MOTIONPROBE: DIRECT LLM PROBING FOR ZERO-SHOT SKELETON ACTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot skeleton action recognition commonly represents unseen actions by generating textual descriptions with an LLM and encoding them into semantic prototypes. However, we find that this two-stage process is not neutral: repeated generations under the same prompt produce different recognition outcomes, while auxiliary text encoders impose biases inherited from their own pretraining. The resulting prototypes therefore depend on both how the LLM verbalizes an action and how another model encodes it, showing that semantic construction is a critical part of zero-shot transfer rather than harmless preprocessing. Since the purpose of introducing an LLM is to access its rich action knowledge rather than the descriptions themselves, we propose MotionProbe, which directly reads prompt-conditioned hidden states from a frozen LLM, avoiding the generation variability and encoder bias introduced by the conventional pipeline. Using class-shared motion prompts, MotionProbe probes each action from complementary perspectives and organizes the resulting hidden states into a semantic constellation. These semantic views are then progressively transformed into task-ready prototypes: seen skeletons ground them in discriminative motion geometry, while unlabeled test skeletons further calibrate each candidate constellation and determine how its views should be combined. Extensive experiments on multiple benchmarks demonstrate that MotionProbe attains state-of-the-art performance.

1 INTRODUCTION

Human action recognition underpins a broad range of applications, from human–computer interaction and healthcare to intelligent surveillance. Among the available sensing modalities, 3D skeletons distill human motion into compact joint trajectories, offering a structured representation that protects privacy and remains robust to variations in appearance and background (Yan et al., 2018). Human behavior, however, does not form a closed world. New gestures, rare activities, and evolving forms of interaction continually expand the label space, making exhaustive annotation impractical. Zero-shot skeleton-based action recognition (ZSSAR) addresses this open setting by transferring knowledge from seen to unseen actions through language semantics. With no unseen training samples, semantic prototypes define the novel classes and determine how knowledge is transferred to them.

Accordingly, recent progress in ZSSAR has focused on enriching these prototypes and strengthening their correspondence with skeleton motion. Early methods match an entire skeleton sequence with a single class embedding (Gupta et al., 2021; Zhou et al., 2023). More recent studies use LLMs to obtain detailed action knowledge at low cost, then pursue finer correspondence through descriptions of individual body regions (Zhu et al., 2024; Chen et al., 2024), learnable textual prompts (Zhu et al., 2025a), object and scene context (Wang et al., 2026), or relations among action classes (Chen et al., 2026). Although their alignment strategies differ, most of these methods construct semantic prototypes through the same two-stage pipeline: an LLM first verbalizes its knowledge as an action description, and CLIP (Radford et al., 2021) or another text encoder then maps that description to a feature representation.

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.

Although effective, this pipeline raises a central question: is semantic construction merely a neutral preprocessing step? As shown in Table 1, we conduct a controlled study based on Flora, the latest open-source state-of-the-art ZSSAR method. We find that repeated LLM generations from the same prompt produce different recognition accuracies, while replacing CLIP with BERT under fixed descriptions and downstream settings also changes performance, revealing generation and auxiliary encoder dependence. More importantly, carefully designed descriptions remain competitive with only a lightweight MLP, showing that prototype quality strongly affects zero-shot transfer. Semantic prototype construction is therefore not a simple, neutral preprocessing step, but a central modeling problem on par with cross-modal alignment.

Table 1: Controlled diagnosis based on Flora, with ZSL accuracy (%) reported.

Setting	NTU-60		NTU-120	
	55/5	48/12	110/10	96/24
Reported reference				
Flora	86.3	65.3	79.6	66.4
Generation dependence (same prompt)				
Generated description 1	**85.2**	**62.1**	**76.5**	62.8
Generated description 2	81.6	59.7	75.9	**64.2**
Generated description 3	83.3	60.2	73.3	61.6
Encoder dependence (fixed descriptions)				
CLIP	85.2	62.1	76.5	62.8
BERT	81.4	58.7	70.4	60.2
Designed semantics with lightweight alignment				
Designed text + MLP	83.5	62.4	76.6	63.4

The observed generation and encoder dependence arise because LLM knowledge passes through two successive transformations. Autoregressive generation determines which motion cues are expressed, how much detail is included, and how they are worded. Repeated generation can therefore produce different linguistic realizations from the same prompt. The auxiliary text encoder then maps each description according to biases inherited from its own pretraining data and objective. Whether learned from generic language or image–text pairs, these biases may not emphasize the motion cues observable from joint trajectories. The final prototype therefore depends on both how the LLM describes an action and how the auxiliary encoder represents that description.

