rather than the method. UCF101-24 frame mAP is computed over annotated frames or over every frame, with every-point or trapezoidal integration, and with inclusive or exclusive box extents, and papers rarely state which. On the same predictions these choices move our checkpoint across 6.6 points, more than separates the methods they are used to rank. Our contributions are:

- YOLO-ST, a dense STAL detector that keeps frame geometry and tube ownership on separate heads over one pyramid and is level with STAR-L on video AP20 and AP50 (Table 1).
- A measurement of UCF101-24 frame-mAP conventions: one set of our predictions scored under each (88.47 to 95.10), with every published result placed beside the convention it uses (Table 1, Appendix A). We exceed the strongest YOWO-family result under its own evaluator, trail BMViT under the all-frame protocol it cites, and trace the gap to detections on action-free frames.
- A held-out selection protocol whose inference settings are frozen before the test evaluation, applied on four benchmarks, with the selection history and negative results reported (Appendix B).

2 Related Work

2.1 Single-Stage Action Detection

YOWO introduced a unified two-branch CNN that fuses 3D clip features with 2D keyframe features (Köpüklü et al., 2019). YOWOv2 adds multilevel, anchor-free detection and dynamic assignment (Yang & Dai, 2023); YOWOv3 updates the detector while reducing its parameter and compute cost (Nguyen et al., 2026). WOO shares a video backbone between actor localization and action classification (Chen et al., 2021), while progressive cross-stream cooperation iteratively exchanges region

proposals and features between RGB and optical-flow streams (Su et al., 2019). YOWOFormer pairs VideoMAE tokens with a YOLO11 detector through a spatial query adapter and a learned fusion module (Thakaew et al., 2026). YOLO-ST draws on the same two families of pretrained networks, but emits dense predictions across the clip, retains explicit temporal rates, and separates frame and tube responsibilities.

2.2 Tube and Transformer Models

MOC represents actions as moving centers with separate center, movement, and box heads (Li et al., 2020c). TubeR directly decodes variable-length tubelets and predicts action switches (Zhao et al., 2022). HIT incorporates person, object, hand, and pose interactions with temporal memory (Faure et al., 2023). EVAD retains keyframe and motion-relevant tokens and refines actor features with scene context (Chen et al., 2023). BMViT applies bipartite matching directly to multiscale backbone tokens (Ntinou et al., 2024), while STAR predicts coherent tubelets end to end without person proposals or non-maximum suppression (NMS) (Gritsenko et al., 2024). These methods show the value of actor-level temporal structure. Our focus is the complementary problem of preserving dense frame geometry while tube ownership is learned.

2.3 Multirate Video Representation

Factorizing a 3D kernel into spatial and temporal operations improves optimization and video representation (Tran et al., 2018). SlowFast (Feichtenhofer et al., 2019) and MViT (Fan et al., 2021) likewise show that video evidence benefits from multiple rates or scales. VideoMAE learns strong video semantics through high-ratio tube masking (Tong et al., 2022). YOLO-ST combines this pretrained semantic prior with an explicit dense motion pyramid. Unlike a spatial feature pyramid network (FPN) (Lin et al., 2017), its levels differ in both spatial scale and temporal rate. MultiSports is especially relevant because rapid motion, occlusion, and simultaneous actors expose failures hidden by UCF101-24 (Li et al., 2021; Singh et al., 2023).

3 Method

3.1 Overview

Let $X \in \mathbb{R}^{3\times T\times H\times W}$ be a $T = 64$ frame clip. YOLO-ST predicts dense outputs $D_l = \{S_l, O_l, B_l\}$ (class scores, objectness, boxes) at pyramid levels $l \in \{3, 4, 5\}$ for every frame, and a set of tube queries $\mathcal{Q}$, each with a class, per-frame boxes and visibility, and an action interval. Figure 1 summarizes the model. The dense heads own frame geometry; the tube queries own actor identity and temporal extent.

3.2 Three Pathways and One Pyramid

Semantics. A VideoMAE-Large encoder (Tong et al., 2022), pretrained and fine-tuned on Kinetics-400 (Kay et al., 2017), reads 32 of the 64 frames, sampled uniformly; its last 8 of 24 blocks are fine-tuned at $0.1\times$ the learning rate. A pointwise projection reduces its tokens to 256 channels and two residual depthwise temporal adapters (kernel 5) mix them over time. The result is interpolated to the P3 grid (all T frames, stride 8) and the P4 grid ($T/2$, stride 16) and refined by 3D convolutions, and P5 ($T/4$, stride 32) is pooled from P4. A residual factorized block of depthwise $(1, 3, 3)$, depthwise $(3, 1, 1)$ and pointwise convolutions, the *pyramid adapter*, then refines each level.

Frame appearance. A frozen COCO-pretrained YOLO11-L backbone and neck (Jocher & Qiu, 2024; Lin et al., 2014) run on 16 uniformly sampled frames; their P3–P5 features enter the pyramid through zero-initialized gated residuals.

Pyramid context. Each level then passes a global temporal context gate and a two-layer temporal attention over one saliency-pooled actor token per frame. P3 also carries a *cross-clip memory*: a GRU updated once per 8-frame chunk of pooled actor tokens modulates the P3 features, is trained

with a consistency loss (weight 0.05) between forward and backward passes over the chunks, and at inference carries its state from one clip to the next.

