ZERO-SHOT SKELETON-BASED TEMPORAL ACTION SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human action segmentation in continuous skeleton streams relies heavily on closed-set annotations, struggling to generalize to novel categories in the wild. While zero-shot learning (ZSL) offers a promising paradigm, existing skeleton-based ZSL methods are strictly confined to clip-level recognition. By encoding entire actions into global semantic prototypes, they inherently erase the fine-grained, localized motion primitives shared between seen and unseen categories, rendering them ineffective for dense frame-level segmentation. In this paper, we formalize and tackle the novel task of zero-shot skeleton-based temporal action segmentation. We propose SkelAtom, a framework that achieves compositional generalization by explicitly decomposing untrimmed actions into sparse combinations of shared motion atoms. To ensure these atoms capture transferable physical movements, we introduce Kinematic Anchor Grounding (KAG), which supervises the latent space with class-agnostic kinematic descriptors derived from body-part dynamics. Furthermore, a Cross-Modal Atom Codebook aligns both visual frames and textual prototypes into the same atomic basis, enabling visual knowledge learned from seen actions to transfer to unseen categories. We establish four zero-shot benchmarks on the MCFS-130 and PKU-MMD datasets. Experimental results show that SkelAtom achieves state-of-the-art performance.

1 INTRODUCTION

Human action segmentation is a core problem in computer vision and underpins a wide range of applications, including video understanding (Ding et al., 2023), human-computer interaction (Terreran et al., 2023), and medical rehabilitation (Guo et al., 2024). Compared with RGB video, skeleton sequences explicitly preserve the spatio-temporal topology of body joints and motion trajectories, while offering privacy protection, compact representation, and robustness to background clutter, which makes them an increasingly important modality for action understanding.

Existing skeleton-based action segmentation methods (Yang et al., 2023; Li et al., 2026; Ji et al., 2026a; Filtjens et al., 2022) mostly rely on abundant annotations and assume that the action categories seen during training and testing are identical, i.e., the closed-set setting. In open and dynamic real-world environments, novel action categories continually emerge, and supplying frame-level boundary annotations for continuous untrimmed sequences is far more costly than labeling trimmed clips. This reliance on annotations for unseen categories forms a key obstacle to the practical deployment of skeleton-based action segmentation.

Zero-shot learning (ZSL) provides a natural way to move beyond the closed-set assumption, as it treats the semantic description of a category as a bridge that transfers knowledge from seen classes to unseen ones. Most existing zero-shot skeleton studies, however, target action recognition on trimmed clips (Gupta et al., 2021; Li et al., 2024a; Zhou et al., 2023; Chen et al., 2025), where the goal is to assign an unseen category to a short sequence containing a single action, and each category is encoded as one global prototype for matching. The genuine bridge toward unseen categories lies instead in the local motion components shared between seen and unseen actions, since elementary movements such as raising an arm, rotating the torso, and shifting the body weight constitute both familiar and never-observed actions, and compressing a category into a global prototype erases exactly this transferable structure. How to advance zero-shot transfer from clip-level recognition to frame-level dense segmentation therefore remains an open question.

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.

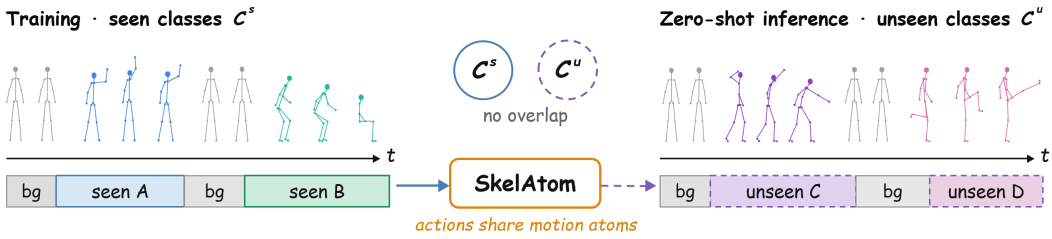


Figure 1: Our zero-shot skeleton-based temporal action segmentation framework. SkelAtom is trained on skeleton streams of seen classes $\mathcal{C}^s$ and tested on streams of unseen classes $\mathcal{C}^u$, with no overlap between them. Since seen and unseen actions share motion atoms, SkelAtom represents each action as a combination of atoms and predicts a label for every frame of the untrimmed stream.

In this paper, we present SkelAtom, a method for zero-shot skeleton-based temporal action segmentation. SkelAtom explicitly decomposes an action into a combination of a small set of shared motion atoms, each of which characterizes the coordinated movement of several body parts within a short time span. Because elementary movements are shared between seen and unseen actions, this set of atoms is shared across categories and carries the decomposition knowledge acquired on seen classes over to unseen ones. Specifically, we propose Kinematic Anchor Grounding (KAG) to endow the atoms with genuine physical meaning. KAG computes deterministic motion descriptors directly from the raw skeleton, covering the distribution of motion energy over body parts together with its variation along the action duration, and uses these descriptors to supervise the frame features. Furthermore, since most current methods share semantics only on the textual side and overlook visual semantics, we design a Cross-Modal Atom Codebook that lets visual frame features and textual prototypes share a single set of atomic bases, so that both the visual and the textual information of seen classes are passed to unseen classes through this codebook (Figure 1).

To assess the effectiveness of our method, we construct four zero-shot settings on MCFS-130 and PKU-MMD, two benchmarks for skeleton-based temporal action segmentation. As no directly comparable method exists for this problem, we adapt representative zero-shot skeleton recognition methods to the same segmentation protocol and use them as baselines. Experiments show that our method yields consistent improvements over these adapted baselines across all four settings.

The main contributions of this paper are summarized as follows:

- We study zero-shot skeleton-based temporal action segmentation, which advances zero-shot transfer from clip-level recognition to frame-level dense segmentation and semantic prediction on continuous untrimmed skeleton streams.
- We ground each atom in a measurable amount of body-part motion, so that an action is decomposed into a combination of motion primitives shared between seen and unseen categories, which provides unseen categories with a transferable intermediate structure that carries knowledge from both the visual and the textual sides.
- We establish baselines under a unified protocol for this new problem, and obtain consistent improvements over adapted zero-shot recognition baselines across four zero-shot settings on MCFS-130 and PKU-MMD.

2 Related Work

Temporal Action Segmentation. Temporal action segmentation divides an untrimmed video into semantically coherent segments and assigns a category label to each of them. Early methods refine frame-level predictions stage by stage with temporal convolutions and smoothing constraints (Lea et al., 2017; Farha & Gall, 2019; Li et al., 2020), and later work introduces explicit boundary modeling to alleviate over-segmentation (Wang et al., 2020; Ishikawa et al., 2021). Transformer-based methods aggregate context adaptively, as FACT (Lu & Elhamifar, 2024) refines a frame-level and an action-level branch through cross-attention and LTContext (Bahrami et al., 2023) examines how

much long-range context the task requires. Generative and structured models offer further perspectives, with DiffAct (Liu et al., 2023) casting segmentation as conditional sequence generation and KARI (Gong et al., 2023) constraining predictions with recursive grammars. All these methods rest on the closed-set assumption and cannot label action semantics outside the training set.

Skeleton-Based Temporal Action Segmentation. Skeleton sequences describe human motion with compact and privacy-preserving joint coordinates that are insensitive to illumination and background. MS-GCN (Filtjens et al., 2022) combines graph convolution with multi-stage temporal modeling and serves as a representative baseline. Subsequent work models joint- and action-level semantics with text embeddings or relational graphs (Ji et al., 2024; 2025), strengthens boundary discrimination and spatio-temporal modeling with frequency filtering, Lagrangian dynamics, and curvature guidance (Ji et al., 2026a;b; Li et al., 2026), or enhances segment-level modeling with composable latent actions (Yang et al., 2023). To reduce the cost of frame-level annotation, other studies explore self-supervised contrastive pre-training (Tian, 2025), training on stitched trimmed clips (Tian & Payeur, 2025), local dimension enhancement (Sun et al., 2026), and unsupervised quantization into motion words (Gökay et al., 2025). Whatever supervision they adopt, these methods remain closed-set and generalize poorly to novel actions.

Zero-Shot Action Understanding. Zero-shot action understanding connects seen and unseen categories through textual semantics. In RGB video, ZSTAD (Zhang et al., 2020) formalizes zero-shot temporal action detection, and subsequent CLIP-based methods mitigate the error propagation between localization and classification through parallel branches or decoupled networks (Nag et al., 2022; Li et al., 2024b; Phan et al., 2024), or improve detection with unified frameworks, fine-grained cross-modal interaction, and text-enhanced actionness (Yan et al., 2023; Lee et al., 2024; Raza et al., 2024). Recent studies further explore generative priors, self-training, chain-of-thought prompting, diffusion denoising, and training-free memory retrieval (Susladkar et al., 2024; Hyun et al., 2025; Zhu et al., 2026a;b; Lee et al., 2026; Jiang et al., 2026; Liberatori et al., 2024). In the skeleton modality, zero-shot action recognition bridges the skeleton and semantic spaces through syntactic guidance, mutual information alignment, or disentangled variational autoencoders (Gupta et al., 2021; Zhou et al., 2023; Li et al., 2024a). Later methods mine finer-grained part and semantic cues (Zhu et al., 2024; Chen et al., 2024; Kuang et al., 2025), build prompt learning and context-evolving representations on pre-trained vision-language models (Zhu et al., 2025a; Wang et al., 2026; Chen et al., 2025), or adopt frequency-semantic enhancement, diffusion-driven alignment, and training-free test-time adaptation (Wu et al., 2025; Do & Kim, 2025; Zhu et al., 2025b). Neither family transfers to our problem, since detection methods take RGB input and produce sparse proposals, whereas skeleton recognition methods label trimmed single-action clips. Frame-level segmentation of untrimmed skeleton streams under the zero-shot setting, which must transfer semantics while keeping the temporal structure plausible, remains unexplored.

3 METHOD

3.1 PROBLEM DEFINITION AND OVERALL FRAMEWORK

Given a skeleton sequence $\mathbf{X} \in \mathbb{R}^{C\times V\times T}$, where C is the number of feature channels per joint, V the number of joints, and T the number of frames, we denote by $\mathbf{x}_{j,t} \in \mathbb{R}^3$ the coordinate of joint j at frame t and by $\mathbf{x}_{\mathrm{root},t}$ the coordinate of the root joint. Zero-shot skeleton-based temporal action segmentation aims to predict a semantic label for every frame. Let $\mathcal{C}^s$ be the set of seen categories and $\mathcal{C}^u$ the set of unseen categories, which are disjoint ($\mathcal{C}^s \cap \mathcal{C}^u = \varnothing$). Training provides frame-level annotations only for the seen action categories and the background class bg, while the candidate set at test time consists of the background class and the unseen action categories. Unlike zero-shot recognition, which makes a single-label decision inside an already segmented clip, this task requires dense prediction for every frame of a continuous untrimmed stream.

We adopt a text-aligned discriminative framework. A skeleton encoder $\Phi(\cdot)$ maps $\mathbf{X}$ to frame-wise features $\{\mathbf{f}_t\}_{t=1}^{T}$ with $\mathbf{f}_t \in \mathbb{R}^d$. A frozen text encoder $\Psi(\cdot)$ produces a textual prototype $\mathbf{p}_c \in \mathbb{R}^d$ for each category c. We encode a detailed description of the category and M descriptions of distinct temporal phases, all generated by a large language model, and average their embeddings into $\mathbf{p}_c$, so that the temporal evolution of the action is retained explicitly on the text side. The prompts are given in the appendix.

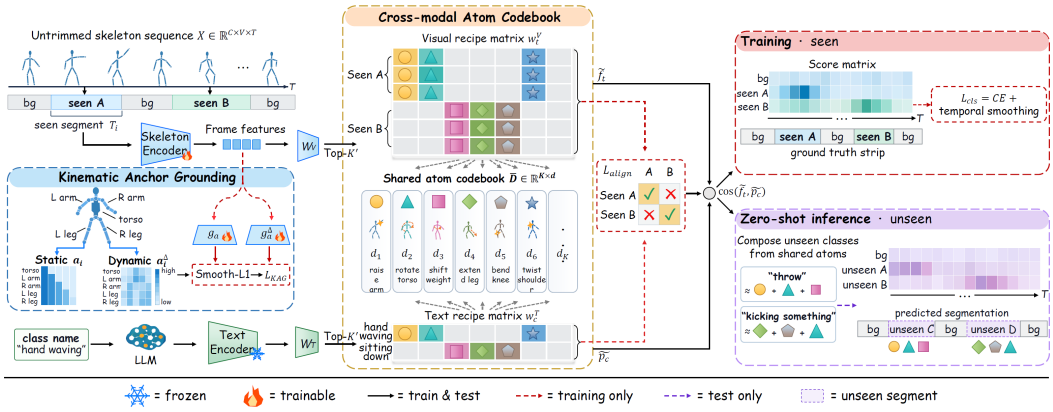


Figure 2: Overview of SkelAtom. Kinematic anchor grounding computes class-agnostic static and dynamic motion descriptors from the raw skeleton and regresses them from segment features, so that the frame features retain measurable body-part dynamics. Visual frames and textual prototypes query one shared atom codebook and are expressed as sparse K'-atom recipes over the same atoms, which are aligned at the recipe level on seen classes. At inference, an unseen class is read as a new combination of the same atoms and matched against every frame. Atom names are illustrative.

The core of SkelAtom is a set of motion atoms shared across categories. Around these atoms we propose two components. Kinematic anchor grounding (Section 3.2) computes measurable motion quantities from the raw skeleton and uses them to supervise the features. The cross-modal atom codebook (Section 3.3) lets visual frames and textual prototypes share one atomic basis, so that the visual and textual knowledge of seen categories transfers through this codebook to unseen categories, for which only textual descriptions are available. Figure 2 presents the overall framework.

3.2 Kinematic Anchor Grounding

If the atom codebook is learned freely under cross-modal alignment alone, it tends to converge to discriminative shortcuts that separate seen categories rather than to motion primitives that transfer across categories. Kinematic anchor grounding (KAG) therefore constrains the features with class-agnostic motion quantities computed directly from the raw skeleton.