Since LLMs are introduced for their broad action knowledge rather than the descriptions themselves, MotionProbe directly reads hidden states conditioned on shared prompts, bypassing autoregressive generation and auxiliary text encoding by design, as shown in Figure 1. The strong performance achieved by carefully designed prototypes with a simple MLP further motivates us to place the modeling emphasis on semantic construction rather than alignment complexity. Accordingly, the probed states are organized into a semantic constellation that serves as the shared representation throughout the method. Seen skeletons shape this constellation into a discriminative, skeleton-grounded form during training, while unlabeled test skeletons further calibrate the location and view composition of each candidate constellation at inference. The final prototypes are thus constructed by progressively refining the same constellation, with cross-modal alignment handled by a simple adapter and projector. Our main contributions are summarized as follows:

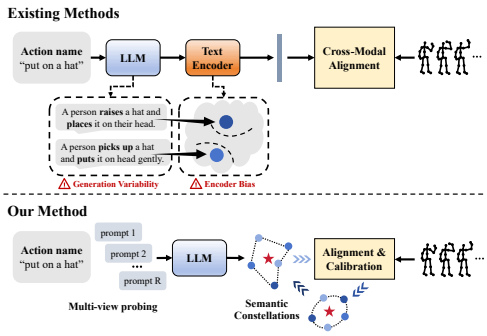


Figure 1: Comparison of semantic transfer paradigms. Existing methods generate and encode descriptions, introducing generation variability and encoder bias. MotionProbe directly probes an LLM to form semantic constellations for alignment and calibration.

- We are the first to directly probe a frozen LLM for semantic prototype construction in ZSSAR, using hidden states conditioned on shared prompts without intermediate text generation or an auxiliary text encoder.
- We propose semantic constellations that transform general LLM representations into task-ready prototypes through complementary motion views, grounding with seen skeletons, and calibration with target data, while keeping cross-modal alignment lightweight.
- Extensive experiments show that our MotionProbe achieves state-of-the-art results in both ZSL and GZSL settings on NTU RGB+D 60, NTU RGB+D 120, and PKU-MMD datasets.

2 Related Work

2.1 Zero-Shot Skeleton Action Recognition

ZSSAR transfers knowledge beyond the training label space by matching skeleton representations with semantic prototypes defined for both seen and unseen actions. Early approaches therefore focused on constructing these prototypes from action names or manually normalized descriptions. SynSE decomposed descriptions into verb and noun components for compositional transfer (Gupta et al., 2021), whereas SMIE used pretrained language embeddings as global anchors and strengthened their correspondence with skeleton dynamics through mutual-information maximization (Zhou et al., 2023). These early efforts established compact linguistic priors as an effective bridge to unseen actions. However, names and short descriptions offer limited detail, which motivated richer accounts of human motion.

With broad world knowledge and flexible language generation, LLMs offer a scalable way to turn action names into detailed semantic descriptions from multiple perspectives. PURLS and STAR used this capacity to describe body regions, temporal evolution, and local motion for finer alignment (Zhu et al., 2024; Chen et al., 2024). InfoCPL, Neuron, and SkeletonContext further expanded semantic coverage through multiple descriptions, iterative interaction, or object and scene context (Xu et al., 2025; Chen et al., 2025; Wang et al., 2026). Meanwhile, SCoPLe adapts CLIP semantics through prompt learning during seen-class training (Zhu et al., 2025a). While these approaches enrich semantic content and improve its alignment with skeleton motion, a dominant branch still generates action descriptions and then encodes them with CLIP or related models. This shared interface remains fragile because the resulting prototypes depend on both how an LLM verbalizes an action and how an auxiliary encoder interprets the description. Our MotionProbe takes a complementary direction, transferring LLM knowledge through direct probing rather than generated descriptions and auxiliary text encoding.

