Size Doesn't Matter: Motion Based Guidance For Video Understanding

Stanford CS229 Project

Stanford University
Department of Computer Science
Team Members: Ayaan Malik, Miko Rimer, Yanav Lall
Emails: ayaan04@stanford.edu, mikor@stanford.edu, yanav@stanford.edu
Category: Computer Vision

1 Introduction

Video models such as X-CLIP (Ni et al., 2022) are specialized in action recognition but are primarily driven by appearance-centric signals (objects, background, coarse pose). In contrast, humans can often recognize actions from skeletal motion alone. We ask whether this human-interpretable signal, joint motion, can act as a compact teacher to improve a larger video model like X-CLIP.

We encode each frame as a 51-D motion vector: for each of the 17 COCO joints (Ji et al., 2022), we compute 2-D velocities and their speed. A small ($\approx$1M-parameter) motion-only network is trained on UCF-101 to predict clip-level action distributions, and its soft labels supervise an X-CLIP student alongside ground-truth labels. We also compare against a more expensive baseline where an X-CLIP teacher supervises an X-CLIP student.

In this work, we demonstrate that a compact, human-interpretable motion network can act as an effective teacher for a large video model.

2 Related Work

Recently, there has been an interest in moving beyond pose as an intermediate representation. Zhao et al. explore the use of pose as a weakly supervised signal for action segmentation, improving perception tasks by treating pose as a structured weak prior (Zhao et al., 2025). However, Zhao et al. focuses on pose itself rather than on motion. Honing in on motion enables us to learn an interpretable emergent property of pose that captures how joints evolve over time. As such, we draw inspiration from teacher–student (TS) motion-derived expert to supervise larger vision models.

Specialist models trained rapidly alongside larger backbones have shown value in knowledge distillation beyond one-hot labels (Hinton et al., 2015). Huang et al. build on this idea with MoCLIP-Lite (Huang et al., 2025), a lightweight supervised motion-vector network that supplies motion-based heuristics to a CLIP-style video recognizer. MoCLIP-Lite supports the trend that cheap motion cues can significantly improve CLIP-based video recognition.

Beyond Huang et al., we constrain ourselves to a human-interpretable pseudo-label pipeline in which joint velocities and speeds are explicitly modeled as the core supervisory signal for a small motion expert. Rather than fusing motion features at the representation level, we use pose-derived motion to train an explicit motion teacher and distill its predictions into a larger X-CLIP student, aligning the learning signal with human-understandable notions of action and motion.

3 Dataset and Features

UCF-101 Dataset: In order to assess the accuracy of our TS models in recognizing everyday actions, we evaluate on UCF-101 (Soomro et al., 2012). UCF-101 (Soomro et al., 2012) contains 13,320 videos from 101 action classes at 320 × 240 and 25 FPS. We use the official split (60/10/30% train/val/test) and apply our weak supervision pipeline to all train and val clips.
Keypoint feature extraction: In order to extract the 2-D keypoints, we first use RT-DETR on full frames to obtain person bounding boxes before running the pose estimation (Lv et al., 2024) (Zhao et al., 2023). Given each detected person's bounding box, we apply a lightweight pose extractor being

ViTPose as our keypoints processor (Xu et al., 2022). This output serves as the basis for computing the speed features which informs our motion pseudo label teacher.

4 Methods

Our pipeline consists of three stages: (i) converting each video into a body-part motion vector; (ii) training the teacher models; and (iii) training a student model supervised by the teacher model. All methods rely on the X-CLIP base model with patch-size 16 as a frozen feature extractor on 32 uniformly sampled frames per video (25 FPS) as a student model (Ni et al., 2022). X-CLIP is video–text model with a ViT-B/16 backbone that applies contrastive learning to videos (Ni et al., 2022).