Motion. The motion pathway is the backbone and neck of a separately trained YOLO-ST $(2+1)$D detector (trained on UCF101-24 split-1 train); it is frozen, and only its fusion into the pyramid is learned. Each block factorizes a 3D kernel (Tran et al., 2018):

$$F_k^m = \sigma_o\big(K_k^t * \sigma_i(K_k^s * F_{k-1}^m)\big)\,, \tag{1}$$

where k indexes successive convolutions, K^s has support $(1,3,3)$, K^t has support $(3,1,1)$, and σ_i, σ_o are BatchNorm followed by ReLU and SiLU. The stem and first stage process all 64 frames; a path-aggregation neck fuses the stages at $T/2$, $T/4$ and $T/8$ and upsamples them temporally to T, $T/2$ and $T/4$, giving motion maps M_3, M_4, M_5 aligned with P3–P5.

Actor-local fusion. Motion added everywhere also injects background change, so we add it last, gated by a soft actor mask A_l: the detached sigmoid of the objectness that the dense head predicts from the pyramid before motion is added. Only the last quarter of the channels is updated:

$$\widetilde{F}_l = \big[F_l^{(1)},\ F_l^{(2)} + \alpha_l \bar{A}_l \odot G_l \odot \big(P_l(M_l) + \beta_l P_l'(\Delta_t M_l)\big)\big], \tag{2}$$

where $F_l^{(1)}$ and $F_l^{(2)}$ hold the first 75% and last 25% of the channels, $\bar{A}_l = 0.1 + 0.9A_l$ keeps a floor of motion everywhere, P_l and P_l' project the motion maps and their signed frame-to-frame differences, G_l is a sigmoid gate computed from $F_l^{(2)}$, both projections, and A_l, and $\alpha_l = \tanh a_l$, $\beta_l = \tanh b_l$ with a_l, b_l initialized to zero.

3.3 Dense and Tube Heads

Dense heads predict class scores, objectness, and a 17-bin distribution over box offsets at every P3–P5 location of every frame, trained with task-aligned assignment, binary cross-entropy (weight 0.5), CIoU (7.5), distribution focal loss (Li et al., 2020b) (1.5), and objectness (1.5); an attention from one query per class to each level refines the class scores. The tube decoder uses eight actor slots with two interval queries each, 16 hypotheses in total, represented at 32 temporal positions; six layers that alternate attention across queries, cross-attention to the pyramid, and attention over time predict class logits, per-frame boxes, visibility, interval endpoints, and boundary logits. Queries are assigned to ground-truth tubes by Hungarian matching (Carion et al., 2020) and trained with

$$\begin{aligned}\mathcal{L}_{\text{tube}} = \lambda_c\mathcal{L}_{\text{cls}} + \lambda_b\mathcal{L}_1 + \lambda_g\mathcal{L}_{\text{GIoU}} + \lambda_v\mathcal{L}_{\text{vis}} + \lambda_d\mathcal{L}_{\text{bnd}} \\ + \lambda_e(\mathcal{L}_{\text{start}} + \mathcal{L}_{\text{end}}) + \lambda_i\mathcal{L}_{\text{int}} + \lambda_o\mathcal{L}_{\text{cov}} + \lambda_f\mathcal{L}_{\text{frag}},\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{cls}}$ is softmax cross-entropy on matched queries, $\mathcal{L}_1$ is smooth-ℓ_1 (β=0.05), $\mathcal{L}_{\text{GIoU}}$ is generalized IoU (Rezatofighi et al., 2019), and $(\lambda_c, \lambda_b, \lambda_g, \lambda_v, \lambda_d, \lambda_e, \lambda_i, \lambda_o, \lambda_f) = (1, 5, 2, 1, 0.25, 0.5, 2, 1, 0.5)$. The coverage and fragmentation terms discourage partial or split ownership. The total loss adds the dense loss and the memory term.

Task isolation. To keep tube training from reshaping the features the dense heads read, tube gradients are stopped at P3 and P5, where small spatio-temporal adapters (hidden ratio 0.25) adapt the detached features for the decoder; P4 passes a shared gradient scaled by 0.2.

3.4 Inference

Clips of 64 frames start every 32 frames, with a final clip aligned to the last frame. Dense detections (objectness times class probability, threshold 0.005) are weighted by a Hann window over each clip, normalized so that a frame's most central clip keeps weight 1, merged across clips, and suppressed class-wise at IoU 0.5. Except in two variants marked in Table 1, no input is augmented and no models are ensembled. Tube queries are masked at visibility 0.35 and interpolated to frame rate; tubelets visible on fewer than 8 frames are dropped, and the rest are linked across clips by a MOC-style linker (link IoU 0.45, tubelet NMS 0.6, top-k 10, minimum length 16, split gap 2) with tube NMS at 0.3. For video metrics, a training-free *geometry snap* then moves each linked tube box 75% of the way toward the highest-IoU same-class dense detection in its frame when that IoU is at least 0.3, and averages the boxes over ± 2 frames with triangular weights. The snap changes no dense prediction and therefore no frame metric.