2.2 Relational and Target-Aware Prototype Refinement

Rich semantic content alone does not determine how a prototype should relate to neighboring classes or target samples. Flora contextualizes semantic anchors through relations among classes and encourages consistent geometry across modalities (Chen et al., 2026). Target data offer complementary evidence by revealing how unseen actions are distributed in the skeleton space. PGFA constructs visual prototypes from pseudo-labeled samples and uses them to realign unseen text features (Zhou et al., 2025), while Skeleton-Cache moves prototype refinement to inference through a skeleton cache (Zhu et al., 2025b). Together, these studies show that class relations reveal which semantic distinctions matter, whereas target structure reveals how those semantics are realized visually. Our MotionProbe brings both forms of evidence into a shared semantic constellation. Relations among actions guide discriminative motion learning, while target skeletons calibrate the locations and view compositions of candidate constellations.

3 Method

3.1 Problem Formulation and Overview

Let $\mathcal{D}^s_{\text{tr}} = \{(\mathbf{X}^s_i, y^s_i)\}_{i=1}^{N_s}$ be the training set from seen classes $\mathcal{Y}^s$. A skeleton sequence $\mathbf{X}^s_i \in \mathbb{R}^{3\times T\times V\times M}$ contains the 3D coordinates of V joints over T frames for at most M people. The unseen classes $\mathcal{Y}^u$ are disjoint from $\mathcal{Y}^s$ and provide no labeled skeletons. Zero Shot Learning (ZSL) predicts among $\mathcal{Y}^u$, whereas Generalized Zero Shot Learning (GZSL) predicts among $\mathcal{Y}^s \cup \mathcal{Y}^u$. At inference, MotionProbe further uses the unlabeled test set for target calibration. In ZSL, the test set contains only unseen-class samples, whereas in GZSL it contains samples from both seen and unseen classes.

Figure 2 presents MotionProbe as a progressive process for constructing semantic prototypes from LLM knowledge. First, class-shared motion prompts directly probe a frozen LLM to form a raw semantic constellation $\mathcal{H}_y$. During training, relations among these constellations identify likely confusions, while seen skeletons reveal the motion directions along which they can be distinguished. A simple semantic adapter and projector then transform $\mathcal{H}_y$ into the skeleton-grounded constellation $\mathcal{E}_y$. During inference, unlabeled test skeletons calibrate each candidate constellation's location and

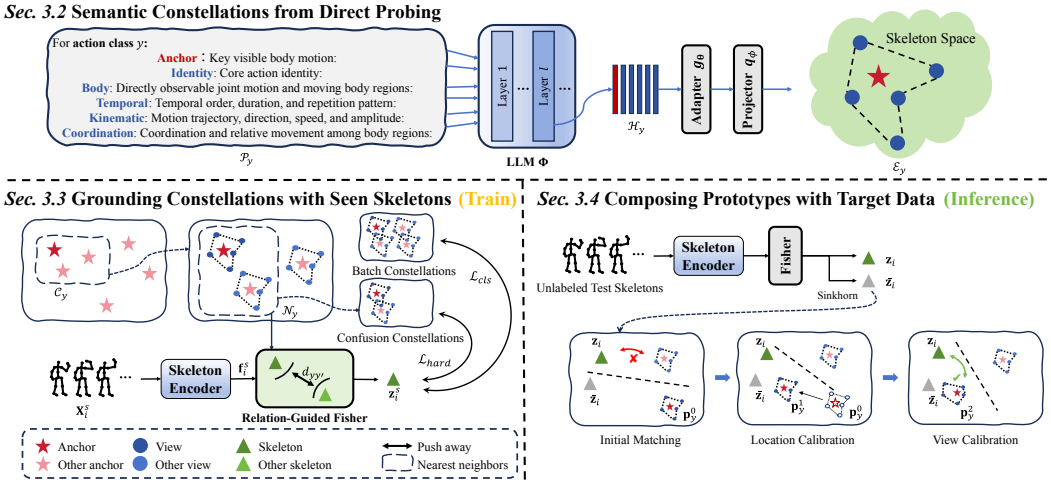


Figure 2: Overview of MotionProbe. Top: class-shared motion prompts directly probe a frozen LLM to form semantic constellations. Bottom left: during training, seen skeletons ground these constellations in discriminative motion geometry. Bottom right: during inference, unlabeled test skeletons calibrate prototype locations and view compositions to obtain the final prototypes.