We partition the body into P parts $\{b_p\}_{p=1}^{P}$, each a group of joints. Let $\mathcal{S}$ be the set of seen action segments, where segment i spans the frames $\mathcal{T}_i = [t_i^s, t_i^e]$ and has category $y_i \in \mathcal{C}^s$. For a joint set b and a frame set $\mathcal{T}$, the average motion energy is

$$E(b, \mathcal{T}) = \frac{1}{|b||\mathcal{T}|} \sum_{j \in b} \sum_{t \in \mathcal{T}} \left\| \mathbf{x}_{j,t} - \mathbf{x}_{j,t-1} \right\|_2, \tag{1}$$

which follows directly from the raw skeleton and requires no category information. On segment i we write $E_p = E(b_p, \mathcal{T}_i)$ for part p and $E_* = E(\bigcup_p b_p, \mathcal{T}_i)$ for the whole body.

The static descriptor summarizes how motion energy is distributed over the body within a segment:

$$\mathbf{a}_i = \left[E_1, \ldots, E_P, \boldsymbol{\rho}_i \right], \tag{2}$$

where $\boldsymbol{\rho}_i$ collects relative-activity ratios, namely the energy shares of the limbs and the torso, the left-right asymmetry of the arms and the legs, and the limb activity relative to the torso, together with the change of the overall intensity from the beginning to the end of the segment and a pose-extent term. The dynamic descriptor captures how this distribution evolves over time. Consistent with the textual prototype, we divide $\mathcal{T}_i$ uniformly into M phases $\{\mathcal{T}_i^{(m)}\}_{m=1}^{M}$ and write $\pi_p^{(m)} = E(b_p, \mathcal{T}_i^{(m)})/E_*^{(m)}$ for the energy share of part p in phase m, where $E_*^{(m)}$ is the whole-body energy of that phase. The descriptor reads

$$\mathbf{a}_i^{\Delta} = \left[\left\{ \pi_p^{(M)} - \pi_p^{(1)} \right\}_p, \left\{ \pi_p^{(m)} - \tfrac{1}{2}\left(\pi_p^{(1)} + \pi_p^{(M)} \right) \right\}_{p,\, 1<m<M}, \boldsymbol{\eta}_i \right], \tag{3}$$

where the first two groups capture the first-order trend and the second-order curvature of the energy shares, and $\boldsymbol{\eta}_i$ records how the overall intensity, the arm-leg dominance, and the left-right symmetry change from the first to the last phase. The body partition and the exact form of $\boldsymbol{\rho}_i$ and $\boldsymbol{\eta}_i$ are given in Appendix B.

Both descriptors are compressed by a signed logarithm and normalized to unit length, $\phi(\mathbf{a}) = \mathbf{u}/\|\mathbf{u}\|_2$ with $\mathbf{u} = \operatorname{sign}(\mathbf{a}) \odot \log(1 + |\mathbf{a}|)$, where $\odot$ denotes element-wise multiplication. Taking the temporal average $\bar{\mathbf{f}}_i = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \mathbf{f}_t$ of each segment as input, two prediction heads g_a and g_a^{Δ} regress these targets:

$$\mathcal{L}_{\mathrm{KAG}} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \Big[\ell_1^{\mathrm{sm}}\big(g_a(\bar{\mathbf{f}}_i), \phi(\mathbf{a}_i)\big) + \ell_1^{\mathrm{sm}}\big(g_a^{\Delta}(\bar{\mathbf{f}}_i), \phi(\mathbf{a}_i^{\Delta})\big) \Big], \tag{4}$$

where ℓ_1^{sm} is the Smooth-L1 loss, and the loss is averaged over the refinement stages of the skeleton encoder. KAG adds only two lightweight heads, which are discarded at inference, while its gradient flows into the skeleton encoder and forces the features to retain measurable motion structure, so that the atoms composed over these features inherit the same physical grounding.

3.3 Cross-Modal Atom Codebook

The cross-modal atom codebook maps visual frames and textual prototypes to sparse recipes over the same set of atoms, so that the atomic basis is shared across the two modalities and, through them, across seen and unseen categories.

Shared atomic basis. Let $\mathbf{D} \in \mathbb{R}^{K \times d}$ be the shared atom codebook with K atoms. Each atom is normalized to unit length as $\bar{\mathbf{D}}_k$, which keeps all atoms on a common scale, so that a recipe is determined by which atoms are selected and how they are weighted rather than by differences in atom norms. We express a feature as a combination of atoms through a sparse cross-attention mechanism: the query comes from a modality-specific projection of the feature, while the keys and values are the same set of normalized atoms $\{\bar{\mathbf{D}}_k\}_{k=1}^{K}$, of which only the K' most responsive ones are kept ($K' \ll K$) so that the resulting recipe stays sparse and identifiable. For a frame feature $\mathbf{f}_t$, let $\alpha_{t,k} = (\mathbf{W}_{\mathrm{V}}\mathbf{f}_t)^{\top}\bar{\mathbf{D}}_k/\sqrt{d}$ be the attention score and let Ω_t be the index set of the K' largest entries of $\{\alpha_{t,k}\}_{k=1}^{K}$. The components of the sparse recipe are

$$w_{t,k}^{\mathrm{V}} = \frac{\mathbb{I}[k \in \Omega_t]\, \exp(\alpha_{t,k})}{\sum_{k' \in \Omega_t} \exp(\alpha_{t,k'})}, \tag{5}$$

where $\mathbb{I}[\cdot]$ is the indicator function. We write the recipe vector as $\mathbf{w}_t^{\mathrm{V}} = [w_{t,k}^{\mathrm{V}}]_{k=1}^{K} \in \mathbb{R}^K$, and the atomic reconstruction of the frame is $\mathbf{r}_t^{\mathrm{V}} = \sum_{k=1}^{K} w_{t,k}^{\mathrm{V}} \bar{\mathbf{D}}_k$. The text side proceeds in the same way, using a separate query projection $\mathbf{W}_{\mathrm{T}}$ on the prototype $\mathbf{p}_c$ to obtain the recipe $\mathbf{w}_c^{\mathrm{T}} \in \mathbb{R}^K$ and the reconstruction $\mathbf{r}_c^{\mathrm{T}}$. The two modalities use their own query projections to accommodate different input distributions, yet share the same codebook $\mathbf{D}$, that is, the same set of atoms as the basis for composition. Keeping the recipe sparse ensures that it carries identifiable compositional semantics rather than a dense weighting over all atoms.

The atoms are injected in a residual manner so that the aligned cosine geometry is preserved:

$$\tilde{\mathbf{f}}_t = \mathbf{f}_t + \gamma\, \mathbf{r}_t^{\mathrm{V}}, \qquad \tilde{\mathbf{p}}_c = \mathbf{p}_c + \gamma\, \mathbf{r}_c^{\mathrm{T}}, \tag{6}$$

where γ is the residual strength.

Cross-modal recipe alignment. A shared basis alone does not guarantee that the two modalities select the same atoms for the same action. To establish a class-conditional correspondence, we require the visual segments of a seen category and its textual prototype to select the same recipe. For segment i, we take the temporal average of its visual recipes $\bar{\mathbf{w}}_i^{\mathrm{V}} = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \mathbf{w}_t^{\mathrm{V}}$ and match it to the textual recipe of its category in the shared recipe space:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \big\| \bar{\mathbf{w}}_i^{\mathrm{V}} - \mathbf{w}_{y_i}^{\mathrm{T}} \big\|_2^2. \tag{7}$$

Both recipes are sparse points of the same simplex, so this term drives the two modalities to select the same atoms with the same weights. Once vision and text correspond at the recipe level on seen categories, the textual prototype of an unseen category can be read at inference as a transferable combination over the same set of atoms.

3.4 Training and Inference

Training. Let $y_t \in \{\mathrm{bg}\} \cup \mathcal{C}^s$ be the ground-truth label of frame t. Frame-to-category scores are given by cosine similarity, and the components of the log-probability vector $\mathbf{q}_t$ of frame t over the training candidate set are

$$q_{t,c} = \log \frac{\exp\big(\cos(\tilde{\mathbf{f}}_t, \tilde{\mathbf{p}}_c)/\tau_c\big)}{\sum_{c' \in \{\mathrm{bg}\} \cup \mathcal{C}^s} \exp\big(\cos(\tilde{\mathbf{f}}_t, \tilde{\mathbf{p}}_{c'})/\tau_c\big)}. \tag{8}$$

The frame-wise classification loss then combines a cross-entropy term with a truncated temporal smoothing term:

$$\mathcal{L}_{\mathrm{cls}} = -\frac{1}{T}\sum_{t=1}^{T} q_{t,y_t} + \lambda_{\mathrm{sm}} \frac{1}{T-1}\sum_{t=2}^{T} \min\big(\|\mathbf{q}_t - \mathbf{q}_{t-1}\|_2^2,\, \epsilon\big), \tag{9}$$

where the second term suppresses prediction jitter between adjacent frames, the truncation threshold ϵ keeps large jumps at action boundaries from being penalized excessively, and λ_{sm} is its weight. The overall objective consists of frame-wise classification, kinematic anchor grounding, and cross-modal recipe alignment:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{KAG}}\mathcal{L}_{\mathrm{KAG}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}, \tag{10}$$

with weights λ_{KAG} and λ_{align}. No text of any unseen category is accessed at any point during training, so the zero-shot constraint holds strictly.

Inference. To reduce the prediction bias between background and action frames, we subtract a calibration scalar β from the background logit and take the frame-wise maximum as the prediction:

$$\hat{y}_t = \arg\max_{c \in \{\mathrm{bg}\} \cup \mathcal{C}^u} \big(\cos(\tilde{\mathbf{f}}_t, \tilde{\mathbf{p}}_c)/\tau_c - \beta\, \mathbb{I}[c = \mathrm{bg}]\big). \tag{11}$$

4 Experiments

4.1 Datasets and Metrics

We evaluate the proposed method on two benchmarks for skeleton-based temporal action segmentation, and repartition the action categories to follow the zero-shot setting.

MCFS-130 (Liu et al., 2021) is a motion-centric figure skating dataset whose action sequences are fine-grained, long in duration, and subject to large intra-class variation. MCFS-130 annotates actions with 130 fine-grained categories, which places a high demand on semantic modeling. We construct two unseen-class splits, U13 with 13 unseen categories and U26 with 26 unseen categories, the latter being more challenging.

PKU-MMD (Liu et al., 2017) is a large-scale benchmark for continuous action understanding, captured by three Kinect v2 cameras. It contains about 20,000 action instances covering 51 daily action categories, and provides the 3D coordinates of 25 body joints for every frame. The dataset offers two standard protocols, cross-subject (X-sub) and cross-view (X-view), which examine robustness to individual differences and to viewpoint changes respectively. Under both protocols we take 10 categories as unseen.

Metrics. We report frame-wise accuracy (Acc), the edit score (Edit), and the segmental F1 score at IoU thresholds of 10%, 25%, and 50%, together with AvgF1, the mean of the three F1 scores.

4.2 Implementation Details

We adopt MS-GCN (Filtjens et al., 2022) as the skeleton encoder Φ and freeze the CLIP ViT-B/32 text encoder. We use $K = 32$ atoms with $K' = 4$ kept per recipe, $M = 3$ temporal phases, and a

Table 1: Comparison with state-of-the-art zero-shot skeleton-based temporal action segmentation baselines on MCFS-130 under the U13 and U26 unseen-class splits. **Bold** and underline denote the best and second-best results.

Method	Venue	MCFS-130 U13						MCFS-130 U26					
		Acc	Edit	F1@{10, 25, 50}			AvgF1	Acc	Edit	F1@{10, 25, 50}			AvgF1
SMIE (Zhou et al., 2023)	MM'23	42.65	41.63	38.32	34.73	22.36	31.80	37.06	25.77	27.80	22.91	13.38	21.36
DVTA (Kuang et al., 2025)	PR'25	48.90	36.17	40.99	39.67	26.58	35.75	39.53	29.58	30.09	26.36	17.31	24.59
SynSE (Gupta et al., 2021)	ICIP'21	52.80	38.26	34.55	32.56	20.27	29.13	42.54	27.22	21.35	18.81	10.96	17.04
SA-DVAE (Li et al., 2024a)	ECCV'24	55.11	33.92	36.59	34.84	21.25	30.89	41.76	26.96	21.47	19.92	11.83	17.74
TDSM (Do & Kim, 2025)	ICCV'25	45.34	30.73	27.15	25.94	17.80	23.63	42.24	29.13	28.50	25.65	14.96	23.04
STALE (Nag et al., 2022)	ECCV'22	56.25	36.95	42.74	35.46	20.41	32.87	43.05	27.96	27.33	24.26	17.56	23.05
Ti-FAD (Lee et al., 2024)	NeurIPS'24	56.26	35.92	42.98	38.72	24.51	35.40	35.83	32.30	34.98	31.77	20.69	29.15
SkelAtom	–	**59.31**	**42.79**	**43.58**	**40.30**	**28.66**	**37.51**	**53.64**	**41.25**	**39.53**	**36.34**	**22.95**	**32.94**

Table 2: Comparison with state-of-the-art zero-shot skeleton-based temporal action segmentation baselines on PKU-MMD under the X-sub and X-view unseen-class splits (U10). **Bold** and underline denote the best and second-best results.