4 EXPERIMENTS

4.1 DATASETS AND METRICS

UCF101-24. We use split 1 with the corrected annotations of Singh et al. (2017): 24 classes, 2,284 training and 910 test videos. Video mAP matches predicted and ground-truth tubes by spatio-temporal IoU at 0.2 and 0.5 (AP20, AP50) and averaged over 0.5:0.05:0.95 (AP50:95), computed with MOC's evaluator (Li et al., 2020c). Frame mAP at IoU 0.5 is computed in several ways in the literature (Section 5.1); we report each.

JHMDB-21. JHMDB-21 (Jhuang et al., 2013) contains 928 trimmed videos of at most 40 frames over 21 classes with three official splits. Every frame is annotated, so the frame-set question does not arise. We report each split and, following Li et al. (2020c), the three-split mean.

AVA v2.2. AVA v2.2 (Gu et al., 2018) annotates 80 atomic actions, 60 of which are evaluated, on keyframes at 1 Hz with multiple labels per actor. We report frame mAP at IoU 0.5 with the official ActivityNet evaluator. Because supervision is keyframe-only, the dense loss is masked to the annotated frame. The 235-video official training split is divided into 201 training and 34 development videos; the official 64-video validation set is never used for selection. Our stronger AVA model adds about 138k training keyframes from AVA-Kinetics (Li et al., 2020a).

MultiSports. MultiSports (Li et al., 2021) has 1,574/555 official training/validation videos and 66 classes, 60 of which the official evaluator scores. We report frame mAP and video AP at 0.2 and 0.5 with our re-implementation of the official evaluator. A 168-video development partition is held out of the training videos for selection.

4.2 IMPLEMENTATION AND SELECTION PROTOCOL

Training uses AdamW (Loshchilov & Hutter, 2019) (learning rate 10^{-4}, weight decay 0.05, cosine schedule with one warm-up epoch) at an effective batch of 32, bfloat16 autocast, gradient clipping at 1.0, and an exponential moving average (decay 0.999) whose weights are evaluated. UCF101-24 training augments with flips, scale crops (0.5–1.5), color jitter, grayscale, random erasing, and temporal strides of 1–3. Inputs are 64-frame clips at 224×224 for UCF101-24 (30 epochs) and AVA (6 epochs), 40-frame clips at 224×224 for JHMDB-21 (40 epochs), and 64-frame clips at 288×512 for MultiSports (12 epochs). The model has 414M parameters, of which the frozen YOLO11-L layers and motion trunk hold 63M. The JHMDB-21, AVA, and MultiSports models reuse the frozen motion trunk, which was trained with UCF101-24 labels.

We hold out capture groups g22–g25 of the UCF101-24 training split: g24/g25 is the *tune* partition on which inference settings are chosen, and g22/g23 is a disjoint *confirm* partition. Three settings were selected this way and frozen in a protocol file before the reported test evaluation: the tube linker (adopted when its confirm AP50 exceeded the default by at least 0.5 without lowering AP20 or AP50:95, on two seeds), the frame clip weighting (chosen among four candidates on tune and confirmed against the default on both partitions of a second seed), and the geometry snap (chosen on tune; on confirm it passed the linker's rule with $+1.24$ AP50). Held-out models train without g22–g25; the reported checkpoint is the same recipe trained on the full training split. Four caveats apply (Appendix B). The architecture and training recipe were fixed earlier by comparisons on split-1 test, before this protocol existed. The frozen motion trunk was trained on the full split, including g22–g25, so held-out numbers that involve it are optimistic, markedly so for the motion ablation (Section 5.2); test numbers are unaffected. The reported checkpoint was scored on test before the freeze (91.64 frame mAP with its raw, non-EMA weights and default weighting). And the first test runs filtered tube candidates at 16 visible frames instead of the 8 used in validation; we report test video numbers recomputed with the validated filter and list the uncorrected ones in the appendix. Every number we report uses the frozen settings.

To bound seed noise on test, three seeds of the held-out recipe were scored on test under the frozen protocol (mean $\pm$ standard deviation): frame mAP is 92.98 ± 0.05 on annotated frames and 89.41 ± 0.12 on every frame (inclusive extents), while video mAP varies far more, 88.79 ± 0.51 AP20,

Frame set	AP integral	Published under this convention	YOLO-ST	+flip[h]	Ens.[h]
Every frame	VOC, ROAD MATLAB[a]	BMViT[b] 90.7	88.47	88.93	88.89
Every frame	VOC, ROAD Python[a]	–	89.30	89.70	89.80
Every frame	Trapezoid, raw curve[c]	MOC 78.0	88.92	89.24	89.46
Annotated	VOC every-point[d]	–	93.21	93.50	93.63
Annotated	Trapezoid, raw curve	–	92.99	93.23	93.44
Annotated	Trapezoid, with end point[e]	YOWOFormer-L 93.35	95.10[f]	95.28	95.26
Not stated	–	STAR-L 90.3, HIT[g] 84.8	–	–	–

Video mAP	AP20	AP50	AP50:95
STAR-L (Gritsenko et al., 2024)	89.8	73.4	35.8
HIT[g] (Faure et al., 2023)	88.8	74.3	–
YOLO-ST	89.82	73.39	34.97