4.1 Deriving Motion Vectors From 2-D Poses

4.1.1 Pose Extraction

For each frame we first run RT-DETR to obtain person bounding boxes (Zhao et al., 2023) (Lv et al., 2024), then apply ViTPose within each box to obtain 17 COCO joints (Xu et al., 2022). Let

$$K_n^{(p)} \in \mathbb{R}^{17\times 2}, \quad C_n^{(p)} \in \mathbb{R}^{17}$$

denote the keypoint locations and confidences for person p in frame n. We keep only the person $p^\star$:

$$p^\star = \arg\max_p \frac{1}{17}\sum_{j=1}^{17} C_{n,j}^{(p)}; \quad K_n := K_n^{(p^\star)}; \quad C_n := C_n^{(p^\star)}$$

4.1.2 Confidence Gating

Following Tang et al. (Tang et al., 2022), we normalize 2D skeletons before computing motion. We apply three simple gates: (i) *multi-person*: select $p^\star$ as above; (ii) *per-joint masking*: joints with $C_{n,j} < 0.8$ are marked invisible; (iii) *minimal visibility*: frames with fewer than 9 visible joints are discarded. Note, discarding is equivalent to setting rows to NaN.

4.1.3 Translation and Scale Invariance

We then make each valid frame translation- and scale-invariant (Tang et al., 2022). For translation, we subtract the origin point o_n for each joint in each frame. We denote o_n as the midpoint of the shoulders and normalize:

$$K'_n(j,:) = K_n(j,:) - o_n.$$

For scale, we compute a torso distance s_t by computing the distance between the shoulders (Tang et al., 2022).

$$K'_n(j,:) := K'_n(j,:)/s_t \quad \text{Frames with } s_t < \epsilon = 8\text{px are discarded}$$

4.1.4 Extracting Motion Vectors

Let $K'_{n,j}$ denote the normalized 2-D COCO joint (Ji et al., 2022) location for joint j in frame n after applying the joint-confidence gating. For each available joint we compute velocities via a centered frame difference, giving $v_j \in \mathbb{R}^2$ for both (x, y) across all frames that were not discarded. From the velocities, we derive the speed $s_j \in \mathbb{R}$ for each body part over frames $n = 2, \ldots, N-1$. Ultimately, this yields an aggregated motion vector $x_n \in \mathbb{R}^{51}$ per frame which forms the basis of our motion derived pseudo labeler.

4.2 Training Teacher Models

To evaluate motion as an interpretable supervisory signal, we compare two teacher–student setups that differ in both teacher size and signal. We deliberately *do not* match the number of ground-truth labels seen by each teacher; instead, our goal is to test whether a compact motion teacher can substitute for a larger X-CLIP-based teacher for distillation. The motion teacher is trained on pose-derived motion features, while the X-CLIP teacher is trained in a controlled k-shot setting on frozen X-CLIP embeddings, and both are then used to produce soft labels for the student training stage (Section 4.3).

4.2.1 Motion-Derived Teacher

We model each video as a bag $B_i = \{x_{it}\}_{t=1}^{T_i}$ of per-frame motion descriptors $x_{it} \in \mathbb{R}^{51}$ with a single clip-level label $y_i \in \{1, \ldots, C\}$, where i is the bag index and t is the frame. The motion teacher learns a clip-level distribution

$$f(B_i) = p_\theta(y \mid B_i) \in \Delta^{C-1}$$

in a multiple instance learning (MIL) setting, where only clip-level labels are available and the model must learn which frames are informative (Wang et al., 2017).

Instance encoding: Each instance encodes per-frame motion statistics (joint velocities, speeds, visibilities). Each motion vector $x_{it} \in \mathbb{R}^{51}$ is projected to a hidden space of dimension $H{=}256$ using a linear layer, layer norm, ReLU, and dropout. We then process the per-frame motion embeddings with an L-layer Transformer encoder with sinusoidal positional encodings (Vaswani et al., 2023).

Gated attention MIL pooling: To obtain a single bag representation, our MIL pooling follows the gated attention operator (Ilse et al., 2018) over the instance embeddings $\{h_{it}\}$:

$$v_{it} = \tanh(V h_{it}), \quad u_{it} = \sigma(U h_{it}), \quad g_{it} = v_{it} \odot u_{it},$$

$$a_{it} = w^\top g_{it}, \quad \alpha_{it} = \frac{\exp(a_{it})}{\sum_{t' : m_{it'}=1} \exp(a_{it'})}, \quad r_i = \sum_{t=1}^{T_i} \alpha_{it}\, h_{it} \in \mathbb{R}^H,$$

where m_{it} masks padded time steps. Gated attention ensures that the motion model focuses on a subset of discriminative frames while down-weighting neutral or ambiguous motion frames.