Method	Venue	PKU-MMD X-sub U10						PKU-MMD X-view U10					
		Acc	Edit	F1@{10, 25, 50}			AvgF1	Acc	Edit	F1@{10, 25, 50}			AvgF1
SMIE (Zhou et al., 2023)	MM'23	39.30	36.57	38.55	35.23	25.93	33.24	38.91	34.10	37.50	34.30	25.74	32.51
DVTA (Kuang et al., 2025)	PR'25	43.09	36.78	38.32	33.96	25.54	32.61	38.24	32.92	37.03	34.97	25.89	32.63
SynSE (Gupta et al., 2021)	ICIP'21	40.80	29.15	32.58	29.70	23.27	28.52	46.80	35.29	37.24	34.67	28.95	33.62
SA-DVAE (Li et al., 2024a)	ECCV'24	44.56	36.86	35.97	31.60	23.92	30.50	46.43	35.52	39.56	36.60	28.12	34.76
TDSM (Do & Kim, 2025)	ICCV'25	41.79	36.38	37.39	33.58	24.15	31.71	41.16	33.72	32.07	29.12	21.86	27.68
STALE (Nag et al., 2022)	ECCV'22	45.94	40.60	44.30	38.55	24.65	35.83	44.90	35.98	43.99	40.09	24.47	36.18
Ti-FAD (Lee et al., 2024)	NeurIPS'24	46.86	41.54	41.82	36.85	26.72	35.13	44.13	36.12	43.57	38.86	26.98	36.47
SkelAtom	–	**48.46**	**43.21**	**45.58**	**39.48**	**29.23**	**38.10**	**47.56**	**41.78**	**44.33**	**40.14**	**30.81**	**38.43**

residual strength $\gamma = 0.5$ for both branches. The loss weights are $\lambda_{\text{KAG}} = 0.1$ and $\lambda_{\text{align}} = 0.5$, and the background scalar β is selected on a validation split of seen categories. The model is trained with Adam for 100 epochs, using a learning rate of 10^{-3}, a batch size of 16, and a weight decay of 10^{-4}. All experiments run on a single RTX 4090. Further details are given in the appendix.

4.3 Comparisons with the State-of-the-Art

We compare our method with state-of-the-art zero-shot action recognition and temporal localization methods on MCFS-130 and PKU-MMD, with detailed results reported in Table 1 and Table 2.

Results on MCFS-130. Table 1 reports the results under the U13 and U26 zero-shot splits of MCFS-130, where SkelAtom outperforms all baselines on every metric. Under U13, it improves frame-wise accuracy by 3.05% and AvgF1 by 1.76% over the second-best method. The margin widens under U26, a split that doubles the number of unseen categories and causes a pronounced degradation for most baselines, whereas SkelAtom reaches 53.64% accuracy and an edit score of 41.25%, and improves AvgF1 by 3.79%. These results indicate that methods relying on a single global prototype per category struggle with the fine-grained figure skating actions of MCFS-130, while sharing local motion atoms establishes a physically grounded bridge between seen and unseen actions and preserves both dense segmentation quality and semantic transfer under the more demanding setting.

Results on PKU-MMD. Table 2 reports the performance under the cross-subject (X-sub) and cross-view (X-view) zero-shot splits (U10) of PKU-MMD, where SkelAtom again achieves consistent state-of-the-art results. Under X-sub, it attains 29.23% on F1@50, a metric that requires accurate action boundaries, exceeding the second-best method Ti-FAD by 2.51% and showing that the temporal extent of unseen actions is localized more precisely. Under X-view, where the test viewpoint changes substantially, SkelAtom reaches an edit score of 41.78% and surpasses the second-best method by 5.66%. Such a gain on the edit score means that the predictions are more coherent in

Table 3: Component ablation on MCFS-130 U13 and PKU-MMD X-sub. Best results are highlighted in bold.

KAG	Codebook	MCFS-130 U13					PKU-MMD X-sub				
		Acc	Edit	F1@{10, 25, 50}			Acc	Edit	F1@{10, 25, 50}		
✗	✗	55.93	39.60	40.30	38.55	24.77	39.53	40.39	40.00	37.49	23.78
✓	✗	59.05	41.78	**43.97**	40.23	27.26	42.62	42.14	44.49	39.20	26.12
✗	✓	57.57	41.06	42.50	40.06	26.76	42.49	42.55	43.51	**40.56**	25.17
✓	✓	**59.31**	**42.79**	43.58	**40.30**	**28.66**	**48.46**	**43.21**	**45.58**	39.48	**29.23**

Table 4: Effect of kinematic supervision.

Supervision	Acc	Edit	F1@{10, 25, 50}		
None	42.49	42.55	43.51	40.56	25.17
Shuffled	45.94	42.74	41.32	37.05	26.26
Static	48.00	42.97	44.87	**40.69**	24.85
Dynamic	47.58	42.27	43.95	39.79	25.84
S+D	**48.46**	**43.21**	**45.58**	39.48	**29.23**

Table 5: Effect of atom codebook design.

Variant	Acc	Edit	F1@{10, 25, 50}		
w/o sparse	45.46	40.43	44.31	37.75	26.71
w/o norm.	47.48	42.44	44.31	38.47	28.19
Independent	48.10	**44.02**	43.59	38.37	24.35
Shared query	47.74	43.39	**45.77**	39.05	28.54
Full	**48.46**	43.21	45.58	**39.48**	**29.23**

temporal structure, which alleviates the over-segmentation commonly observed in zero-shot dense prediction. We attribute this behavior to kinematic anchor grounding, since the energy and phase quantities computed directly from the raw skeleton are inherently robust to body size and viewpoint changes, so that the learned motion atoms remain stable when they are extracted and recombined across subjects and views.

4.4 Ablation Studies

Effect of components. Table 3 reports the contribution of kinematic anchor grounding (KAG) and the cross-modal atom codebook. Without either component, frame-wise accuracy on PKU-MMD is 39.53%. Introducing KAG or the codebook alone already brings a clear gain, and the two together give the best overall performance, which points to a complementary relationship between them. KAG endows the features with measurable physical meaning and prevents spurious correlations, while the codebook carries the resulting decomposition from seen categories to unseen ones.

Effect of kinematic anchor grounding. Table 4 examines the supervision used in KAG on PKU-MMD X-sub. Removing it (None) lowers accuracy from 48.46% to 42.49%. Supervising the features with descriptors of mismatched segments (Shuffled) recovers only part of this gap, which indicates that the gain relies on the motion quantities of each segment rather than on the extra regression objective alone. The two descriptors are complementary, as the static one yields higher accuracy and the dynamic one a higher F1@50, and combining them (S+D) gives the best accuracy.

Effect of cross-modal atom codebook. Table 5 ablates the design of the codebook on PKU-MMD X-sub. Removing the sparsity constraint (w/o sparse) causes the largest drop in accuracy, since dense attention blends all atoms and destroys the identifiable composition that unseen categories rely on. A separate codebook for each modality (Independent) mainly hurts segment quality and lowers F1@50 from 29.23% to 24.35%, confirming that a shared basis is the bridge for cross-modal transfer. Sharing the query projection (Shared query) or removing the atom normalization (w/o norm.) also degrades performance, showing that the two modalities need their own projections and that the atoms should stay on a common scale.

4.5 Qualitative Results

Sensitivity analysis. Figure 3 reports the sensitivity to the number of atoms K and the sparsity K' on PKU-MMD X-sub. A small codebook forces different movements to share the same atom and clearly degrades F1@50 and the edit score, whereas enlarging it beyond 32 atoms brings little change, suggesting that 32 atoms suffice to cover the motion primitives in the data. Performance peaks at $K' = 4$. Selecting a single atom per frame removes the compositional structure, while

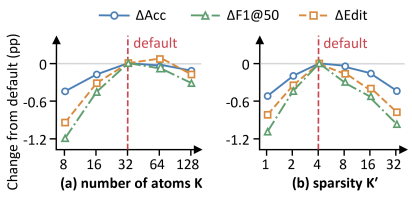


Figure 3: Sensitivity to the number of atoms K and the sparsity K' on PKU-MMD X-sub.

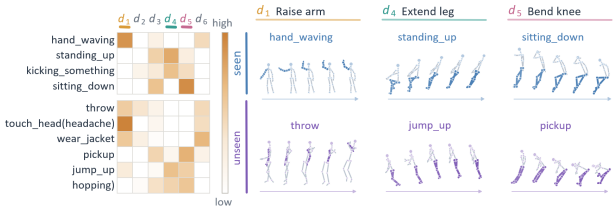


Figure 4: Shared atoms on PKU-MMD. Left: recipe weights of seen and unseen classes over six atoms. Right: segments that most activate three atoms.

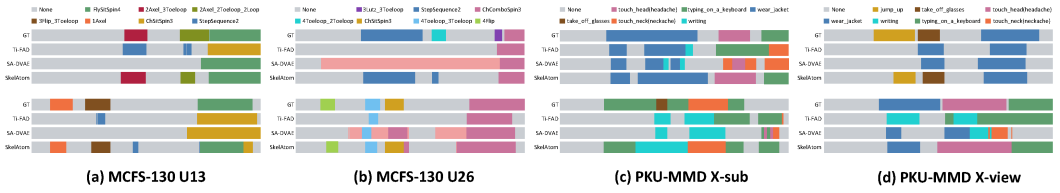


Figure 5: Qualitative segmentation on unseen categories under the four zero-shot settings.

denser recipes mix more atoms and degrade performance gradually. In both cases, the segment-level metrics are more sensitive than frame-wise accuracy, and performance remains stable for $K \in [16, 128]$ and $K' \in [2, 16]$.

Visualization of shared atoms. Figure 4 illustrates how the atoms are shared across categories on PKU-MMD. The recipe matrix on the left shows that unseen classes reuse atoms already active on seen classes rather than introducing new ones. For instance, the unseen classes *throw*, *touch head*, and *wear jacket* rely mainly on d_1, which also dominates *hand waving*; *pick up* shares d_3 and d_5 with *sitting down*; and *jump up* and *hopping* combine d_4 and d_5, which are also active on *standing up* and *kicking something*. The right side shows the segments that most strongly activate three atoms, and each atom captures a consistent movement in both seen and unseen sequences, namely arm raising for d_1, leg extension for d_4, and knee bending for d_5. The atoms therefore encode motion primitives that transfer to unseen categories rather than cues specific to the seen ones.

Qualitative segmentation results. Figure 5 compares the frame-wise predictions of SkelAtom with those of Ti-FAD and SA-DVAE on two test sequences from each zero-shot setting. The baselines tend either to assign a long stretch of frames to a wrong category or to split one action into many short fragments, as in the MCFS-130 U26 and PKU-MMD sequences. SkelAtom recognizes most unseen actions correctly and produces coherent segments whose boundaries closely follow the ground truth, consistent with its higher edit and F1 scores. The remaining errors mainly involve very short segments, such as the brief jumps in the first U26 sequence, and actions with similar hand motion, such as typing and writing in the second X-sub sequence, which are likely composed of similar atoms.

5 Conclusion

In this paper, we introduce SkelAtom, the first framework for zero-shot skeleton-based temporal action segmentation, which decomposes actions into sparse combinations of motion atoms shared across seen and unseen categories. Our proposed kinematic anchor grounding supervises the features with class-agnostic motion descriptors computed from the raw skeleton, so that the atoms are composed over a feature space that retains measurable body-part dynamics. The cross-modal atom codebook expresses visual frames and textual prototypes over the atoms and aligns their recipes on seen categories, which allows an unseen category to be recognized as a new combination of known atoms. We construct four zero-shot settings on MCFS-130 and PKU-MMD with a unified protocol and adapted baselines, on which SkelAtom achieves consistent improvements. We hope that this task and its benchmarks will support further research on open-world action segmentation.

AI USE STATEMENT

We used GPT-4 to generate the textual descriptions of the action categories, as described in Section 3.1 and listed in the appendix. Apart from this, large language models were used to improve the grammar, clarity, and overall readability of the manuscript. All scientific contributions, including the core ideas, the method, the experiments, and the conclusions, are the work of the human authors, and the use of these tools did not alter the scientific content or its interpretation.

ETHICS STATEMENT

This work uses only the public benchmarks MCFS-130 (Liu et al., 2021) and PKU-MMD (Liu et al., 2017) under their original terms, and no new data were collected from human subjects. Both datasets represent motion as skeleton joint coordinates, which contain no facial appearance or other identifying visual information. Although action segmentation could be misused for surveillance without consent, our method only recognizes coarse action categories from anonymized skeletons, and we encourage its use in consented settings such as rehabilitation and human-computer interaction.

REPRODUCIBILITY STATEMENT

Sections 3.2–3.4 and Appendix B describe the model and all components of the kinematic descriptors. Section 4.2 gives the skeleton encoder, the main hyperparameters, and the training schedule, and the appendix provides the remaining hyperparameters, the prompts for generating class descriptions, the protocol for adapting the baselines, and the unseen categories of all four zero-shot splits. The code for preprocessing, training, and evaluation will be released upon acceptance.

REFERENCES

Emad Bahrami, Gianpiero Francesca, and Juergen Gall. How much temporal long-term context is needed for action segmentation? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10351–10361, 2023.

David Barber and Felix Agakov. The IM algorithm: A variational approach to information maximization. In *Advances in Neural Information Processing Systems*, volume 16, 2003.

Yang Chen, Jingcai Guo, Tian He, Xiaocheng Lu, and Ling Wang. Fine-grained side information guided dual-prompts for zero-shot skeleton action recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 778–786, 2024.

Yang Chen, Jingcai Guo, Song Guo, and Dacheng Tao. Neuron: Learning context-aware evolving representations for zero-shot skeleton action recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8721–8730, 2025.

Guodong Ding, Fadime Sener, and Angela Yao. Temporal action segmentation: An analysis of modern techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2): 1011–1030, 2023.

Jeonghyeok Do and Munchurl Kim. Bridging the skeleton-text modality gap: Diffusion-powered modality alignment for zero-shot skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12757–12768, 2025.

Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3575–3584, 2019.

Benjamin Filtjens, Bart Vanrumste, and Peter Slaets. Skeleton-based action segmentation with multi-stage spatial-temporal graph convolutional neural networks. *IEEE Transactions on Emerging Topics in Computing*, 12(1):202–212, 2022.

Uzay Gökay, Federico Spurio, Dominik R Bach, and Juergen Gall. Skeleton motion words for unsupervised skeleton-based temporal action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12101–12111, 2025.

Dayoung Gong, Joonseok Lee, Deunsol Jung, Suha Kwak, and Minsu Cho. Activity grammars for temporal action segmentation. *Advances in Neural Information Processing Systems*, 36:75413–75433, 2023.