view composition to obtain the final prototype $\mathbf{p}_y^2$. Thus, $\mathcal{H}_y$, $\mathcal{E}_y$, and $\mathbf{p}_y^2$ denote successive stages of the same semantic prototype.

3.2 Semantic Constellations from Direct Probing

Direct multi-view probing. For each class y, a reference prompt $p^0(y)$ provides a stable anchor, while $R = 5$ shared prompts $\{p^r(y)\}_{r=1}^R$ query action identity, involved body regions, temporal pattern, trajectory, and coordination. We denote the complete prompt set by $\mathcal{P}_y = \{p^r(y)\}_{r=0}^R$. Rather than asking the LLM to generate a description, we directly read the hidden state at the final prompt token:

$$\mathbf{h}_y^r = \Phi_{\text{LLM}}^{(l)}(p^r(y)), \quad r = 0, \ldots, R, \tag{1}$$

where $\Phi_{\text{LLM}}^{(l)}$ denotes layer l of the frozen LLM. We collect these states as the *raw semantic constellation* $\mathcal{H}_y = \{\mathbf{h}_y^r\}_{r=0}^R$, which records prompt-conditioned action knowledge in the frozen LLM space.

Projecting the semantic constellation. The states in $\mathcal{H}_y$ reside in the frozen LLM space and therefore cannot be directly used with skeleton representations. Let d_{ske} denote the dimension of the fixed skeleton representation space. We introduce a shared semantic adapter g_θ and projector q_ϕ, both implemented as MLPs, to map each state into this space:

$$\mathbf{e}_y^r = q_\phi\big(g_\theta(\mathbf{h}_y^r)\big) \in \mathbb{R}^{d_{\text{ske}}}. \tag{2}$$

The resulting set $\mathcal{E}_y = \{\mathbf{e}_y^r\}_{r=0}^R$ preserves the anchor–view structure of the raw constellation. Its anchor provides a stable class reference, while the remaining views provide complementary motion perspectives.

3.3 Grounding Constellations with Seen Skeletons

The raw constellations encode broad action knowledge but are not yet grounded in what can be observed from joint motion. We ground them in three steps. First, relations among raw constellations identify the class confusions that matter. Next, seen skeleton statistics locate the motion directions that resolve them. Finally, the semantic adapter and projector are optimized using this geometry to shape $\mathcal{E}_y$ into its skeleton-grounded form.

Semantic confusion graph. Grounding begins by identifying the semantic distinctions for which skeleton evidence is most needed. We construct the graph from the raw constellations $\mathcal{H}_y$ in the frozen LLM space, before any seen-class adaptation. For each seen class y, the raw anchor retrieves a shortlist $\mathcal{C}_y$ of its $2K$ nearest classes. The motion views then verify whether each candidate remains similar when the two actions are considered from multiple motion perspectives. Let $A_{yy'} \in \mathbb{R}$ denote the agreement obtained by matching each view of one constellation to its closest view in the other, in both directions. We score a candidate by

$$s_{yy'} = \cos(\mathbf{h}_y^0, \mathbf{h}_{y'}^0) + \log \frac{1 + A_{yy'}}{2}. \tag{3}$$

The top K candidates form the confusion neighborhood $\mathcal{N}_y$. The anchor proposes classes with similar overall identity, while the motion views retain only those whose similarity is also supported by finer action semantics. The graph thus selects the local class distinctions that the skeleton space should resolve.