Table 1: UCF101-24 split-1 test (%). Top: the reported checkpoint's predictions (YOLO-ST column) scored under each frame-mAP convention, beside the published results that use it. [a]ROAD's two released evaluators (Singh et al., 2017), both scoring every frame: the MATLAB script its README recommends for published results widens detections by one pixel and keeps those above 0.01, at most 20 per class and frame after NMS at 0.45 (89.55 without this filter); the Python script uses exclusive box extents (89.54 with the inclusive extents of the other rows). [b]Ntinou et al. (2024) state that they follow ROAD's protocol but not which evaluator; their repository contains no code. [c]MOC's released frame evaluator (Li et al., 2020c). [d]The VOC routine of YOWO's released evaluator (Köpüklü et al., 2019) on all annotated frames; YOWO's own frame list omits the last frame of each tube, which gives 93.32. [e]The released evaluator of YOWOv3 (Nguyen et al., 2026) and YOWOFormer (Thakaew et al., 2026): exclusive extents on boxes rescaled to 224×224; YOWOv3-L reports 88.64 under it. [f]Our re-implementation of that evaluator from its released source (Appendix A); it also counts ground truth on frames without detections, which the released code drops (95.30 if dropped). [g]HIT uses external person proposals, so its numbers are not end-to-end. [h]Not single-model, single-pass: frame detections fused with those of the horizontally flipped video (+flip, gated on held-out data), or of the reported model and four held-out variants trained without g22–g25 (Ens., fixed before testing; Appendix B). Video mAP is unchanged.

70.92 ± 2.02 AP50, and 33.29 ± 2.12 AP50:95. These seeds train without g22–g25 and bound the run-to-run noise of the recipe rather than replicating the reported checkpoint.

5 Results

5.1 UCF101-24 Frame mAP Depends on the Evaluation Convention

Table 1 scores one set of predictions from the reported checkpoint under each convention. The frame set dominates: with box extents held fixed, scoring every frame rather than only annotated ones costs 3.67 points (93.21 to 89.54). Of the 20,762 test frames that contain no annotated action, 14,728 keep at least one detection, and 4,789 of the detections on these frames score above 0.5; the annotated-frame convention never sees these frames. The AP integral, the box extents, and the detection filter of ROAD's MATLAB script move the number by a further 0.2 to 1.9 points. Compared under the convention each method uses or cites, YOLO-ST exceeds YOWOFormer-L and YOWOv3-L under the YOWO-family evaluator (95.10 against 93.35 and 88.64) and trails BMViT under both of ROAD's all-frame evaluators (88.47 and 89.30 against 90.7). STAR-L does not state its frame set: our checkpoint is above it on annotated frames and below it on every frame. On video mAP, which these works compute by spatio-temporal tube overlap, we measure $+0.02$ AP20, -0.01 AP50 and -0.83 AP50:95 against STAR-L, each within one standard deviation of the three seeds in Section 4.2, so the video comparison is unresolved. Two frame-level variants, fusing each video with its horizontal flip (gated on held-out data) and fusing five models (fixed before testing), add 0.16 to 0.54 points in every row of Table 1. They narrow the gap to BMViT by 0.40 to 0.50 under either ROAD evaluator without closing it, and neither is a single-model, single-pass result.

Change	Metric	Tune	Confirm	Test
Adopted				
Clip weighting: default → frozen	Frame	95.70 → **96.38**	94.09 → **95.45**	–
Geometry snap	AP50	80.21 → **82.05**	77.29 → **78.53**	72.93 → 73.39
Geometry snap	AP50:95	43.20 → **45.76**	41.51 → **43.89**	34.18 → 34.97
Measured and rejected				
Tube-score frame rescoring	Frame	96.37 → 96.39	–	–
Trainable keyframe stream	Frame	95.71 → 95.49	–	–
Trainable keyframe stream, full LR	Frame	95.71 → 62.84	–	–
Backbone: VideoMAE-L → H	Frame	95.71 → 95.27	–	–
Tube-extent frame gating	All-frame	94.46 → 94.50	91.69 → 91.67	–
Action-free training clips, 3×	All-frame	94.46 → 94.16	91.68 → 91.18	–
Segment-end erosion	All-frame	94.46 → 94.47	–	–
Module removed (frame mAP)				
None (held-out control)	Frame	96.38	95.45	92.92
Motion pathway	Frame	92.21	92.32	92.62
Task adapters	Frame	96.45	95.75	–
Cross-clip memory	Frame	96.53	95.44	–
Pyramid adapter	Frame	96.20	95.85	–

Table 2: Ablations of the reported recipe (%), measured on the held-out control (seed 17, trained without g22–g25); Tune is g24/g25, Confirm is g22/g23, and Test is split-1 test. Frame is annotated-frame mAP, All-frame scores every frame. Adopted changes were selected on Tune and verified on Confirm before the reported test evaluation. Each change is compared with the control scored in the same pipeline, which accounts for the 0.01 differences between controls. The trainable keyframe stream uses $0.1\times$ the learning rate unless marked. VideoMAE-Huge makes the whole model $1.8\times$ larger (742M against 414M parameters); its change is within the three-seed Tune standard deviation of 0.61 under default weighting. Tube-extent gating (180 settings), action-free training clips, and segment-end erosion (30 settings) were pre-registered and failed their gates. Module rows are single-seed; the three-seed Confirm standard deviation is 0.07. Because the frozen motion trunk saw Tune and Confirm (Section 4.2), the control and the motion row are also scored on test, which it never saw.