Rui Guo, Zheng Xie, Chencheng Zhang, and Xiaohua Qian. Causality-enhanced multiple instance learning with graph convolutional networks for parkinsonian freezing-of-gait assessment. *IEEE Transactions on Image Processing*, 33:3991–4001, 2024.

Pranay Gupta, Divyanshu Sharma, and Ravi Kiran Sarvadevabhatla. Syntactically guided generative embeddings for zero-shot skeleton action recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 439–443. IEEE, 2021.

Jeongseok Hyun, Su Ho Han, Hyolim Kang, Joon-Young Lee, and Seon Joo Kim. Exploring scalability of self-training for open-vocabulary temporal action localization. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9406–9415. IEEE, 2025.

Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2322–2331, 2021.

Haoyu Ji, Bowen Chen, Xinglong Xu, Weihong Ren, Zhiyong Wang, and Honghai Liu. Language-assisted skeleton action understanding for skeleton-based temporal action segmentation. In *European Conference on Computer Vision*, pp. 400–417. Springer, 2024.

Haoyu Ji, Bowen Chen, Weihong Ren, Wenze Huang, Zhihao Yang, Zhiyong Wang, and Honghai Liu. Text-derived relational graph-enhanced network for skeleton-based action segmentation. *IEEE Transactions on Image Processing*, 34:7305–7320, 2025.

Haoyu Ji, Bowen Chen, Zhihao Yang, Wenze Huang, Yu Gao, Xueting Liu, Weihong Ren, Zhiyong Wang, and Honghai Liu. Spectral scalpel: Amplifying adjacent action discrepancy via frequency-selective filtering for skeleton-based action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12849–12859, 2026a.

Haoyu Ji, Xueting Liu, Yu Gao, Wenze Huang, Zhihao Yang, Weihong Ren, Zhiyong Wang, and Honghai Liu. Lady: Lagrangian-dynamic informed network for skeleton-based action segmentation via spatial-temporal modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34449–34459, 2026b.

Han Jiang, Haoyu Tang, Xiaoxuan Mu, Chen Li, and Jihua Zhu. Memory matters: Boosting training-free zero-shot temporal action localization with a learnable lookup table. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9762–9772, 2026.

Jidong Kuang, Hongsong Wang, Chaolei Han, Yang Zhang, and Jie Gui. Zero-shot skeleton-based action recognition with dual visual-text alignment. *Pattern Recognition*, pp. 112342, 2025.

Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 156–165, 2017.

Yearang Lee, Ho-Joong Kim, and Seong-Whan Lee. Text-infused attention and foreground-aware modeling for zero-shot temporal action detection. *Advances in Neural Information Processing Systems*, 37:9864–9884, 2024.

Yearang Lee, Ho-Joong Kim, and Seong-Whan Lee. Tf-cade: Foreground-concentrated text-video alignment for zero-shot temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2843–2852, 2026.

Guozhang Li, Xinran Duan, Mei Wang, Lizhi Wang, and Hua Huang. Curvature-guided task synergy for skeleton based temporal action segmentation. In *International Conference on Learning Representations (ICLR)*, 2026.

Sheng-Wei Li, Zi-Xiang Wei, Wei-Jie Chen, Yi-Hsin Yu, Chih-Yuan Yang, and Jane Yung-jen Hsu. Sa-dvae: Improving zero-shot skeleton-based action recognition by disentangled variational autoencoders. In *European conference on computer vision*, pp. 447–462. Springer, 2024a.

Shijie Li, Yazan Abu Farha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6647–6658, 2020.

Zhiheng Li, Yujie Zhong, Ran Song, Tianjiao Li, Lin Ma, and Wei Zhang. Detal: Open-vocabulary temporal action localization with decoupled networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7728–7741, 2024b.

Benedetta Liberatori, Alessandro Conti, Paolo Rota, Yiming Wang, and Elisa Ricci. Test-time zero-shot temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18720–18729, 2024.

Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for skeleton-based human action understanding. In *Proceedings of the workshop on visual analysis in smart and connected communities*, pp. 1–8, 2017.

Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffusion action segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10139–10149, 2023.

Shenglan Liu, Aibin Zhang, Yunheng Li, Jian Zhou, Li Xu, Zhuben Dong, and Renhao Zhang. Temporal segmentation of fine-gained semantic action: A motion-centered figure skating dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 2163–2171, 2021.

Zijia Lu and Ehsan Elhamifar. Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18175–18185, 2024.

Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Zero-shot temporal action detection via vision-language prompting. In *European conference on computer vision*, pp. 681–697. Springer, 2022.

Thinh Phan, Khoa Vo, Duy Le, Gianfranco Doretto, Donald Adjeroh, and Ngan Le. Zeetad: Adapting pretrained vision-language model for zero-shot end-to-end temporal action detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 7046–7055, 2024.

Asif Raza, Bang Yang, and Yuexian Zou. Zero-shot temporal action detection by learning multimodal prompts and text-enhanced actionness. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11000–11012, 2024.

Shaofan Sun, Lilang Lin, Jiahang Zhang, Ling-Yu Duan, and Jiaying Liu. Local dimension enhancement representation learning for skeleton-based action segmentation. *IEEE Transactions on Image Processing*, 2026.

Onkar Kishor Susladkar, Gayatri Sudhir Deshmukh, Vandan Gorade, and Sparsh Mittal. Grizal: Generative prior-guided zero-shot temporal action localization. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 19046–19059, 2024.

Matteo Terreran, Leonardo Barcellona, and Stefano Ghidoni. A general skeleton-based action and gesture recognition framework for human–robot collaboration. *Robotics and Autonomous Systems*, 170:104523, 2023.

Haitao Tian. Duoclr: Dual-surrogate contrastive learning for skeleton-based human action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13772–13782, 2025.

Haitao Tian and Pierre Payeur. Stitch, contrast, and segment: Learning a human action segmentation model using trimmed skeleton videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7365–7373, 2025.

Ning Wang, Tieyue Wu, Naeha Sharif, Farid Boussaid, Guangming Zhu, Lin Mei, Mohammed Bennamoun, and Liang Zhang. Skeletoncontext: Skeleton-side context prompt learning for zero-shot skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20170–20180, 2026.

Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. Boundary-aware cascade networks for temporal action segmentation. In *European Conference on Computer Vision*, pp. 34–51. Springer, 2020.

Wenhan Wu, Zhishuai Guo, Chen Chen, Hongfei Xue, and Aidong Lu. Frequency-semantic enhanced variational autoencoder for zero-shot skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11122–11131, 2025.

Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13623–13633, 2023.

Di Yang, Yaohui Wang, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Lac-latent action composition for skeleton-based action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13679–13690, 2023.

Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Sen Wang, Zongyuan Ge, and Alexander Hauptmann. Zstad: Zero-shot temporal activity detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 879–888, 2020.

Yujie Zhou, Wenwen Qiang, Anyi Rao, Ning Lin, Bing Su, and Jiaqi Wang. Zero-shot skeleton-based action recognition via mutual information estimation and maximization. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 5302–5310, 2023.

Anqi Zhu, Qiuhong Ke, Mingming Gong, and James Bailey. Part-aware unified representation of language and skeleton for zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18761–18770, 2024.

Anqi Zhu, Jingmin Zhu, James Bailey, Mingming Gong, and Qiuhong Ke. Semantic-guided cross-modal prompt learning for skeleton-based zero-shot action recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13876–13885, 2025a.

Jingmin Zhu, Anqi Zhu, Hossein Rahmani, Jun Liu, Mohammed Bennamoun, and Qiuhong Ke. Boosting skeleton-based zero-shot action recognition with training-free test-time adaptation. *arXiv preprint arXiv:2512.11458*, 2025b.

Sa Zhu, Wanqian Zhang, Lin Wang, Xiaohua Chen, Chenxu Cui, Jinchao Zhang, and Bo Li. Decompose and transfer: Cot-prompting enhanced alignment for open-vocabulary temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20334–20344, 2026a.

Sa Zhu, Wanqian Zhang, Lin Wang, Jinchao Zhang, Cong Wang, and Bo Li. Denoise and align: Diffusion-driven foreground knowledge prompting for open-vocabulary temporal action detection. *arXiv preprint arXiv:2604.18313*, 2026b.

Appendix

A Zero-Shot Benchmarks and Evaluation Protocol **14**

B Complete Specification of Kinematic Anchor Grounding **15**

B.1 Body Partition and Base Quantities 15

B.2 Static and Dynamic Descriptors 15

B.3 Regression Targets 16

C Properties of the Kinematic Descriptors **17**

C.1 Invariances 17

C.2 Time Reversal 18

C.3 Structural Properties 19

C.4 Coordinate Noise 19

C.5 Why KAG Discourages Shortcuts 20

D Analysis of the Cross-Modal Atom Codebook **21**

D.1 Sparse Recipes and Residual Injection 21

D.2 Matching Score and Recipe Alignment 21

D.3 Capacity 22

E Class Descriptions **22**

F Implementation Details **23**

G Additional Quantitative Results **23**

G.1 Enlarging the Unseen Label Space 24

G.2 Hyperparameter Sensitivity 24

H Failure Cases and Limitations **24**

I Additional Visualizations **25**

I.1 Segmentation with Keyframes 25

I.2 Motion Atoms 26

A Zero-Shot Benchmarks and Evaluation Protocol

Unseen categories. Table 6 lists the unseen categories. The two PKU-MMD protocols share the same ten unseen categories, and the 26 unseen categories of MCFS-130 contain the 13 of U13, so the two MCFS splits differ only in how much of the label space is withheld.

Protocol. Training sequences keep the segments of seen categories and background and remove those of unseen categories; test sequences keep the segments of unseen categories and background and remove those of seen categories. Within a source video, the retained segments are concatenated

Table 6: Unseen categories of the four zero-shot splits. MCFS-130 U26 consists of the 13 categories of U13 together with the 13 categories listed in the last column.

PKU-MMD, X-sub and X-view (U10)	MCFS-130 U13	MCFS-130 U26, additional categories
hopping(one_foot_jumping)	StepSequence2	3Lutz_3Toeloop
jump_up	FlySitSpin4	StepSequence4
pickup	3Flip_3Toeloop	3Salchow
take_off_glasses	2Axel_3Toeloop	ChComboSpin3
throw	3Lutz_3Loop	3Lutz_2Toeloop
touch_head(headache)	ChSitSpin3	3Axel_2Toeloop
touch_neck(neckache)	4Toeloop_3Toeloop	2Axel_2Toeloop
typing_on_a_keyboard	3Flip_Euler_3Salchow	3Salchow_3Toeloop
wear_jacket	1Axel	ChCamelSpin3
writing	4Toeloop_2Toeloop	3Flip_2Toeloop_2Loop
	FlyCamelSpin2	4Flip
	2Toeloop	4Lutz_3Toeloop
	2Axel_2Toeloop_2Loop	FlyChComboSpin3

in their original order and joined by four-frame transitions following Tian & Payeur (2025); the transitions carry an ignore label and are excluded from the loss and from the evaluation. Training thus never observes a frame or a description of an unseen category. Table 7 reports the resulting benchmark sizes. We report Acc, Edit, F1@{10, 25, 50}, and AvgF1, the mean of the three F1 scores; background segments are not counted in Edit and F1.

Table 7: Size of the four zero-shot benchmarks. Frames exclude the ignore-labeled transitions and include background; segments are the maximal runs of a category label and exclude background.

	Training (seen)			Test (unseen)		
Setting	sequences	frames	segments	sequences	frames	segments
PKU-MMD X-sub	696	407,901	3,541	197	57,910	393
PKU-MMD X-view	602	347,315	3,086	285	78,513	566
MCFS-130 U13	151	871,891	1,390	57	63,863	94
MCFS-130 U26	151	725,428	1,170	79	145,449	215

B Complete Specification of Kinematic Anchor Grounding

B.1 Body Partition and Base Quantities

We group the joints into $P = 5$ disjoint parts: the torso b_{tor}, the arms $b^{\mathrm{l}}_{\mathrm{arm}}, b^{\mathrm{r}}_{\mathrm{arm}}$, and the legs $b^{\mathrm{l}}_{\mathrm{leg}}, b^{\mathrm{r}}_{\mathrm{leg}}$ (Figure 6), and write $b_g = b^{\mathrm{l}}_g \cup b^{\mathrm{r}}_g$ for a limb group $g \in \{\mathrm{arm}, \mathrm{leg}\}$. On MCFS-130 the four facial keypoints are not assigned to any part; whole-body quantities always use the full joint set $\mathcal{J}$. Coordinates are 3D for PKU-MMD and, for MCFS-130, 2D keypoints normalized by the frame size and centered at the root. KAG uses the motion energy E of Eq. (1) and the pose extent

$$\Pi(b, \mathcal{T}) = \frac{1}{|b||\mathcal{T}|} \sum_{j \in b} \sum_{t \in \mathcal{T}} \left\| \mathbf{x}_{j,t} - \mathbf{x}_{\mathrm{root},t} \right\|_2, \tag{12}$$

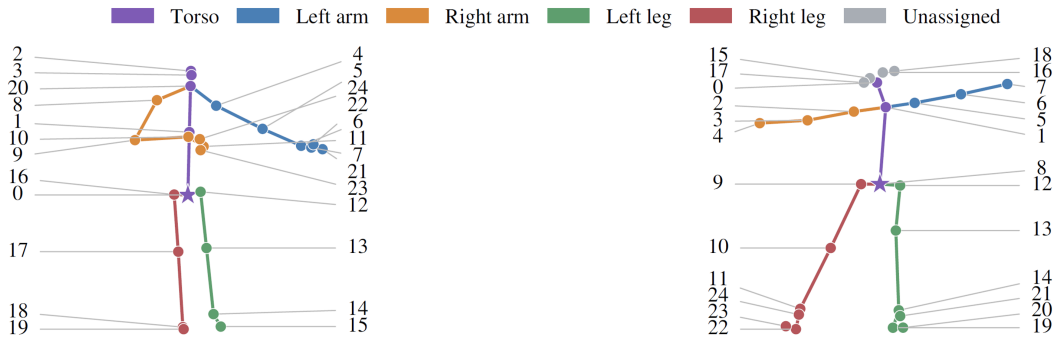
and on segment i we write

$$\begin{aligned} &E_p = E(b_p, \mathcal{T}_i), \quad \Pi_p = \Pi(b_p, \mathcal{T}_i), \quad E_* = E(\mathcal{J}, \mathcal{T}_i), \quad E_{\mathrm{tor}} = E(b_{\mathrm{tor}}, \mathcal{T}_i), \\ &E^{\mathrm{l}}_g = E(b^{\mathrm{l}}_g, \mathcal{T}_i), \quad E^{\mathrm{r}}_g = E(b^{\mathrm{r}}_g, \mathcal{T}_i), \quad E_g = \tfrac{1}{2}\big(E^{\mathrm{l}}_g + E^{\mathrm{r}}_g\big), \qquad g \in \{\mathrm{arm}, \mathrm{leg}\}. \end{aligned} \tag{13}$$