Discriminative motion geometry. The graph specifies which classes are likely to be confused, but not how to separate them through joint motion. We therefore translate its edges into directions measured from seen skeletons. A fixed skeleton encoder maps $\mathbf{X}_i^s$ to a pooled feature $\mathbf{f}_i^s \in \mathbb{R}^{d_{\text{ske}}}$. We first compute a regularized Fisher whitener $\mathbf{W}_0 \in \mathbb{R}^{d_{\text{ske}} \times d_{\text{ske}}}$ from seen skeletons to suppress within-class variation. Let $\boldsymbol{\mu}_y, \bar{\boldsymbol{\mu}} \in \mathbb{R}^{d_{\text{ske}}}$ denote the class and global means. After whitening and normalization, each graph edge (y, y') yields a unit direction $\mathbf{d}_{yy'} \in \mathbb{R}^{d_{\text{ske}}}$ along the skeleton-space difference between the two classes. We aggregate these directions as

$$\boldsymbol{\Sigma}_c = \sum_{y \in \mathcal{Y}^s} \sum_{y' \in \mathcal{N}_y} \alpha_{yy'} \mathbf{d}_{yy'} \mathbf{d}_{yy'}^\top \in \mathbb{R}^{d_{\text{ske}} \times d_{\text{ske}}}, \tag{4}$$

where $\alpha_{yy'} \in \mathbb{R}$ is the normalized semantic score of the edge. Directions that repeatedly separate strongly related classes receive larger support in $\boldsymbol{\Sigma}_c$. Let $\boldsymbol{\Sigma}_c = \mathbf{U} \operatorname{diag}(\boldsymbol{\lambda}) \mathbf{U}^\top$. We convert the normalized eigenvalues into bounded directional gains and combine them with the Fisher whitener:

$$\mathbf{W} = \mathbf{U} \operatorname{diag}\left(\sqrt{1 + \boldsymbol{\lambda} / \lambda_{\max}} \right) \mathbf{U}^\top \mathbf{W}_0 \in \mathbb{R}^{d_{\text{ske}} \times d_{\text{ske}}}, \tag{5}$$

where $\lambda_{\max} \in \mathbb{R}$ is the largest eigenvalue. The seen-skeleton representation is then $\mathbf{z}_i^s = (\mathbf{f}_i^s - \bar{\boldsymbol{\mu}}) \mathbf{W}^\top \in \mathbb{R}^{d_{\text{ske}}}$. The resulting relation-guided discriminative transform $\mathbf{W}$ reshapes skeleton representations around the distinctions selected by the semantic graph, providing the motion geometry used to ground $\mathcal{E}_y$.

Skeleton-grounded constellation learning. With this geometry established, we ground the projected constellation $\mathcal{E}_y$ using seen skeletons. The compatibility between $\mathbf{z}_i^s$ and class y is

$$\ell_{iy} = \tau_p \log \sum_{r=0}^{R} \exp\left(\frac{\cos(\mathbf{z}_i^s, \mathbf{e}_y^r)}{\tau_p} \right), \tag{6}$$

where τ_p controls the concentration of the matching over semantic views. This formulation allows each skeleton to draw more strongly on the constellation element most consistent with its observed motion.

Global and hard-neighbor grounding. For a batch $\mathcal{B}$, let $\mathcal{Y}_\mathcal{B}$ denote the classes appearing in it. The global classification loss grounds each constellation in the overall seen-class structure:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(\ell_{iy_i^s})}{\sum_{c \in \mathcal{Y}_\mathcal{B}} \exp(\ell_{ic})}. \tag{7}$$

The classes selected by the confusion graph serve as local hard competitors, so the same difficult relations also focus the grounding objective:

$$\mathcal{L}_{\text{hard}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(\ell_{iy_i^s})}{\sum_{c \in \{y_i^s\} \cup \mathcal{N}_{y_i^s}} \exp(\ell_{ic})}. \tag{8}$$

The training objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_h \mathcal{L}_{\text{hard}}. \tag{9}$$

The two mechanisms exploit the same difficult class relations from complementary sides: $\mathbf{W}$ emphasizes the skeleton-space directions that separate them, while $\mathcal{L}_{\text{hard}}$ encourages the motion views to encode the distinctions needed to resolve them. Together with $\mathcal{L}_{\text{cls}}$, jointly optimizing g_θ and q_ϕ shapes each $\mathcal{E}_y$ into the *skeleton-grounded semantic constellation*.