5.2 What the Model Owes to Each Choice

Table 2 separates the changes that survive held-out confirmation from those that do not. The geometry snap improves strict localization, $+1.24$ AP50 and $+2.38$ AP50:95 on confirm and $+0.46$ and $+0.79$ on test, while AP20 moves by at most 0.09, as expected of a box correction. The clip weighting changes inference but not the weights.

On the held-out partitions, removing the motion pathway looks decisive: it costs 4.17 and 3.13 frame mAP. Its frozen trunk, however, was trained on those partitions. On the test set, which the trunk never saw, the same removal costs 0.30 (0.34 when every frame is scored), small though beyond the seed standard deviation of 0.05, and it leaves video mAP within seed noise (89.16 / 72.56 / 34.69 against 88.90 / 72.24 / 34.56). Most of the held-out gap is therefore leakage, a caution for any protocol that selects on data a frozen component was trained on. Removing the task adapters or the pyramid adapter raises Confirm frame mAP by 0.30 and 0.40, several standard deviations, while their Tune changes ($+0.07$, -0.18) are within noise, and removing the memory changes neither; their effect on video mAP is not reported, and the recipe was fixed before these ablations ran. A larger backbone does not help.

We also tested whether the tubes' own temporal extent could suppress the dense detections on action-free frames that cost 3.67 points under the all-frame protocol. A pre-registered rule kept a dense detection only when a linked tube covered its frame, over 180 settings of tube score threshold, temporal dilation, down-weighting, and class dependence. No setting gained more than 0.07 all-frame points on tune, and the selected one gained 0.04 on tune and lost 0.02 on confirm against a required gain of 1.0 (Table 2). The linked tubes cover the same action-free frames; both heads read the same pyramid. Two further pre-registered attempts failed as well. Keeping the action-free clips that training normally drops, repeated three times, lowered every-frame mAP by 0.30 on tune

Split	Frame	AP20	AP50	AP50:95
1	93.49	93.41	92.21	59.45
2	91.97	91.49	90.63	57.56
3	89.10	90.53	88.80	56.97
Mean (ours)	91.52	91.81	90.55	57.99
STAR/CSN-152 (Gritsenko et al., 2024)	**93.5**	**96.3**	**95.4**	–
STAR-L (Gritsenko et al., 2024)	92.1	93.1	92.6	–

Table 3: JHMDB-21 official test splits (%), from the model registered before any test score. Video metrics link dense detections with a linker frozen on a development partition of split 1. STAR/CSN-152 is STAR with a CSN-152 backbone (Tran et al., 2019). STAR does not state whether its results are split 1 or a three-split mean; our mean is below both STAR models on all three shared metrics, and our split-1 frame mAP is 0.01 below STAR/CSN-152.

Model	Training data	mAP@0.5
YOLO-ST, dev	AVA subset	27.74
YOLO-ST, *official val*	AVA subset	28.66
YOLO-ST, dev	+ AVA-Kinetics	32.37
YOLO-ST, *official val*	+ AVA-Kinetics	33.23
YOWOFormer-L (Thakaew et al., 2026)	AVA	26.32
STAR-L (Gritsenko et al., 2024)	AVA-Kinetics	42.5
VideoMAE V2 (Wang et al., 2023)	AVA	**42.6**

Table 4: AVA v2.2 frame mAP at IoU 0.5 (%), six training epochs. Rows marked *official val* use the official validation set (64 videos) with the official evaluator and are directly comparable to the published results; *dev* rows use 34 held-out training videos and are not. AVA subset rows train on the 201-video partition; all our models also use the frozen motion trunk trained on UCF101-24. Against YOWOFormer-L we gain 2.34 points on AVA data alone and 6.91 with AVA-Kinetics. Per-class results are in Appendix C.

and 0.50 on confirm, and annotated-frame mAP by 0.61 and 0.78. Eroding the temporal ends of each detected segment gained at most 0.01 on tune. The confident detections on action-free frames lie close to annotated actions (88% within 32 frames on tune), where frames are already trained as negatives and where detected segments do not systematically overrun the annotation.

5.3 Other Benchmarks

On JHMDB-21 (Table 3), AVA v2.2 (Table 4), and MultiSports (Table 5) we follow each official protocol, and substantial gaps to the strongest published results remain. On AVA the model exceeds

Model	Frame	AP20	AP50
YOLO-ST 224^2, 12 epochs, dev	35.53	38.34	16.82
YOLO-ST 224^2, 24 epochs, dev	40.82	44.38	18.85
YOLO-ST 288×512, 6 epochs, dev	37.42	41.56	16.53
YOLO-ST 288×512, 12 epochs, dev	44.45	49.02	20.12
YOLO-ST 224^2, 24 epochs, *official val*	37.23	41.91	15.70
YOLO-ST 288×512, 12 epochs, *official val*	42.11	46.66	18.29
STAR-L (Gritsenko et al., 2024)	**59.3**	**62.0**	**36.2**
STAR/CSN-152 (Gritsenko et al., 2024)	45.4	50.1	18.6

Table 5: MultiSports (%), our re-implementation of the official evaluator (60 of the 66 classes scored). Rows marked *official val* use the official validation split (555 videos), as the published results do; *dev* rows use 168 held-out training videos, and their video APs are the best of eight linker settings on dev. Both models evaluated on official val are shown; the 288×512 model was chosen on dev before its official-val run, and its linker was pinned on dev. Video APs link dense detections.