B.2 Static and Dynamic Descriptors

The static descriptor of Eq. (2) is $\mathbf{a}_i = [E_1, \ldots, E_P, \boldsymbol{\rho}_i]$ with

$$\begin{aligned} \boldsymbol{\rho}_i = \Big[&\frac{E_{\mathrm{arm}}}{E_*}, \frac{E_{\mathrm{leg}}}{E_*}, \frac{E_{\mathrm{tor}}}{E_*}, \frac{|E^{\mathrm{l}}_{\mathrm{arm}} - E^{\mathrm{r}}_{\mathrm{arm}}|}{E_{\mathrm{arm}}}, \frac{|E^{\mathrm{l}}_{\mathrm{leg}} - E^{\mathrm{r}}_{\mathrm{leg}}|}{E_{\mathrm{leg}}}, \\ &\frac{E_{\mathrm{arm}} + E_{\mathrm{leg}}}{2E_{\mathrm{tor}}}, \zeta_i, \frac{1}{P} \sum_{p=1}^{P} \frac{\Pi_p}{1 + E_*} \Big], \end{aligned} \tag{14}$$



(a) Kinect-25 (PKU-MMD) (b) BODY_25 (MCFS-130)

Figure 6: Body partition on the 3D Kinect v2 skeleton of PKU-MMD (left) and on the 2D OpenPose BODY_25 skeleton of MCFS-130 (right). Unassigned keypoints are drawn in gray and the root joint is marked with a star. The same part colors are used in all appendix figures that refer to body parts.

where ζ_i compares the whole-body energy of the first and the last $\lfloor |\mathcal{T}_i|/3 \rfloor$ frames, $\mathcal{T}_i^{\text{early}}$ and $\mathcal{T}_i^{\text{late}}$,

$$\zeta_i = \frac{E_{\text{late}} - E_{\text{early}}}{E_{\text{late}} + E_{\text{early}}}, \qquad E_{\text{early}} = E\big(\mathcal{J}, \mathcal{T}_i^{\text{early}}\big), \quad E_{\text{late}} = E\big(\mathcal{J}, \mathcal{T}_i^{\text{late}}\big). \tag{15}$$

Besides the part energies, $\mathbf{a}_i$ thus contains energy shares, left–right asymmetries, the limb-to-torso activity, the intensity change, and the pose extent; with $P = 5$, $\mathbf{a}_i \in \mathbb{R}^{13}$.

For the dynamic descriptor, each segment is split uniformly into M phases with whole-body energy $E_*^{(m)}$ and part shares $\pi_p^{(m)} = E(b_p, \mathcal{T}_i^{(m)})/E_*^{(m)}$; for a limb group, $\pi_g^{\text{l},(m)}$ and $\pi_g^{\text{r},(m)}$ are the shares of its two sides and $\pi_g^{(m)}$ their mean. With the arm–leg dominance and the left–right asymmetry of a phase,

$$\varrho^{(m)} = \frac{\pi_{\text{arm}}^{(m)} - \pi_{\text{leg}}^{(m)}}{\pi_{\text{arm}}^{(m)} + \pi_{\text{leg}}^{(m)}}, \qquad \varsigma^{(m)} = \frac{1}{2} \sum_{g \in \{\text{arm},\text{leg}\}} \frac{\big|\pi_g^{\text{l},(m)} - \pi_g^{\text{r},(m)}\big|}{\pi_g^{(m)}}, \tag{16}$$

the dynamic descriptor of Eq. (3) reads

$$\begin{aligned} \mathbf{a}_i^{\Delta} &= \Big[\big\{ \pi_p^{(M)} - \pi_p^{(1)} \big\}_{p=1}^{P}, \ \Big\{ \pi_p^{(m)} - \tfrac{1}{2}\big(\pi_p^{(1)} + \pi_p^{(M)}\big) \Big\}_{p=1,\dots,P;\ 1<m<M}, \ \boldsymbol{\eta}_i \Big], \\ \boldsymbol{\eta}_i &= \Big[\frac{E_*^{(M)} - E_*^{(1)}}{E_*^{(M)} + E_*^{(1)}}, \ \varrho^{(M)} - \varrho^{(1)}, \ \varsigma^{(M)} - \varsigma^{(1)} \Big]. \end{aligned} \tag{17}$$

Its three groups are the trend and the curvature of the shares and the changes of intensity, dominance, and asymmetry; with $P = 5$ and $M = 3$, matching the three phase descriptions on the text side, $\mathbf{a}_i^{\Delta} \in \mathbb{R}^{13}$.

B.3 Regression Targets

Both descriptors are mapped to targets by $\phi(\mathbf{a}) = \mathbf{u}/\|\mathbf{u}\|_2$ with $\mathbf{u} = \operatorname{sign}(\mathbf{a}) \odot \log(1 + |\mathbf{a}|)$, which compresses the unbounded energies while preserving the sign of the dynamic components. Denominators are guarded by $\delta = 10^{-6}$, and KAG is applied to seen segments of at least 12 frames. The heads g_a and g_a^{Δ} are two-layer MLPs with a hidden width of 128, shared across the refinement stages, and their outputs are ℓ_2-normalized before the Smooth-L1 loss. Table 8 summarizes all components and their invariances, which Appendix C proves, and Algorithm 1 gives the complete procedure.

Table 8: Components of the kinematic descriptors ($P = 5$, $M = 3$). Rig.: invariance to a rigid motion $\mathbf{x} \mapsto \mathbf{R}\mathbf{x} + \mathbf{t}$ of the whole sequence in the coordinate space of the dataset (3D for PKU-MMD, normalized 2D for MCFS-130). Scale: invariance to $\mathbf{x} \mapsto s\mathbf{x}$. Mirror: behavior when the action is performed with the other side of the body; l↔r means that the left and right entries are exchanged. Rev.: parity under time reversal of the segment, up to a single boundary displacement; odd components change sign. $\omega_p = |b_p|/|\mathcal{J}|$.

Component	Dim.	Meaning	Range	Rig.	Scale	Mirror	Rev.
Static descriptor $\mathbf{a}_i$							
E_p	P	part energy	$[0, \infty)$	✓	✗ ($\times s$)	l↔r	even
E_g/E_*	3	energy share (arm/leg/tor)	$[0, \lvert\mathcal{J}\rvert/\lvert b_g\rvert]$	✓	✓	✓	even
$\lvert E_g^{\mathrm{l}} - E_g^{\mathrm{r}}\rvert/E_g$	2	L–R asymmetry (arm, leg)	$[0, 2]$	✓	✓	✓	even
$(E_{\mathrm{arm}} + E_{\mathrm{leg}})/(2E_{\mathrm{tor}})$	1	limb/torso activity	$[0, \infty)$	✓	✓	✓	even
ζ_i in Eq. (15)	1	intensity change (thirds)	$[-1, 1]$	✓	✓	✓	odd
$\frac{1}{P}\sum_p \Pi_p/(1+E_*)$	1	pose extent	$[0, \infty)$	✓	✗	✓	even
Dynamic descriptor $\mathbf{a}_i^{\Delta}$							
$\pi_p^{(M)} - \pi_p^{(1)}$	P	share trend	$[-\omega_p^{-1}, \omega_p^{-1}]$	✓	✓	l↔r	odd
$\pi_p^{(m)} - \frac{1}{2}(\pi_p^{(1)}+\pi_p^{(M)})$	$P(M-2)$	share curvature	$[-\omega_p^{-1}, \omega_p^{-1}]$	✓	✓	l↔r	even
$\eta_{i,1}$ in Eq. (17)	1	intensity change (phases)	$[-1, 1]$	✓	✓	✓	odd
$\varrho^{(M)} - \varrho^{(1)}$	1	dominance change	$[-2, 2]$	✓	✓	✓	odd
$\varsigma^{(M)} - \varsigma^{(1)}$	1	asymmetry change	$[-2, 2]$	✓	✓	✓	odd

Algorithm 1 Kinematic descriptors of a seen segment

Require: joint coordinates $\{\mathbf{x}_{j,t}\}$; segment $\mathcal{T}_i$ with $|\mathcal{T}_i| \geq 12$; parts $\{b_p\}_{p=1}^{P}$; $M = 3$; $\delta = 10^{-6}$
Ensure: regression targets $\phi(\mathbf{a}_i)$ and $\phi(\mathbf{a}_i^{\Delta})$

$e_{j,t} \leftarrow \|\mathbf{x}_{j,t} - \mathbf{x}_{j,t-1}\|_2$ and $d_{j,t} \leftarrow \|\mathbf{x}_{j,t} - \mathbf{x}_{\mathrm{root},t}\|_2$ for all $j \in \mathcal{J}$, $t \in \mathcal{T}_i$ ▷ speed, extent
$E(b, \mathcal{T}) \leftarrow$ mean of $e_{j,t}$ over $j \in b$, $t \in \mathcal{T}$; $\Pi(b, \mathcal{T}) \leftarrow$ mean of $d_{j,t}$ over $j \in b$, $t \in \mathcal{T}$
compute E_p, Π_p, E_*, E_{tor}, E_g^{l}, E_g^{r}, E_g, and ζ_i on $\mathcal{T}_i$ ▷ Eqs. (13), (15)
$\mathbf{a}_i \leftarrow [E_1, \ldots, E_P, \boldsymbol{\rho}_i]$ ▷ Eq. (14)
split $\mathcal{T}_i$ uniformly into phases $\mathcal{T}_i^{(1)}, \ldots, \mathcal{T}_i^{(M)}$
for $m = 1, \ldots, M$ **do**
    $E_*^{(m)} \leftarrow E(\mathcal{J}, \mathcal{T}_i^{(m)})$; $\pi_p^{(m)} \leftarrow E(b_p, \mathcal{T}_i^{(m)})/\max(E_*^{(m)}, \delta)$ for $p = 1, \ldots, P$
    compute $\varrho^{(m)}$ and $\varsigma^{(m)}$ ▷ Eq. (16)
end for
assemble $\mathbf{a}_i^{\Delta}$ from trends, curvatures, and $\boldsymbol{\eta}_i$ ▷ Eq. (17)
return $\phi(\mathbf{a}_i)$ and $\phi(\mathbf{a}_i^{\Delta})$, where $\phi(\mathbf{a}) = \mathbf{u}/\|\mathbf{u}\|_2$, $\mathbf{u} = \mathrm{sign}(\mathbf{a}) \odot \log(1 + |\mathbf{a}|)$

C Properties of the Kinematic Descriptors

We prove the properties of Table 8, ignoring the numerical guards and boundary effects, which concern single frames only. We write $\mathbf{z}_i = \phi(\mathbf{a}_i)$ and $\mathbf{z}_i^{\Delta} = \phi(\mathbf{a}_i^{\Delta})$ for the targets, both unit vectors in $\mathbb{R}^{13}$.

C.1 Invariances

Proposition C.1 (Rigid motion). *If $\mathbf{x}'_{j,t} = \mathbf{R}\mathbf{x}_{j,t} + \mathbf{o}$ for all joints and frames, with $\mathbf{R}$ orthogonal, then $E'(b, \mathcal{T}) = E(b, \mathcal{T})$ and $\Pi'(b, \mathcal{T}) = \Pi(b, \mathcal{T})$ for every joint set b and frame set $\mathcal{T}$; hence $\mathbf{a}'_i = \mathbf{a}_i$ and $\mathbf{a}_i^{\Delta\prime} = \mathbf{a}_i^{\Delta}$.*

Proof. $\mathbf{x}'_{j,t} - \mathbf{x}'_{j,t-1} = \mathbf{R}(\mathbf{x}_{j,t} - \mathbf{x}_{j,t-1})$ and $\mathbf{x}'_{j,t} - \mathbf{x}'_{\mathrm{root},t} = \mathbf{R}(\mathbf{x}_{j,t} - \mathbf{x}_{\mathrm{root},t})$, and $\mathbf{R}$ preserves norms. All components are functions of E and Π on fixed joint and frame sets. □

For the 3D skeletons of PKU-MMD, a change of camera viewpoint is a rigid motion up to sensor noise and occlusion; for MCFS-130, the proposition covers rotations and translations of the normalized 2D coordinates.

Proposition C.2 (Scaling). *If $\mathbf{x}'_{j,t} = s\,\mathbf{x}_{j,t}$ with $s > 0$, then $E' = sE$ and $\Pi' = s\Pi$. The part energies are multiplied by s, the pose-extent term becomes $\frac{1}{P}\sum_p s\Pi_p/(1+sE_*)$, and all other components of $\mathbf{a}_i$ and $\mathbf{a}_i^\Delta$ are invariant.*

Proof. E and Π are averages of norms and thus positively homogeneous of degree one. All other components are ratios of degree-one quantities, or differences of such ratios, and thus of degree zero. □

Hence 20 of the 26 components are invariant to a global scaling, which models differences in body size to first order.