Classifier, training, and soft labels: We feed r_i to a small MLP classifier:

$$h_i^{\text{cls}} = \text{Dropout}(\text{ReLU}(W_1 r_i + b_1)), \qquad z_i = W_2 h_i^{\text{cls}} + b_2,$$

and obtain clip-level probabilities $p_\theta(y \mid B_i) = \text{softmax}(z_i)$. The motion teacher is trained with cross-entropy and label smoothing (with $\varepsilon = 0.1$) (Szegedy et al., 2015). After training, we use $p_\theta(y \mid B_i)$ as soft labels for all clips; these motion-based soft labels serve as an interpretable teacher signal for the student model and let us compare a compact pose-motion teacher to a larger X-CLIP-based teacher in distillation.

4.2.2 X-CLIP Derived Teacher

We follow the X-CLIP training protocol (Ni et al., 2022). For training, we apply standard video augmentations: resize (short side 256), random 224 × 224 crop, horizontal flip (probability 0.5), color jitter, occasional grayscale, and ImageNet statistics normalization. For validation and test, we use resize (short side 256), center crop, and the same normalization.

To train this teacher, we define a K-shot supervised subset $\mathcal{L}$ containing $N_L = K \times 101$ labeled videos with K clips per UCF-101 class. We train a linear multinomial logistic regression classifier (Bishop, 2006) with weight matrix $\mathbf{W} \in \mathbb{R}^{101 \times 512}$ and bias $\mathbf{b} \in \mathbb{R}^{101}$:

$$\min_{\mathbf{W},\mathbf{b}} \left[\frac{1}{N_L} \sum_{i=1}^{N_L} \ell_{\text{CE}}(\mathbf{W}\,\phi(x_i) + \mathbf{b},\, y_i) + \frac{1}{2}\,\|\mathbf{W}\|_F^2 \right], \qquad \ell_{\text{CE}}(\mathbf{z}, y) = -\log\left(\frac{e^{z_y}}{\sum_k e^{z_k}}\right).$$

We optimize this objective with L-BFGS and apply early stopping on validation accuracy, requiring an improvement of at least $\delta = 10^{-3}$ over 20 iterations.

4.3 Training Students on Teacher Pseudo Labels

To incorporate weak motion-based supervision without degrading the primary classifier, we use a multi-head architecture on top of the frozen X-CLIP backbone. For each video x_i, we extract a 512-D embedding $\phi(x_i)$ and attach two linear heads: a supervised head trained on the sparsely labeled set $\mathcal{L}$, and a weak head trained on the densely pseudo-labeled set $\mathcal{L}_{\text{weak}}$ (Hinton et al., 2015). The heads produce

$$q_i^{\text{sup}} = \text{softmax}(\mathbf{W}_{\text{sup}}\phi(x_i) + \mathbf{b}_{\text{sup}}), \qquad q_i^{\text{weak}} = \text{softmax}(\mathbf{W}_{\text{weak}}\phi(x_i) + \mathbf{b}_{\text{weak}}).$$

The supervised head is trained with cross-entropy on ground-truth labels y_i for $i \in \mathcal{L}$, while the weak head matches the teacher distribution $\pi_i \in \Delta^{C-1}$ (from either the motion MIL teacher or the X-CLIP teacher) (Hinton et al., 2015). We use

$$\mathcal{L}_{\text{sup}} = -\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} \log q^{\text{sup}}_{i,y_i}, \qquad \mathcal{L}_{\text{weak}} = -\frac{1}{|\mathcal{L}_{\text{weak}}|}\sum_{i\in\mathcal{L}_{\text{weak}}}\sum_{c=1}^{C} \pi_{i,c} \log q^{\text{weak}}_{i,c},$$

and minimize the weighted sum $\mathcal{L} = \alpha\mathcal{L}_{\text{sup}} + \beta\mathcal{L}_{\text{weak}}$, where α and β control the relative contribution of ground-truth and teacher supervision (Zhao et al., 2025).

4.4 Evaluation Bounds

Lower Bound: Zero-Shot Classification: We use X-CLIP in a zero-shot setting as a lower bound. For each class, we define $M = 5$ text prompts (e.g "a video of a person [action]") and obtain 512-D ℓ_2-normalized text embeddings $\mathbf{t}_{c,m}$ via X-CLIP's text encoder. Class prototypes ψ and predictions $\hat{y}$ are computed as

$$\psi(c) = \text{norm}\left(\frac{1}{M}\sum_{m} \mathbf{t}_{c,m}\right), \quad \hat{y} = \arg\max_{c} \phi(x)^{\top}\psi(c).$$

Upper Bound: Full-Supervised Linear Probe: We train a logistic regression classifier on all labeled training videos and apply the augmentations above on frozen X-CLIP features. This approximates an upper bound for linear probes on frozen X-CLIP features, since it uses all available labels.