YOWOFormer-L under the same evaluator. Video APs on JHMDB-21 and MultiSports come from linking dense detections, which outperformed the tube queries on the development splits of both benchmarks.

6 Limitations and Conclusion

The largest limitation is temporal. YOLO-ST fires on frames outside the annotated action intervals; these frames are invisible to the annotated-frame convention and cost 3.67 frame mAP when every frame is scored. Neither gating by the tubes' own temporal extent, nor training on action-free clips, nor eroding detected segments removes them, because they sit within a few frames of annotated actions; sharper temporal boundaries are the natural next step. Second, the frozen motion trunk was trained on the held-out groups, which inflated the held-out value of the motion pathway about tenfold, and the recipe was fixed on split-1 test before our protocol existed; a clean protocol needs held-out groups that no pretrained component has seen. Third, UCF101-24 and MultiSports are sports-dominated and JHMDB-21 is small; on AVA v2.2 and MultiSports substantial gaps remain, which we do not attribute to a single cause.

We introduced YOLO-ST, a dense action detector that separates semantic ownership, high-rate motion, and frame geometry within one model and matches STAR-L on video AP20 and AP50. Its evaluation shows that UCF101-24 frame-mAP comparisons are not meaningful without the convention behind each number, which we report for every result.

AI Use Statement

In this work, we used generative AI tools to implement methods, to provide feedback on the design of experiments, to clean and reformat dataset files, and to interpret results. We have not used generative AI tools to generate synthetic datasets, to develop theoretical models or conceptual frameworks, to formulate mathematical claims, to provide ingredients for proving mathematical claims, to assist in the writing of proofs, or to propose or refine the hypotheses of this work; the remaining required-disclosure tasks, namely translation and qualitative or thematic data analysis, are not applicable to this work.

We have reviewed all AI-assisted work. The research questions, the method design, and the architectural choices are the authors'. AI-generated code was reviewed by the authors before it produced any reported result. Every reported metric was emitted by that training and evaluation code, and the summary metric of every reported model is written to an append-only ledger together with its evaluation partition and evaluator; no number in this paper was generated, estimated, or altered by a generative AI tool, and the inference settings were frozen before the reported evaluations. Competitor figures in the comparison tables were checked against the papers they are attributed to, and figures that could not be verified under our evaluation protocol are marked as such.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility Statement

Every summary metric of the models in this paper is recorded in an append-only experiment ledger together with its evaluation partition and exact evaluator. Inference settings are selected on held-out capture groups g24/g25 and confirmed on g22/g23; the linker, clip weighting, and geometry snap were frozen in a protocol file whose timestamps precede the reported test evaluations, and Appendix B records every earlier test evaluation and selection decision. Code, configurations, the protocol file, and the ledger are available at `https://anonymous.4open.science/r/YOLO-ST-F1BB` and in the supplementary material.

Ethics Statement

This work uses four public action-detection datasets (UCF101-24, JHMDB-21, MultiSports, and AVA v2.2) and the AVA-Kinetics annotations, together with models pretrained on Kinetics-400 and COCO; all contain images or video of people and were released for research use. We did not collect

new data and did not attempt to identify any individual. Action detectors of this kind can be applied to surveillance, and the datasets used here are sports- and film-oriented benchmarks rather than surveillance corpora; we make no claim that the method is appropriate for deployment in settings that affect individual rights without separate ethical and legal review.

REFERENCES

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pp. 213–229, 2020.

Lei Chen, Zhan Tong, Yibing Song, Gangshan Wu, and Limin Wang. Efficient video action detection with token dropout and context refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10388–10399, 2023.

Shoufa Chen, Peize Sun, Enze Xie, Chongjian Ge, Jiannan Wu, Lan Ma, Jiajun Shen, and Ping Luo. Watch only once: An end-to-end video action detection framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8178–8187, 2021.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6824–6835, 2021.

Gueter Josmy Faure, Min-Hung Chen, and Shang-Hong Lai. Holistic interaction transformer network for action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3340–3350, 2023.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, 2019.

Alexey A. Gritsenko, Xuehan Xiong, Josip Djolonga, Mostafa Dehghani, Chen Sun, Mario Lučić, Cordelia Schmid, and Anurag Arnab. End-to-end spatio-temporal action localisation with video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18373–18383, 2024.

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6047–6056, 2018.

Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3192–3199, 2013.

Glenn Jocher and Jing Qiu. Ultralytics YOLO11. `https://github.com/ultralytics/ultralytics`, 2024.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified CNN architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019.

Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The AVA-Kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020a.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21002–21012, 2020b.

Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *Proceedings of the European Conference on Computer Vision*, pp. 68–84, 2020c.

Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. MultiSports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13536–13545, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pp. 740–755, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Dang Duc Manh Nguyen, Bui Duc Nhan, Jia-Ching Wang, and Viet-Hang Duong. YOWOv3: An efficient and generalized framework for spatiotemporal action detection. *IEEE Intelligent Systems*, 41(1):75–85, 2026. doi: 10.1109/MIS.2025.3581100.

Ioanna Ntinou, Enrique Sanchez, and Georgios Tzimiropoulos. Multiscale vision transformers meet bipartite matching for efficient single-stage action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18827–18836, 2024.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2019.