Proposition C.3 (Mirroring). *Let $\iota : \mathcal{J} \to \mathcal{J}$ exchange corresponding left and right joints and fix the midline joints, and let $\mathbf{Q}$ be a reflection. For the mirrored performance $\mathbf{x}'_{j,t} = \mathbf{Q}\mathbf{x}_{\iota(j),t}$, $\boldsymbol{\rho}_i$ and $\boldsymbol{\eta}_i$ are invariant, and the left and right entries of the part energies, trends, and curvatures are exchanged.*

Proof. As $\mathbf{Q}$ preserves norms and the root lies on the midline, $E'(b,\mathcal{T}) = E(\iota(b),\mathcal{T})$ and $\Pi'(b,\mathcal{T}) = \Pi(\iota(b),\mathcal{T})$. Since ι maps $\mathcal{J}$ and b_{tor} onto themselves and exchanges b_g^{l} and b_g^{r}, the quantities E_*, E_{tor}, E_g, $\sum_p \Pi_p$, E_{early}, E_{late}, $\pi_g^{(m)}$, $|E_g^{\mathrm{l}} - E_g^{\mathrm{r}}|$, and $|\pi_g^{\mathrm{l},(m)} - \pi_g^{\mathrm{r},(m)}|$ are unchanged; they determine $\boldsymbol{\rho}_i$ and $\boldsymbol{\eta}_i$. □

C.2 Time Reversal

Proposition C.4 (Time reversal). *Let $\mathbf{x}'_{j,t} = \mathbf{x}_{j,\,t_i^s + t_i^e - t}$ for $t \in \mathcal{T}_i$ and $M = 3$. Up to one boundary displacement and the rounding of the phase boundaries,*

$$\mathbf{a}'_i = \mathbf{S}\,\mathbf{a}_i, \quad \mathbf{S} = \mathrm{diag}(\mathbf{I}_{11}, -1, 1), \qquad \mathbf{a}_i^{\Delta\prime} = \mathbf{S}^\Delta \mathbf{a}_i^\Delta, \quad \mathbf{S}^\Delta = \mathrm{diag}(-\mathbf{I}_P,\, \mathbf{I}_P,\, -\mathbf{I}_3), \tag{18}$$

and likewise $\mathbf{z}'_i = \mathbf{S}\mathbf{z}_i$ and $\mathbf{z}_i^{\Delta\prime} = \mathbf{S}^\Delta \mathbf{z}_i^\Delta$.

Proof. Reversal negates every displacement and maps the segment onto itself, so all components of $\mathbf{a}_i$ computed on the whole segment are unchanged, while $\mathcal{T}_i^{\mathrm{early}}$ and $\mathcal{T}_i^{\mathrm{late}}$ are exchanged and ζ_i changes sign. Phase m is mapped to phase $M+1-m$, so the trends and $\boldsymbol{\eta}_i$ change sign, while for $M = 3$ the curvature $\pi_p^{(2)} - \frac{1}{2}(\pi_p^{(1)} + \pi_p^{(3)})$ is unchanged. Finally, $\mathrm{sign}(\mathbf{S}\mathbf{a}) = \mathbf{S}\,\mathrm{sign}(\mathbf{a})$, $|\mathbf{S}\mathbf{a}| = |\mathbf{a}|$, and $\|\mathbf{S}\mathbf{u}\|_2 = \|\mathbf{u}\|_2$ for any diagonal $\mathbf{S}$ with entries ± 1, so $\phi(\mathbf{S}\mathbf{a}) = \mathbf{S}\phi(\mathbf{a})$. □

Corollary C.5 (Separation of time-reversed actions). *If action B is approximately action A played backwards, then*

$$\|\mathbf{z}_A - \mathbf{z}_B\|_2 \approx 2\,|z_{A,12}|, \qquad \|\mathbf{z}_A^\Delta - \mathbf{z}_B^\Delta\|_2 \approx 2\,\|\mathbf{z}_{A,\mathrm{odd}}^\Delta\|_2, \tag{19}$$

where $z_{A,12}$ is the entry of ζ and $\mathbf{z}_{A,\mathrm{odd}}^\Delta$ collects the eight sign-changing entries of $\mathbf{z}_A^\Delta$.

Proof. $\mathbf{z}_A - \mathbf{z}_B \approx (\mathbf{I} - \mathbf{S})\mathbf{z}_A$, and $\mathbf{I} - \mathbf{S}$ is twice the projection onto the sign-changing entries; likewise for $\mathbf{S}^\Delta$. □

Figures 7 and 8 confirm this on four such pairs of PKU-MMD: the static class-mean targets of the two actions are only 0.04 to 0.12 apart and nearly coincide, whereas the dynamic ones are 0.69 to 1.35 apart, 11 to 30 times farther.

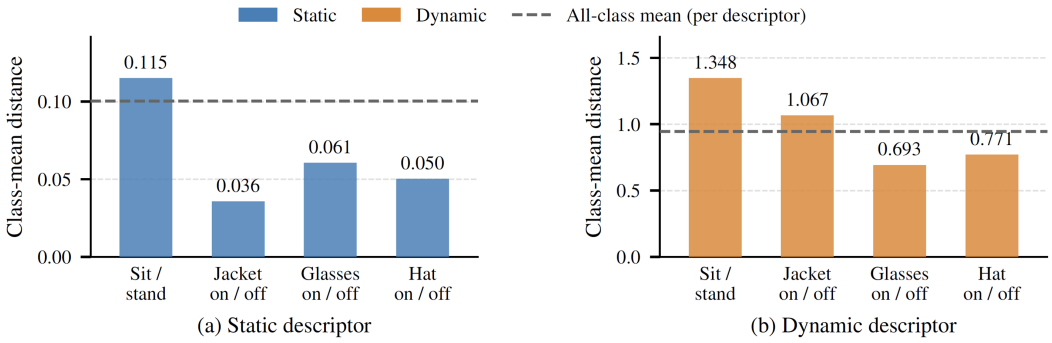


Figure 7: Distance between the class-mean targets of four time-reversed category pairs of PKU-MMD for the static (a) and the dynamic (b) descriptor. A class-mean target is the normalized average of the targets of all segments of the category. Dashed lines: average distance over all pairs of categories, computed separately for each descriptor.

C.3 Structural Properties

Lemma C.6 (Energy shares). *Let $\omega_p = |b_p|/|\mathcal{J}|$ and $\omega_0 = |\mathcal{J}_0|/|\mathcal{J}|$ with $\mathcal{J}_0 = \mathcal{J} \setminus \bigcup_p b_p$ ($\omega_0 = 0$ on PKU-MMD, $4/25$ on MCFS-130). For every frame set $\mathcal{T}$,*

$$E(\mathcal{J}, \mathcal{T}) = \sum_{p=1}^{P} \omega_p \, E(b_p, \mathcal{T}) + \omega_0 \, E(\mathcal{J}_0, \mathcal{T}), \tag{20}$$

so $0 \le \pi_p^{(m)} \le \omega_p^{-1}$ and $\sum_p \omega_p \pi_p^{(m)} = 1 - \omega_0 \pi_0^{(m)} \le 1$, where $\pi_0^{(m)}$ is the share of $\mathcal{J}_0$. On PKU-MMD, the trends and, for $M = 3$, the curvatures therefore have zero ω-weighted sum.

Proof. $E(b, \mathcal{T})$ averages over the joints of b, so the average over $\mathcal{J}$ is the size-weighted average over the disjoint subsets $b_1, \dots, b_P, \mathcal{J}_0$. Dividing by $E_*^{(m)}$ gives the share identity, and linearity gives the constraints. □

Lemma C.7 (Difference and log-ratio forms). *For $M = 3$, the trend and the curvature are the first and second differences of the shares,*

$$\pi_p^{(3)} - \pi_p^{(1)} = \Delta\pi_p^{(1)} + \Delta\pi_p^{(2)}, \qquad \pi_p^{(2)} - \tfrac{1}{2}\big(\pi_p^{(1)} + \pi_p^{(3)}\big) = -\tfrac{1}{2}\,\Delta^2\pi_p, \tag{21}$$

with $\Delta\pi_p^{(m)} = \pi_p^{(m+1)} - \pi_p^{(m)}$ and $\Delta^2\pi_p = \pi_p^{(1)} - 2\pi_p^{(2)} + \pi_p^{(3)}$. Moreover, $\frac{a-b}{a+b} = \tanh\big(\frac{1}{2}\log\frac{a}{b}\big)$ for $a, b > 0$, so ζ_i and $\eta_{i,1}$ are bounded log-ratios of energies.

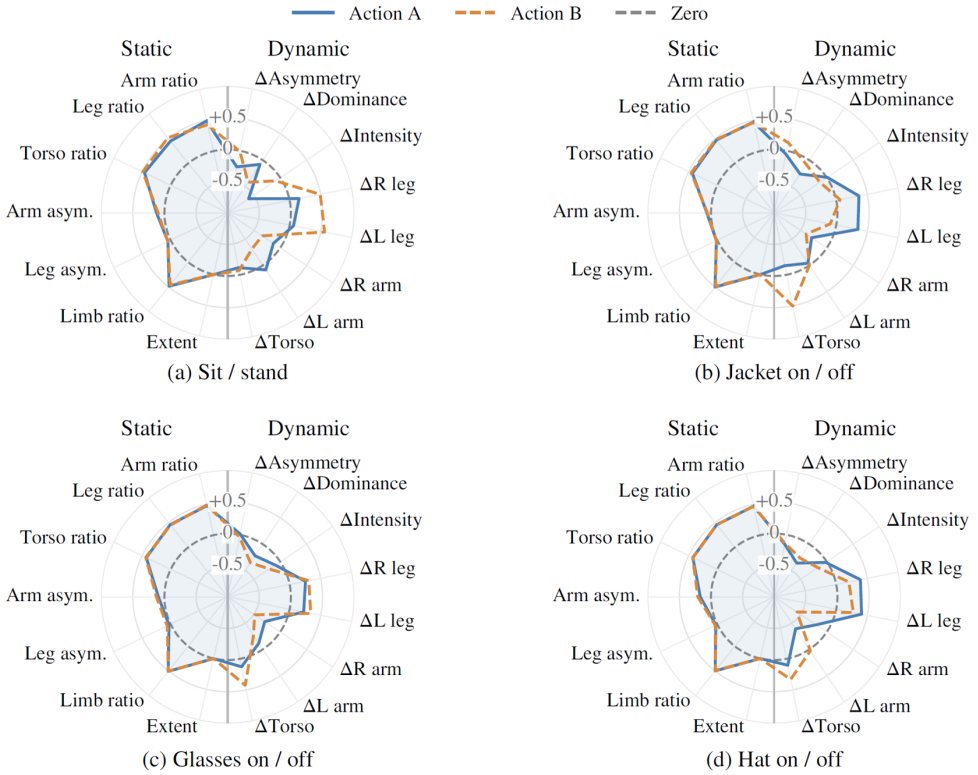
Proof. Expansion, and $\tanh x = (e^{2x} - 1)/(e^{2x} + 1)$ with $e^{2x} = a/b$. □

C.4 Coordinate Noise

Proposition C.8 (Coordinate noise). *Let $\tilde{\mathbf{x}}_{j,t} = \mathbf{x}_{j,t} + \boldsymbol{\xi}_{j,t}$ with independent $\boldsymbol{\xi}_{j,t} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_c)$, and $\mathbf{d} = \mathbf{x}_{j,t} - \mathbf{x}_{j,t-1}$. Then $\mathbb{E}\|\mathbf{d} + \boldsymbol{\xi}_{j,t} - \boldsymbol{\xi}_{j,t-1}\|_2 \ge \|\mathbf{d}\|_2$, and*

$$\begin{aligned} \mathbb{E}\|\boldsymbol{\xi}_{j,t} - \boldsymbol{\xi}_{j,t-1}\|_2 &= \frac{2\sigma\,\Gamma\big(\frac{c+1}{2}\big)}{\Gamma\big(\frac{c}{2}\big)} && \text{if } \mathbf{d} = \mathbf{0}, \\ \mathbb{E}\|\mathbf{d} + \boldsymbol{\xi}_{j,t} - \boldsymbol{\xi}_{j,t-1}\|_2 &= \|\mathbf{d}\|_2 + \frac{(c-1)\sigma^2}{\|\mathbf{d}\|_2} + O\Big(\frac{\sigma^4}{\|\mathbf{d}\|_2^3}\Big) && \text{if } \|\mathbf{d}\|_2 \gg \sigma, \end{aligned} \tag{22}$$

where the first value is $4\sigma/\sqrt{\pi} \approx 2.26\sigma$ for $c = 3$ and $\sqrt{\pi}\,\sigma \approx 1.77\sigma$ for $c = 2$.



(a) Sit / stand (b) Jacket on / off

(c) Glasses on / off (d) Hat on / off

Figure 8: Class-mean targets of the time-reversed pairs of Figure 7. In each panel, the left half shows the six time-symmetric ratio components and the pose-extent term of the static target and the right half the eight components of the dynamic target that change sign under time reversal (Proposition C.4). Action A (solid) is the first action named in each panel, e.g., sitting down, and action B (dashed) the second; Δ denotes the change from the first to the last phase, and the gray dashed circle marks zero. The static profiles of the two actions nearly coincide, whereas their dynamic profiles differ.

Proof. $\boldsymbol{\xi}_{j,t} - \boldsymbol{\xi}_{j,t-1} \sim \mathcal{N}(\mathbf{0}, 2\sigma^2\mathbf{I}_c)$ has zero mean, so the lower bound is Jensen's inequality. For $\mathbf{d} = \mathbf{0}$, its norm follows a chi distribution with c degrees of freedom scaled by $\sqrt{2}\sigma$. For $\|\mathbf{d}\|_2 \gg \sigma$, a second-order expansion of the norm around $\mathbf{d}$ gives the bias $(c-1)\,2\sigma^2/(2\|\mathbf{d}\|_2)$. □

Noise thus adds a floor to the energy of slow joints but only a vanishing bias to fast joints.

C.5 Why KAG Discourages Shortcuts

Let $\hat{g} = g/\|g\|_2$ be the normalized output of a head and ℓ the Smooth-L1 loss averaged over coordinates; the following holds for either descriptor.