3.4 Composing Prototypes with Target Data

Seen skeletons ground $\mathcal{E}_y$ in discriminative motion geometry, but they cannot reveal how the candidate actions are distributed in the target data. We therefore use unlabeled test skeletons to further calibrate each candidate constellation in two stages; in GZSL, calibration is performed over the complete test distribution containing both seen and unseen samples. The first adjusts its global location, and the second refines its internal view composition.

Constellation-level calibration. We first summarize each candidate constellation with its anchor and the mean of its motion views: $\mathbf{p}_y^0 = \text{Norm}\Big(\mathbf{e}_y^0 + \frac{1}{R}\sum_{r=1}^{R} \mathbf{e}_y^r\Big)$. We then match these initial prototypes with the unlabeled test skeletons. Sinkhorn normalization balances the assignments, preventing a few initially favored classes from absorbing most samples. The most confident assignments to class y are averaged to form its target center $\bar{\mathbf{z}}_y$, which calibrates the initial prototype as

$$\mathbf{p}_y^1 = \text{Norm}\big((1-\beta)\mathbf{p}_y^0 + \beta \bar{\mathbf{z}}_y\big)\,. \tag{10}$$

Here, $\text{Norm}(\cdot)$ denotes ℓ_2 normalization, and β controls the contribution of target evidence. This step moves the prototype toward the target support while retaining the semantic prior supplied by the constellation.

View-level calibration. The relocated prototype does not yet indicate which motion views best reflect the target action. We therefore use $\mathbf{p}_y^1$ to estimate their relevance to the observed target motion:

$$\rho_y^r = \underset{r=1,\dots,R}{\text{softmax}}\big(\cos(\mathbf{p}_y^1, \mathbf{e}_y^r)\big)\,. \tag{11}$$

More relevant motion views receive larger weights, while the anchor remains a fixed class reference. Repeating the balanced assignment with the updated view composition refines $\bar{\mathbf{z}}_y$. The final prototype is

$$\mathbf{p}_y^2 = \text{Norm}\left((1-\beta)\,\text{Norm}\left(\mathbf{e}_y^0 + \sum_{r=1}^{R} \rho_y^r \mathbf{e}_y^r\right) + \beta \bar{\mathbf{z}}_y\right). \tag{12}$$

Target skeletons therefore first relocate each candidate constellation and then determine how its motion views contribute to the final prototype. This completes the construction of $\mathbf{p}_y^2$ from the same semantic constellation.

ZSL and GZSL prediction. At inference, let $\mathbf{z}_i \in \mathbb{R}^{d_{\text{ske}}}$ denote the representation of a test skeleton obtained using the same fixed encoder and transform $\mathbf{W}$. We classify it by cosine similarity to the final prototypes:

$$\widehat{y}_i = \underset{y\in\mathcal{Y}}{\arg\max}\,\big[\cos(\mathbf{z}_i, \mathbf{p}_y^2) - \gamma\,\mathbb{I}[y \in \mathcal{Y}^s]\big]\,. \tag{13}$$

For ZSL, $\mathcal{Y} = \mathcal{Y}^u$ and $\gamma = 0$. For GZSL, $\mathcal{Y} = \mathcal{Y}^s \cup \mathcal{Y}^u$, and calibrated stacking (Chao et al., 2016) uses γ to reduce the bias toward seen classes.

4 Experiments

4.1 Experimental Setup

Datasets and protocols. We evaluate MotionProbe on three widely used skeleton action recognition benchmarks. NTU RGB+D 60 (NTU-60) contains 56,880 sequences from 60 action classes performed by 40 subjects (Shahroudy et al., 2016). NTU RGB+D 120 (NTU-120) extends it to

114,480 sequences, 120 classes, and 106 subjects (Liu et al., 2020). PKU-MMD contains approximately 20,000 sequences from 51 action classes (Liu et al., 2017) and we use its first phase following prior ZSSAR studies. We adopt the established class splits: 55/5 and 48/12 on NTU-60, 110/10 and 96/24 on NTU-120, and 46/5 and 39/12 on PKU-MMD, where the two numbers denote the counts of seen and unseen classes. Results are reported under the cross-subject protocol for NTU and under both the cross-subject (Xsub) and cross-view (Xview) protocols for PKU-MMD I. Target calibration follows the test-time adaptation setting: it uses the unseen test set in ZSL and the complete test set containing both seen and unseen samples in GZSL.