Gurkirt Singh, Suman Saha, Michael Sapienza, Philip H. S. Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3637–3646, 2017.

Gurkirt Singh, Vasileios Choutas, Suman Saha, Fisher Yu, and Luc Van Gool. Spatio-temporal action detection under large motion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6009–6018, 2023.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Rui Su, Wanli Ouyang, Luping Zhou, and Dong Xu. Improving action localization by progressive cross-stream cooperation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12016–12025, 2019.

Vasu Thakaew, Sitapa Watcharapinchai, and Punnarai Siricharoen. YOWOFormer: Bridging video transformers and CNN detectors for one-stage spatio-temporal action detection. *IEEE Access*, 14: 85709–85723, 2026. doi: 10.1109/ACCESS.2026.3700732.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5552–5561, 2019.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14549–14560, 2023.

Jianhua Yang and Kun Dai. YOWOv2: A stronger yet efficient multi-level detection framework for real-time spatio-temporal action detection. *arXiv preprint arXiv:2302.06848*, 2023.

Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, Ivan Marsic, Cees G. M. Snoek, and Joseph Tighe. TubeR: Tubelet transformer for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13598–13607, 2022.

A UCF101-24 Frame-mAP Conventions

The YOLO-ST column of Table 1 scores the same recorded detections of the reported checkpoint: 205,532 per-frame boxes over the 910 test videos, after clip fusion and class-wise NMS. The conventions differ in three ways.

- **Frame set.** ROAD's two released evaluators (Singh et al., 2017) and MOC's frame evaluator (Li et al., 2020c) score every frame of every test video, 159,289 frames, so a detection on a frame without a ground-truth box of its class is a false positive. YOWO's evaluator (Köpüklü et al., 2019) and its successors score only annotated frames, 138,527 in the full set; YOWO's released frame list has 137,557 because it omits the last frame of each tube.
- **AP integral.** ROAD and YOWO use VOC every-point interpolation up to the highest recall reached. MOC applies the trapezoid rule to the raw precision–recall points. The YOWOv3 and YOWOFormer evaluators apply the trapezoid rule to the monotonised curve after appending a point at recall 1 with precision 0, which credits area beyond the highest recall reached.
- **Box overlap and matching.** YOWO and MOC match at native resolution with inclusive pixel extents. ROAD's Python script uses exclusive extents; its MATLAB script widens detections by one pixel, which for our boxes matches the inclusive convention, after keeping only detections above 0.01 and at most 20 per class and frame after NMS at 0.45. The YOWOv3 and YOWOFormer evaluators also use exclusive extents, on boxes rescaled to 224×224; the rescaling itself does not change IoU, which is invariant to per-axis scaling of axis-aligned boxes, so the difference is the extent convention. They also assign matches in order of IoU rather than of detection score, and they drop the ground truth of frames on which the detector outputs nothing. All numbers here use the same detections, thresholded at 0.005, which is also the threshold of those pipelines.

On our predictions the frame set accounts for 3.67 points (93.21 to 89.54 with VOC integration and inclusive extents), the exclusive extents of ROAD's Python script for a further 0.23 (89.30; ROAD reads its own copy of the annotations, so a one-pixel difference in box convention could return part of this), the detection filter of ROAD's MATLAB script for 1.08 (89.55 to 88.47, porting the script and mapping our annotation copy, 0-based with inclusive right edges, to its format), the AP integral for 0.22 on annotated frames and 0.62 on every frame, and the YOWO-family evaluator adds 1.89 on annotated frames. Two limits apply to the 95.10. Pretrained weights for YOWOFormer-L were listed as forthcoming in its repository when we checked, so we could not score its checkpoint under the other conventions. And because its evaluator is fused with its inference loop, we re-implemented it from the released source rather than executing it, on a frame list we built from the corrected annotations. That list is the full annotated-frame set, whereas the released pipelines read a separately distributed test list, probably YOWO's 137,557 frames; unlike the released code, the re-implementation keeps the 145 ground-truth boxes on the 139 annotated frames where we output nothing, which lowers the score from 95.30 to 95.10. The 95.10 should be read as the effect of the evaluator on our predictions rather than as a head-to-head result. STAR (Gritsenko et al., 2024) states the corrected annotations and per-frame evaluation, but not the IoU threshold it uses on UCF101-24, its frame set or its integral, and its code is not released. BMViT (Ntinou et al., 2024) states that it follows ROAD's protocol, which places it with ROAD's every-frame evaluators without saying which one, and its repository contains no code.

B SELECTION AND TEST-EVALUATION RECORD

This section lists every decision that touched a reported partition.

Recipe. Before the held-out partitions existed, the recipe of the reported model was fixed by comparisons on UCF101-24 split-1 test: VideoMAE-L over VideoMAE-B, a distribution box head, 224×224 inputs, eight fine-tuned blocks at $0.1\times$ learning rate, exponential moving averaging, and the tube-query decoder, which was promoted when its test AP50 (72.63) cleared the 71.22 of the previous model.