5 Experiments / Results / Discussion

We evaluate our motion-based teacher student (TS) approach on UCF-101 using the official train/val/test splits. Our comparisons in Table 1 include: (i) zero-shot baselines (Random, CLIP, X-CLIP); (ii) supervised X-CLIP linear probes with $k \in \{2, 8\}$; (iii) our motion-only teacher alone; (iv) our Motion TS setup where the motion teacher provides soft labels to an X-CLIP student; and (v) an X-CLIP TS baseline where an X-CLIP-derived teacher supervises the same X-CLIP student. We also report MoCLIP-Lite and fully supervised X-CLIP results from prior work as reference upper bounds.

5.1 Results

	UCF-101 Top-1 Accuracy (%)	
	Method	**Top-1**
Zero-shot	Random Guessing	0.99
	CLIP	64.5
	X-CLIP	72.0
Few-shot	X-CLIP ($k = 2$)	86.73
	X-CLIP ($k = 8$)	94.3
Motion TS (ours)	Raw Motion	79.35
	Motion (teacher) + X-CLIP ($k = 2$)	**91.12**
	Motion (teacher) + X-CLIP ($k = 8$)	**94.7**
X-CLIP TS (ours)	X-CLIP ($k = 2$ teacher) + X-CLIP	**95.72**
	X-CLIP ($k = 8$ teacher) + X-CLIP	**96.2**
Upper bound	MoCLIP-Lite	89.2*
	Fully supervised X-CLIP	98.2**

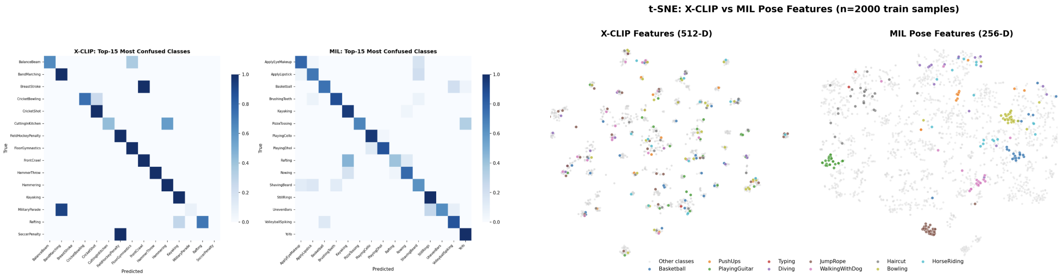
Table 1: Top-1 accuracy on UCF-101. Teacher–student (TS) variants outperform the MoCLIP-Lite motion baseline at a similar parameter scale. Params (M): ours ≈1, X-CLIP ≈200, MoCLIP-Lite ≈0.97. * (Huang et al., 2025); ** (Yang et al., 2024).

Zero-shot performance sets a relatively strong lower baseline (CLIP: 64.5, X-CLIP: 72.0), but it does not approach the high-accuracy regime. In the low-parameter setting, our Motion TS models ($\sim$1M trainable parameters in the motion teacher) outperform MoCLIP-Lite ($\sim$0.97M), reaching 91.1 (at k=2) and 94.7 (at k=8) versus the 89.2 reported by Huang et al. (2025). This comparison is not fully apples-to-apples: our student is a stronger video backbone (X-CLIP) and is allowed k labeled examples per class. MoCLIP-Lite reinforces the notion that cheap motion cues can boost large models; our additional gains suggest that turning pose-derived motion into an explicit teacher for a video–language student, rather than just an extra feature stream, can further improve accuracy while keeping the added learnable component compact.

Compared to the motion teacher, the X-CLIP teacher uses a much larger backbone and achieves higher top-1 accuracy (up to 96.2% vs. 94.7%). However, it is considerably more expensive to train. Our results thus position the motion teacher as a practical alternative when computational or memory budgets preclude training a full X-CLIP teacher.