Evaluation metrics. For ZSL, we report Top-1 accuracy over unseen classes. For GZSL, we report the accuracies on seen and unseen classes, denoted by S and U, together with their harmonic mean $H = \frac{2SU}{S+U}$.

Implementation details. We sample each skeleton sequence to 64 frames and use Shift-GCN (Cheng et al., 2020), pretrained only on the seen classes, as the frozen skeleton encoder. We probe the penultimate layer of a frozen Qwen3-4B (Yang et al., 2025) using one reference prompt and five shared motion prompts. The semantic adapter and projector are optimized with AdamW using a batch size of 64, a learning rate of 10^{-3}, and a weight decay of 10^{-4}. We set the confusion neighborhood size K to 8 and the hard-neighbor loss weight λ_h to 2. At inference, one refinement round uses the top 30% assignments for each class to estimate its target center, with 50 Sinkhorn iterations.

Table 2: Comparison with state-of-the-art methods on NTU RGB+D 60 and NTU RGB+D 120 under the ZSL and GZSL settings. Best results are shown in bold.

		NTU RGB+D 60								NTU RGB+D 120							
		55/5 Split				48/12 Split				110/10 Split				96/24 Split			
Method	Venue	ZSL	S	U	H	ZSL	S	U	H	ZSL	S	U	H	ZSL	S	U	H
ReViSE	ICCV 2017	69.5	40.8	50.2	45.0	24.0	21.8	14.8	17.6	19.8	0.6	14.5	1.1	8.5	3.4	1.5	2.1
JPoSE	ICCV 2019	73.7	66.5	53.5	59.3	27.5	28.6	18.7	22.6	57.3	53.6	11.6	19.1	38.1	41.0	3.8	6.9
CADA-VAE	CVPR 2019	76.9	56.1	56.0	56.0	32.1	50.4	25.0	33.4	52.5	50.2	43.9	46.8	38.7	48.3	27.5	35.1
SynSE	ICIP 2021	71.9	51.3	47.4	49.2	31.3	44.1	22.9	30.1	52.4	57.3	43.2	49.5	41.9	48.1	32.9	39.1
SMIE	ACMMM 2023	77.9	–	–	–	41.5	–	–	–	61.3	–	–	–	42.3	–	–	–
PURLS	CVPR 2024	79.2	–	–	–	41.0	–	–	–	72.0	–	–	–	52.0	–	–	–
STAR	ACMMM 2024	81.4	69.0	69.9	69.4	45.1	62.7	37.0	46.6	63.3	59.9	52.7	56.1	44.3	51.2	36.9	42.9
SA-DVAE	ECCV 2024	82.4	62.8	70.8	66.3	41.4	50.2	36.9	42.6	68.8	61.1	59.8	60.4	46.1	58.8	35.8	44.5
SCoPLe	CVPR 2025	84.1	69.6	71.9	70.8	53.0	54.5	61.8	57.9	74.5	63.5	61.1	62.3	52.2	53.3	51.2	52.2
Neuron	CVPR 2025	86.9	69.1	73.8	71.4	62.7	61.6	56.8	59.1	71.5	**67.6**	59.5	63.3	57.1	**67.5**	44.4	53.6
TDSM	ICCV 2025	86.5	–	–	–	56.0	–	–	–	74.2	–	–	–	65.1	–	–	–
FS-VAE	ICCV 2025	86.9	77.0	74.5	75.7	57.2	56.2	48.6	52.1	74.4	59.2	67.9	63.3	62.5	57.8	51.9	54.7
Skeleton-Cache	NeurIPS 2025	89.4	64.6	**79.2**	71.2	47.8	53.2	44.4	48.4	74.3	62.2	65.8	64.0	53.1	59.9	44.1	50.8
DVTA	PR 2026	79.3	–	–	–	44.1	–	–	–	74.9	–	–	–	51.8	–	–	–
Flora	CVPR 2026	86.3	75.9	78.8	**77.4**	65.3	63.7	57.5	60.5	79.6	66.2	66.0	66.1	66.4	55.9	50.7	53.2
SkeletonContext	CVPR 2026	**89.6**	**78.3**	76.1	77.1	64.4	65.5	57.1	61.1	74.2	61.1	65.5	63.1	60.1	64.6	49.6	56.1
MotionProbe	–	84.8	76.2	74.4	75.3	**75.9**	**73.2**	**62.0**	**67.1**	**83.6**	65.8	**69.2**	**67.5**	**66.9**	60.9	**52.6**	**56.4**