Proposition C.9 (Class-collapse lower bound)**.** *If the segment features depend only on the label, $\bar{\mathbf{f}}_i = h(y_i)$, then for any head g*

$$\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \ell\big(\hat{g}(\bar{\mathbf{f}}_i), \mathbf{z}_i\big) \;\geq\; B_\ell := \sum_{c\in\mathcal{C}^s} \frac{n_c}{|\mathcal{S}|} \min_{\|\mathbf{u}\|_2=1} \frac{1}{n_c} \sum_{i:\, y_i=c} \ell(\mathbf{u}, \mathbf{z}_i), \tag{23}$$

where n_c is the number of segments of class c. For the squared loss, $B = 2 - 2\sum_c \frac{n_c}{|\mathcal{S}|}\bar{R}_c$ with the mean resultant length $\bar{R}_c = \|\bar{\mathbf{z}}_c\|_2$ of the class-mean target $\bar{\mathbf{z}}_c = \frac{1}{n_c}\sum_{i:\, y_i=c} \mathbf{z}_i$.

Proof. $\hat{g}(\bar{\mathbf{f}}_i)$ is one unit vector $\mathbf{u}_c$ per class, so the loss is bounded by its per-class minimum over $\mathbf{u}_c$. For unit vectors, $\|\mathbf{u}-\mathbf{z}\|_2^2 = 2 - 2\mathbf{u}^\top\mathbf{z}$, whose class average $2 - 2\mathbf{u}^\top\bar{\mathbf{z}}_c$ is minimized by $\mathbf{u} = \bar{\mathbf{z}}_c/\|\bar{\mathbf{z}}_c\|_2$. □

Features that merely separate the seen categories, which cross-modal alignment alone can produce, thus cannot reduce $\mathcal{L}_{\mathrm{KAG}}$ below the within-class dispersion of the targets; doing so requires features that encode how each segment moves. The Shuffled targets of Table 4 carry no information about their segment and therefore cannot provide this signal.

Remark C.10 (Information-theoretic view). Smooth-L1 with threshold β_{sm} is quadratic below β_{sm}. As prediction and target are unit vectors, all coordinate residuals are below β_{sm} once $\cos(\hat{g},\mathbf{z}) > 1 - \beta_{\mathrm{sm}}^2/2$, and then $\ell(\hat{g},\mathbf{z}) = \big(1-\cos(\hat{g},\mathbf{z})\big)/(13\,\beta_{\mathrm{sm}})$. With the von Mises–Fisher density $q(\mathbf{z}\mid\bar{\mathbf{f}}) = C(\kappa)\exp(\kappa\,\hat{g}(\bar{\mathbf{f}})^\top\mathbf{z})$ on the unit sphere, whose normalizer does not depend on $\hat{g}$, the bound of Barber & Agakov (2003) gives for every $\kappa > 0$

$$I(\bar{\mathbf{f}};\mathbf{z}) \;\geq\; H(\mathbf{z}) + \mathbb{E}\log q(\mathbf{z}\mid\bar{\mathbf{f}}) = H(\mathbf{z}) + \log C(\kappa) + \kappa - \kappa\,\mathbb{E}\big[1-\cos(\hat{g}(\bar{\mathbf{f}}),\mathbf{z})\big]. \tag{24}$$

Since $H(\mathbf{z})$ is fixed by the data, minimizing $\mathcal{L}_{\mathrm{KAG}}$ in this regime maximizes a lower bound on the mutual information between the features and the kinematic targets.

D Analysis of the Cross-Modal Atom Codebook

We analyze the codebook of Section 3.3, dropping the frame index where no confusion arises.

D.1 Sparse Recipes and Residual Injection

Lemma D.1 (Sparse recipes). *The recipe $\mathbf{w}^{\mathrm{V}}$ lies on the probability simplex with exactly K' nonzero entries, so $\mathbf{r}^{\mathrm{V}} = \sum_k w_k^{\mathrm{V}}\bar{\mathbf{D}}_k$ lies in the convex hull of the selected atoms and $\|\mathbf{r}^{\mathrm{V}}\|_2 \leq 1$; the same holds on the text side. Where the selected set Ω is locally constant, $\partial w_k^{\mathrm{V}}/\partial\alpha_l = w_k^{\mathrm{V}}(\mathbb{I}[k=l] - w_l^{\mathrm{V}})$ for $k,l\in\Omega$ and 0 otherwise, so an atom outside Ω receives no gradient from the frame.*

Proof. Eq. (5) is a softmax restricted to Ω, and $\|\mathbf{r}^{\mathrm{V}}\|_2 \leq \sum_k w_k^{\mathrm{V}}\|\bar{\mathbf{D}}_k\|_2 = 1$. An atom $k\notin\Omega$ enters neither the normalization of Eq. (5) nor $\mathbf{r}^{\mathrm{V}}$. □

Two recipes therefore overlap only on shared atoms, $\langle\mathbf{w}^{\mathrm{V}},\mathbf{w}^{\mathrm{T}}\rangle = \sum_{k\in\Omega^{\mathrm{V}}\cap\Omega^{\mathrm{T}}} w_k^{\mathrm{V}}w_k^{\mathrm{T}}$.

Proposition D.2 (Bounded angular deviation). *If $\gamma\|\mathbf{r}\|_2 < \|\mathbf{f}\|_2$, then $\angle(\mathbf{f}+\gamma\mathbf{r},\mathbf{f}) \leq \arcsin(\gamma\|\mathbf{r}\|_2/\|\mathbf{f}\|_2)$. Consequently, if $\gamma < \min(\|\mathbf{f}\|_2,\|\mathbf{p}\|_2)$, then $\big|\angle(\tilde{\mathbf{f}},\tilde{\mathbf{p}}) - \angle(\mathbf{f},\mathbf{p})\big| \leq \arcsin(\gamma/\|\mathbf{f}\|_2) + \arcsin(\gamma/\|\mathbf{p}\|_2)$.*

Proof. $\mathbf{f}+\gamma\mathbf{r}$ lies in the ball of radius $\gamma\|\mathbf{r}\|_2$ around $\mathbf{f}$, whose largest angle to $\mathbf{f}$ is attained on a tangent from the origin. The second claim follows from $\|\mathbf{r}\|_2 \leq 1$ (Lemma D.1) and the triangle inequality for angles. □

The residual strength thus bounds how far atom injection can rotate frames and prototypes, so Eq. (6) refines the aligned cosine geometry rather than replacing it.

D.2 Matching Score and Recipe Alignment

Proposition D.3 (Score decomposition). *Let $\bar{\mathbf{G}} = \bar{\mathbf{D}}\bar{\mathbf{D}}^\top$ be the Gram matrix of the normalized atoms and $\mu = \max_{k\neq l}|\bar{\mathbf{D}}_k^\top\bar{\mathbf{D}}_l|$ its coherence. Then*

$$\begin{aligned}\tilde{\mathbf{f}}^\top\tilde{\mathbf{p}} &= \mathbf{f}^\top\mathbf{p} + \gamma\big(\mathbf{f}^\top\mathbf{r}^{\mathrm{T}} + \mathbf{r}^{\mathrm{V}\top}\mathbf{p}\big) + \gamma^2\,\mathbf{w}^{\mathrm{V}\top}\bar{\mathbf{G}}\,\mathbf{w}^{\mathrm{T}},\\ \big|\mathbf{w}^{\mathrm{V}\top}\bar{\mathbf{G}}\,\mathbf{w}^{\mathrm{T}} - \langle\mathbf{w}^{\mathrm{V}},\mathbf{w}^{\mathrm{T}}\rangle\big| &\leq \mu\big(1-\langle\mathbf{w}^{\mathrm{V}},\mathbf{w}^{\mathrm{T}}\rangle\big).\end{aligned} \tag{25}$$

Proof. Expand with $\mathbf{r}^{\mathrm{V}} = \bar{\mathbf{D}}^\top\mathbf{w}^{\mathrm{V}}$ and $\mathbf{r}^{\mathrm{T}} = \bar{\mathbf{D}}^\top\mathbf{w}^{\mathrm{T}}$. The unit diagonal of $\bar{\mathbf{G}}$ contributes $\langle\mathbf{w}^{\mathrm{V}},\mathbf{w}^{\mathrm{T}}\rangle$, and the off-diagonal entries contribute at most $\mu\sum_{k\neq l}w_k^{\mathrm{V}}w_l^{\mathrm{T}} = \mu(1-\langle\mathbf{w}^{\mathrm{V}},\mathbf{w}^{\mathrm{T}}\rangle)$ in absolute value, since both recipes are nonnegative and sum to one. □

Corollary D.4 (Recipe alignment and the atom term)**.** *The atom term averaged over the frames of segment i is $\gamma^2 \bar{\mathbf{w}}_i^{\mathrm{V}\top} \bar{\mathbf{G}}\, \mathbf{w}_c^{\mathrm{T}}$, and*

$$\bar{\mathbf{w}}_i^{\mathrm{V}\top} \bar{\mathbf{G}}\, \mathbf{w}_c^{\mathrm{T}} \;\geq\; \langle \bar{\mathbf{w}}_i^{\mathrm{V}}, \mathbf{w}_c^{\mathrm{T}} \rangle - \mu = \tfrac{1}{2}\big(\|\bar{\mathbf{w}}_i^{\mathrm{V}}\|_2^2 + \|\mathbf{w}_c^{\mathrm{T}}\|_2^2 - \|\bar{\mathbf{w}}_i^{\mathrm{V}} - \mathbf{w}_c^{\mathrm{T}}\|_2^2\big) - \mu. \tag{26}$$

Proof. The atom term is linear in the visual recipe, and $\bar{\mathbf{w}}_i^{\mathrm{V}}$ is again nonnegative and sums to one, so Proposition D.3 applies; the equality is the polarization identity. □

For $c = y_i$, the last term is the summand of $\mathcal{L}_{\mathrm{align}}$ in Eq. (7), so, for given recipe norms, minimizing $\mathcal{L}_{\mathrm{align}}$ raises this lower bound on the atom term of the correct category. Its gradient reaches the encoder through the attention scores $\alpha_{t,k} = (\mathbf{W}_{\mathrm{V}}\mathbf{f}_t)^\top \bar{\mathbf{D}}_k/\sqrt{d}$, pushing the frames of every seen category to select the atoms that its textual prototype selects, so that actions built from the same motion primitives tend to receive similar sparse codes (Figure 14).

D.3 Capacity

Proposition D.5 (Capacity)**.** *There are $\binom{K}{K'}$ sets of K' atoms. If the number of seen categories exceeds $\binom{K}{K'}$, at least two seen categories select the same set; for $K' = 1$ their textual recipes coincide and cannot distinguish them.*

Proof. Pigeonhole principle; for $K' = 1$ a recipe is the one-hot vector of its selected atom. □

With $K = 32$, $K' = 4$ gives $\binom{32}{4} = 35{,}960$ sets, far more than the number of categories, whereas $K' = 1$ gives only 32 recipes, fewer than the seen categories of either dataset; this is consistent with the drop at $K' = 1$ in Figure 3.

E Class Descriptions

For every category, GPT-4 generates one overall description and three descriptions of the start, middle, and end phases (Figure 9), matching the $M = 3$ phases of the dynamic descriptor. The prompt asks for body parts, joint movements, motion direction, and temporal change, and excludes appearance, background, and camera information, which skeletons cannot express. With the frozen CLIP ViT-B/32 text encoder Ψ and $N(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|_2$, the prototype of category c is

$$\mathbf{d}_c = N\big(\Psi(\mathrm{detail}_c)\big), \qquad \mathbf{h}_c = N\Big(\tfrac{1}{3}\textstyle\sum_{m=1}^{3} N\big(\Psi(\mathrm{phase}_{c,m})\big)\Big), \qquad \mathbf{p}_c = N\big(\mathbf{d}_c + \mathbf{h}_c\big), \tag{27}$$

so that the overall and the phase descriptions contribute equally. Table 9 shows examples.

```
Describe the human action "{action_name}" in the context of
"{dataset_context}".

Write one short overall description and three short descriptions for
its start, middle, and end phases. Focus on body parts, joint
movements, motion direction, and temporal changes. Avoid appearance,
background, and camera details. Use concise, concrete English.

Return JSON only:
{
  "action": "...",
  "segment": "...",
  "start": "...",
  "middle": "...",
  "end": "..."
}
```

Figure 9: Prompt used to generate the class descriptions.

Table 9: Generated descriptions of two seen (top) and two unseen (bottom) categories of PKU-MMD.

Category	Generated description
`bow`	*Overall:* The upper body bends forward at the hips and then returns to an upright position. *Start:* The trunk begins tilting forward while the feet remain planted. *Middle:* The trunk reaches its deepest forward bend, with the head following the torso. *End:* The trunk and head rise back to an upright position.
`brushing_hair`	*Overall:* One arm lifts toward the head, and the hand makes repeated brushing movements before lowering. *Start:* The shoulder lifts and the elbow bends, bringing the hand toward the top of the head. *Middle:* The hand makes repeated strokes along the head through coordinated shoulder and elbow movements. *End:* The hand moves away from the head and the arm lowers toward the side.
`hopping`	*Overall:* The body repeatedly rises and lands on one supporting foot while the other foot remains lifted. *Start:* Weight shifts onto one leg, and its knee bends as the opposite foot lifts. *Middle:* The supporting leg repeatedly extends for takeoff and bends on landing, producing vertical bounces. *End:* The bounces stop and the supporting leg steadies the body.
`jump_up`	*Overall:* Both legs bend and extend to lift the body upward, followed by a two-foot landing. *Start:* The hips and knees bend, lowering the body in preparation for takeoff. *Middle:* Both legs extend rapidly, lifting both feet off the ground as the body rises. *End:* Both feet land and the knees bend to absorb the downward movement.

F Implementation Details

Configuration. In addition to Section 4.2, the MS-GCN encoder uses two refinement stages with ten temporal layers and 64 channels, projected to $d = 512$; the classification temperature τ_c is initialized at 0.07 and learned; the smoothing term uses $\lambda_{\text{sm}} = 1$ and $\epsilon = 16$; and the codebook is initialized from $\mathcal{N}(0, 0.02^2)$. No data augmentation is used.