Test evaluations of the reported checkpoint. It was scored on test with its raw weights, default frame weighting and the previous model's linker (91.64 frame; 89.70 / 72.63 / 34.14 video) before any setting was frozen; with its EMA weights and the frozen clip weighting and linker (93.21 frame); with the frozen snap, together with three ablations of it (no box blending, no smoothing, neither); once more after the candidate-filter correction below, with the frozen snap and with the snap disabled; and a last time in a re-run that reproduced every reported number exactly and supplies the predictions of the +flip and Ens. columns. We report only frozen settings. The linker, clip weighting, and snap were frozen at 05:02 and 13:11 UTC on 14 September and 13:00 UTC on 16 September 2026, each before the corresponding reported test score. Before the clip weighting was selected, all four candidates were scored on three test videos; the candidate favoured there was later rejected on the held-out partitions. The frozen weighting was chosen on tune and beats the default on confirm, but it ties the unweighted merge there (95.45 against 95.46 for seed 17, 95.38 against 95.39 for seed 29), which scored lower on tune.

Candidate filter. The held-out tube caches on which the linker and snap were validated keep query tubelets with at least 8 visible frames; the first test caches of the reported checkpoint and of the three seeds applied the linker's minimum length of 16 at this stage as well. All reported test video numbers come from caches rebuilt with the validated filter of 8, with the linker's minimum length unchanged, in one re-run of each model scored with the frozen snap and with the snap disabled. Before this correction the reported checkpoint scored 89.48 / 73.17 / 34.98 with the snap, and the seeds 88.48 ± 0.57 / 70.03 ± 2.19 / 31.90 ± 2.30 without it.

Fixes for action-free detections. Tube-extent gating (180 settings), training on action-free clips, and segment-end erosion (30 settings) were each registered with their settings, selection rule, and promotion gate (confirm every-frame mAP $+1.0$) before any gated number was computed. All three were rejected on the held-out partitions and never applied to test; the full-data model trained alongside the action-free-clip variant was stopped before completion and never evaluated. Three boundary-negative loss weightings (best confirm every-frame $+0.45$) and three variants of the model's temporal processing (every-frame on tune 94.21–94.37 against the control's 94.46) were rejected the same way; a fourth variant was not completed and has no result.

Test-time variants. Each gated variant was registered before any of its held-out numbers existed. Averaging the weights of the last two EMA snapshots (epochs 24 and 29) raised confirm annotated-frame mAP by only $+0.04$ to $+0.09$ and failed its $+0.10$ gate. Fusing each video's frame detections and tubes with those of its horizontal flip raised frame mAP ($+0.36$ and $+0.40$ annotated on confirm) but lost up to 2.4 video AP50 on tune; keeping the extent of the strongest tube when fusing also failed its check on a fresh model, so video mAP is never fused. Fusing frame detections only, decided on two models whose flipped evaluations did not yet exist ($+0.38$ and $+0.47$ annotated on confirm), passed and was applied to test once (Table 1, +flip). The five-model ensemble (Ens.) had no gate: its members were fixed before any test run after an exploratory held-out check, in which fusing three of them raised annotated-frame mAP by $+0.37$ over their mean on both tune and confirm. They are the reported model and four held-out models (the control and three one-variable variants), whose registrations had excluded test use; this use is recorded in the ledger.

Held-out leakage. The frozen motion trunk of every held-out model was trained on all of split-1 train, including g22–g25, so held-out numbers that involve it are optimistic. Removing the motion pathway costs 4.17 and 3.13 frame mAP on tune and confirm but 0.30 on test, which is disjoint from

the trunk's training data. The motion ablation was scored on test once, with the corrected candidate filter.

JHMDB-21. A 288×288 variant was trained after the reported 224×224 model had been scored on the test splits; it scores 91.34 frame and 92.04 / 90.55 / 58.80 video on the three-split mean, and we report the model registered first.

MultiSports. The 288×512, 12-epoch model was registered before any official-validation run. The 224^2, 24-epoch model was scored on official validation first; the 288×512 model then outscored it on the development partition and was scored on official validation second. Both official-validation results are reported in Table 5.

C AVA v2.2 Per-Class Behaviour on the Official Validation Set

Overall mAP@0.5 on the official validation set is 33.23 with AVA-Kinetics training. Per-class AP is far from uniform.

Strongest class	AP	Weakest class	AP
stand	88.7	put down	3.3
sit	87.7	push (another person)	3.2
walk	84.0	hand wave	3.0
talk to (e.g., self, a person, a group)	82.7	lift/pick up	2.5
answer phone	79.5	listen (e.g., to music)	2.4
swim	78.0	hand shake	2.3
play musical instrument	74.1	pull (an object)	2.0
watch (a person)	73.9	point to (an object)	0.8

Table 6: The eight strongest and eight weakest of the 60 evaluated AVA classes, from the same checkpoint and evaluation that produces the official validation mAP in Table 4. Of the 60 classes, 12 score below 5 AP and 1 below 1 AP, while 16 exceed 50 AP; the median is 30.1.

Ten classes exceed 70 AP, led by the frequent postural and locomotion classes, and twelve fall below 5 AP, mostly rare fine-grained interactions. Training data moves individual classes sharply: swim has 106 training instances in our AVA training videos and scores 28.2 without AVA-Kinetics, which adds 2,763 more, and 78.0 with it, the largest per-class gain (followed by climb, $+27.0$, and jump/leap, $+20.8$). Backbone pretraining, input resolution, training-data scale, and our six-epoch schedule differ from STAR and VideoMAE V2 (YOWOFormer-L shares our VideoMAE-L backbone and input size), and we do not claim to separate them here.