5.2 Visualization Experiments

We ran two visualization experiments to compare X-CLIP and our motion-only MIL teacher: (A) class-level error structure and (B) feature geometry (Fig. 1).



(a) Top-15 class confusions.

(b) t-SNE feature geometry.

Figure 1: Visualization experiments for X-CLIP and the motion MIL teacher. (A) Row-normalized confusion matrices for the top-15 most confused classes. (B) t-SNE of 2,000 training clips in the X-CLIP (512-D) and motion MIL (256-D) feature spaces.

A. Error structure. X-CLIP confuses actions that share scene appearance or objects (e.g `BandMarching`/`MilitaryParade`) indicating a strong reliance on background rather than fine-grained motion. In contrast, our motion teacher confuses actions with similar joint dynamics (e.g. `ApplyEyeMakeup`/`ApplyLipstick`) despite visual context differences. Both models struggle on settings (e.g. `Kayaking`/`Rafting`), where 2D pose is noisy and motion patterns are semantically similar. These confusion patterns support our claim that the compact motion teacher captures information complementary to X-CLIP and improves the student when used for distillation.

B. Feature geometry. The t-SNE plots show broad, minimally overlapping X-CLIP clusters, consistent with higher top-1 accuracy Table 1). The motion space gives tighter clusters for motion-driven classes (`JumpRope` and `PushUps`), but has a weaker separation for fine-motor activities (`Typing` and `Haircut`). Together these visualizations support our claim that the motion teacher provides complementary supervision to X-CLIP, but does not cluster features in the same way.

6 Conclusion / Future Work

Ultimately, we see that a compact, human-interpretable motion teacher can improve the performance of a large video model on action recognition. Using a pose-derived joint-velocity and speed vector, we train a $\sim$1M-parameter MIL-based motion network that provides soft pseudo-labels to an X-CLIP student. On UCF-101, this Motion TS setup outperforms prior motion-based baselines such as MoCLIP-Lite in a comparable parameter regime and significantly improves few-shot X-CLIP when k is small.

Analyses of confusion matrices and feature geometries suggest that the motion teacher captures a complementary, action-centric representation focused on joint dynamics, while X-CLIP remains more appearance-driven. This complementarity highlights motion-based teachers as a practical, interpretable source of supervision that can deliver competitive few-shot gains with far fewer additional parameters than a full-scale video teacher. In future work, we plan to extend our framework to 3D meshes and more challenging action datasets. In addition, we can explore video segment level classification rather than clip-level labeling.

7 Contributions

- **Ayaan Malik**: Designed our motion model, and set up our evaluation pipeline. Ayaan contributed to the writing of sections 4.2, 4.4, 5.2 & 6
- **Miko Rimer**: Ran data pre-processing: RT-DETR + ViTPose preprocessing and 2D skeleton normalization. Miko contributed to the writing of sections 2, 3, 4.1, 5.1 & 6
- **Yanav Lall**: Implemented the X-CLIP as a teacher and weak supervision pipeline. Yanav contributed to the writing of sections 1, 4.2, 4.3, 4.4, 5 & 6

References

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, USA.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Binhua Huang, Ni Wang, Arjun Pakrashi, and Soumyabrata Dev. 2025. Moclip-lite: Efficient video recognition by fusing clip with motion vectors.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning.

Zhangjian Ji, Zilong Wang, Ming Zhang, Yapeng Chen, and Yuhua Qian. 2022. 2d human pose estimation with explicit anatomical keypoints structure constraints.

Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. 2024. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *European conference on computer vision*, pages 1–18. Springer.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision.

Wei Tang, Peter MA van Ooijen, Deborah A Sival, and Natasha M Maurits. 2022. 2d gait skeleton data normalization for quantitative assessment of movement disorders from freehand single camera video recordings. *Sensors*, 22(11):4245.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. 2017. Untrimmednets for weakly supervised action recognition and detection.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. Vitpose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*.

Zhaoqilin Yang, Gaoyun An, Zhenxing Zheng, Shan Cao, and Fengjuan Wang. 2024. Epk-clip: External and priori knowledge clip for action recognition. *Expert Systems with Applications*, 252:124183.

Seth Z. Zhao, Reza Ghoddoosian, Isht Dwivedi, Nakul Agarwal, and Behzad Dariush. 2025. Pose-aware weakly-supervised action segmentation.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. 2023. Detrs beat yolos on real-time object detection.