Background calibration. Following Eq. (11), β is subtracted from the background logit $\cos(\tilde{\mathbf{f}}_t, \tilde{\mathbf{p}}_{\text{bg}})/\tau_c$, so the decision is the arg max of the softmax of Eq. (8) after the background probability is multiplied by $e^{-\beta}$; β thus corrects the prior imbalance between background and action frames. Figure 11(d) shows that the results change little for β between 0.5 and 1.5.

Truncated smoothing. The gradient of the truncated term of Eq. (9) is

$$\nabla_{\mathbf{q}_t} \min\left(\|\mathbf{q}_t - \mathbf{q}_{t-1}\|_2^2, \epsilon\right) = 2(\mathbf{q}_t - \mathbf{q}_{t-1})\,\mathbb{I}\big[\|\mathbf{q}_t - \mathbf{q}_{t-1}\|_2^2 < \epsilon\big], \tag{28}$$

so adjacent frames whose predictions differ by more than $\sqrt{\epsilon}$ receive no gradient: jitter inside a segment is suppressed while the jumps at action boundaries are left untouched.

Baselines. No published method performs zero-shot segmentation of untrimmed skeleton sequences, so we adapt the recognition methods SMIE, DVTA, SynSE, SA-DVAE, and TDSM and the detection methods STALE and Ti-FAD to frame-level prediction. All of them use the same skeleton encoder, data processing, temporal smoothing loss, and evaluation; each applies its own visual–text alignment objective to the frames of seen actions, and at test time every frame is matched to the prototypes of the unseen categories by cosine similarity with background calibration.

G Additional Quantitative Results

SkelAtom is the best method on all 24 combinations of setting and metric in Tables 1 and 2; averaged over the four settings, its margins over the best baseline of each entry are $+4.00$ Acc, $+4.36$ Edit, $+2.18$ F1@50, and $+2.45$ AvgF1.

G.1 Enlarging the Unseen Label Space

Doubling the unseen categories of MCFS-130 from 13 to 26 adds candidate categories and removes training categories at the same time. SkelAtom loses 9.6% of its accuracy and 12.2% of its AvgF1, whereas the strongest baselines on U13 degrade much more, e.g., Ti-FAD loses 36.3% of its accuracy and STALE 29.9% of its AvgF1 (Figure 10).

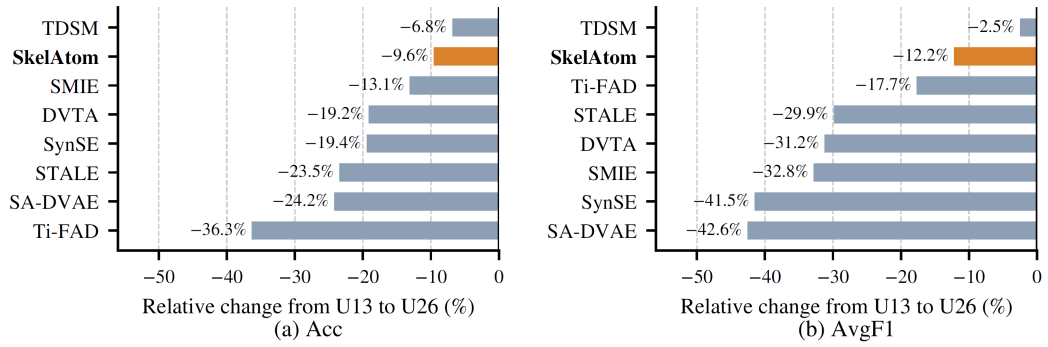


Figure 10: Relative change of frame-wise accuracy (a) and AvgF1 (b) when the number of unseen categories of MCFS-130 grows from 13 (U13) to 26 (U26), computed from Table 1.

G.2 Hyperparameter Sensitivity

Figure 11 extends Figure 3 to the loss weights, the residual strength γ, and the background scalar β. Over the tested ranges, λ_{KAG}, λ_{align}, and β change every metric by less than half a point, while the defaults of all four hyperparameters give the best or near-best results.

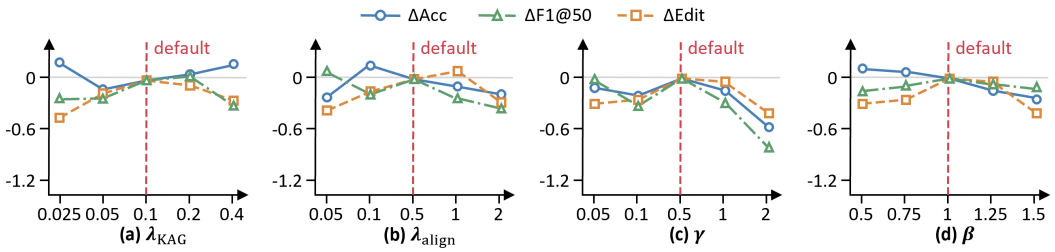


Figure 11: Sensitivity to the loss weights λ_{KAG} (a) and λ_{align} (b), the residual strength γ (c), and the background scalar β (d) on PKU-MMD X-sub, reported as the change of Acc, F1@50, and Edit from the default setting in percentage points. The dashed line marks the default value.

H Failure Cases and Limitations

Categories with similar motion. In Figure 12(a), typing, taking off glasses, and touching the neck are all performed with the hands close to the upper body while the legs stay still; SkelAtom labels the first two as writing and the third as typing, and Ti-FAD makes the same kind of errors. The descriptors measure motion over five body parts and three phases and cannot resolve differences confined to the hands, for which a finer partition is a natural extension. In (b), SkelAtom locates the flying sit spin `FlySitSpin4` accurately but labels it as another flying spin, `FlyCamelSpin2`.

Low-dynamic and short elements. The step sequence in (b) is largely predicted as background, because continuous footwork without a distinct takeoff, rotation, or landing resembles the gliding that fills a program; such low-dynamic elements are also hard for supervised segmentation (Li et al.,

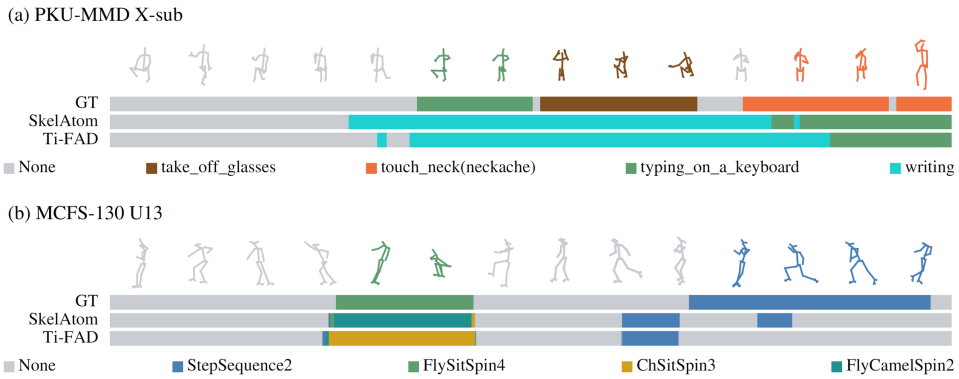


Figure 12: Failure cases on PKU-MMD X-sub (a) and MCFS-130 U13 (b). Keyframes sampled uniformly in time are colored by their ground-truth category (gray for background). Below them are the ribbons of the ground truth, SkelAtom, and Ti-FAD; the categories are listed under each panel.

2026). Short actions, such as the brief jumps in Figure 5, tend to be absorbed by their neighbors, since the temporal refinement favors long homogeneous segments and segments shorter than 12 frames receive no kinematic supervision.

Limitations. Unseen categories are described only by text, so the quality of the descriptions bounds how well a category can be matched; the descriptors do not capture hand and finger motion; and the invariances of Appendix C hold fully only for 3D skeletons.

I Additional Visualizations

I.1 Segmentation with Keyframes

Figure 13 adds keyframes and category names to the qualitative results of Figure 5. In (a), SkelAtom labels all three unseen actions correctly, whereas Ti-FAD splits `wear_jacket` and assigns the other two actions to wrong categories. In (b), SkelAtom labels `StepSequence2` and `3Flip_Euler_3Salchow` correctly, while Ti-FAD breaks the sequence into short pieces of several categories.

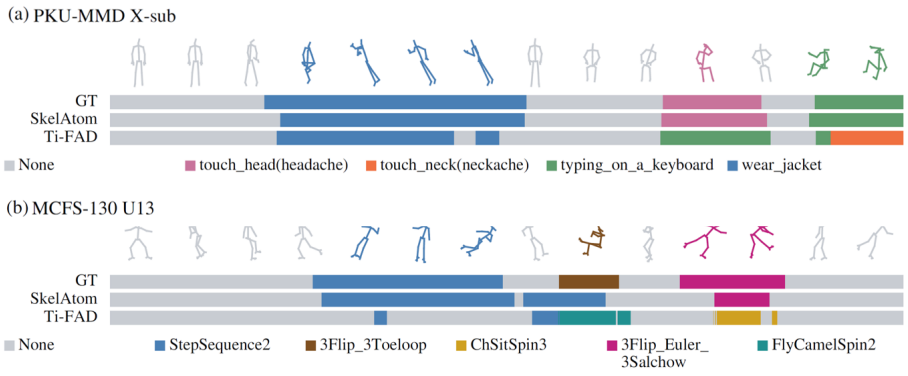


Figure 13: Segmentation of unseen categories with keyframes on PKU-MMD X-sub (a) and MCFS-130 U13 (b). Keyframes sampled uniformly in time are colored by their ground-truth category (gray for background). Below them are the ribbons of the ground truth, SkelAtom, and Ti-FAD; the categories are listed under each panel.

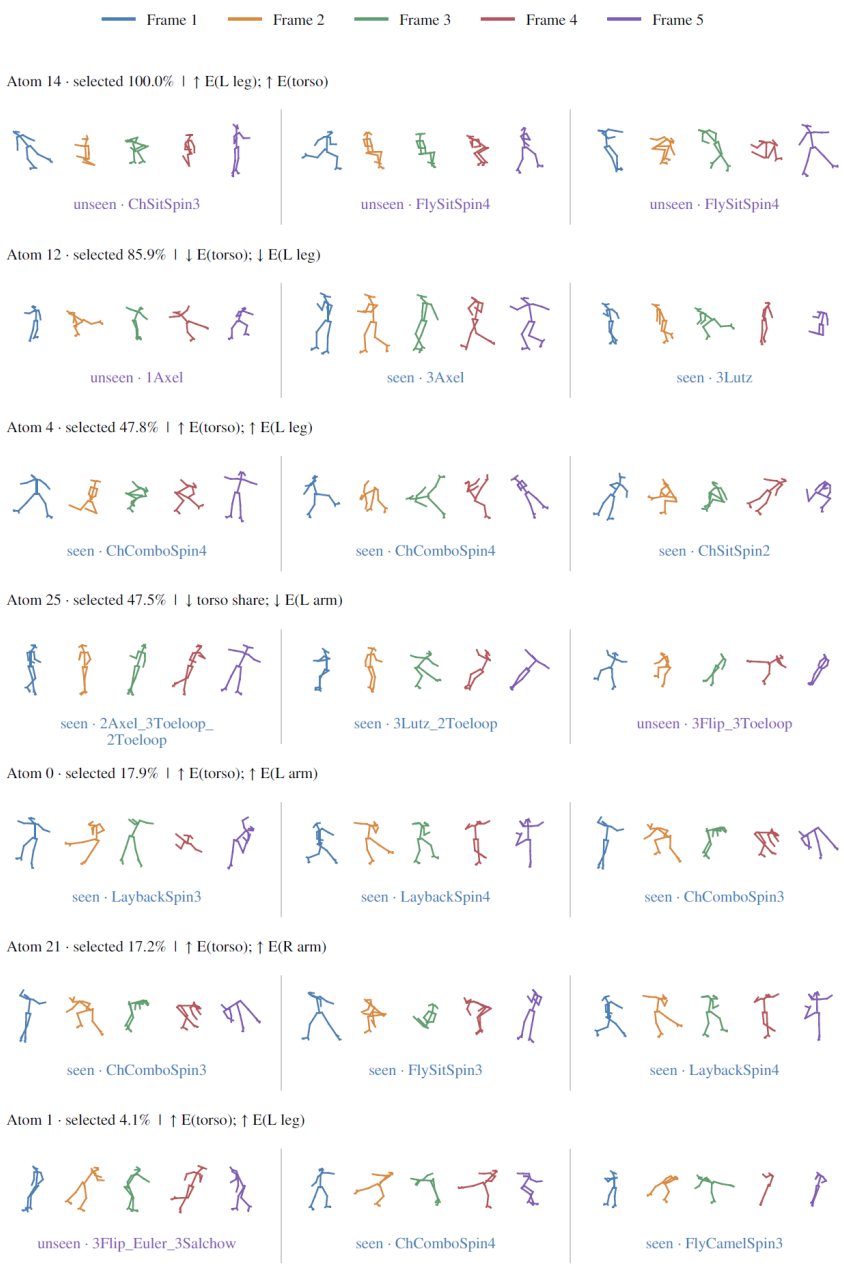


Figure 14: Motion atoms of the MCFS-130 U13 model. Each row shows one atom, the fraction of frames that select it, and the two descriptor components most correlated with its weight (↑ positive, ↓ negative correlation; E denotes a part energy). The atom is illustrated by the three segments with the largest average weight on it, each drawn as five frames sampled uniformly in time and colored from blue to purple, together with its split and category.

I.2 Motion Atoms

Figure 14 shows the atoms of the MCFS-130 U13 model, each illustrated by the three segments that place the largest weight on it. Atom 12 is selected by jumps, atom 25 by jump combinations, and atoms 0, 4, 14, and 21 by spins; unseen elements reuse the atoms of their family. For example, `1Axel` shares atom 12 with `3Axel` and `3Lutz`. Consistently, the spin atoms correlate positively and the jump atoms negatively with the torso energy or share, since the rotation of a spin appears as torso motion in root-centered coordinates.