Table 3: Comparison with state-of-the-art methods on PKU-MMD I under the Xsub and Xview protocols for ZSL and GZSL. Best results are shown in bold.

		PKU-MMD I (Xsub)								PKU-MMD I (Xview)							
		46/5 Split				39/12 Split				46/5 Split				39/12 Split			
Method	Venue	ZSL	S	U	H	ZSL	S	U	H	ZSL	S	U	H	ZSL	S	U	H
ReViSE	ICCV 2017	54.2	44.9	34.5	39.1	19.3	35.7	13.0	19.0	54.1	50.7	39.9	44.6	12.7	34.5	9.4	14.8
JPoSE	ICCV 2019	57.4	67.0	43.0	52.4	27.0	64.8	26.5	37.6	53.1	72.9	42.5	53.7	22.8	57.6	20.2	29.9
CADA-VAE	CVPR 2019	73.9	76.2	51.8	61.7	33.7	69.0	29.3	41.1	74.5	79.9	61.5	69.5	29.5	62.4	28.3	39.0
SynSE	ICIP 2021	69.5	77.8	40.2	53.0	36.5	71.9	30.0	42.3	71.7	69.9	51.1	59.0	25.4	61.9	22.6	33.1
SMIE	ACMMM 2023	72.9	–	–	–	44.2	–	–	–	71.6	–	–	–	40.7	–	–	–
STAR	ACMMM 2024	76.3	59.1	72.3	65.0	50.2	72.7	44.7	55.4	75.4	73.5	72.2	72.8	50.5	69.8	47.5	56.5
Neuron	CVPR 2025	89.2	78.9	75.9	77.4	61.4	78.3	**56.8**	65.9	88.2	78.7	77.8	78.3	62.2	75.4	57.1	65.0
Flora	CVPR 2026	79.1	76.0	65.9	70.6	55.4	74.5	52.3	61.5	76.3	76.0	71.4	73.7	58.7	**77.2**	55.6	64.6
MotionProbe	–	**95.2**	**86.9**	**85.1**	**86.0**	**63.2**	**84.5**	54.6	**66.3**	**94.6**	**90.8**	**80.9**	**85.6**	**63.7**	75.2	**58.3**	**65.7**

4.2 Comparison with the State of the Art

Tables 2 and 3 compare MotionProbe with widely adopted zero-shot baselines and recent ZSSAR methods under the established class splits (Tsai et al., 2017; Wray et al., 2019; Schönfeld et al., 2019; Li et al., 2024; Do & Kim, 2025; Wu et al., 2025; Kuang et al., 2026).

Zero-shot learning. MotionProbe achieves the best ZSL accuracy on seven of the eight protocols. The largest gain appears on NTU-60 under 48/12, where it exceeds the previous best by 10.6 points. It also leads on both NTU-120 splits and all four PKU-MMD I protocols. These results show that class-shared prompting and constellation-based prototype construction remain effective across different dataset scales, unseen-class ratios, and evaluation protocols.

Generalized zero-shot learning. MotionProbe ranks first in harmonic mean on seven protocols, including every setting except NTU-60 under 55/5. On NTU-60 under 48/12, it improves the previous best by 6.0 points. Its consistently strong harmonic mean indicates balanced performance across seen and unseen classes rather than a preference for either side of the label space.

Table 4: Component ablation on NTU-60 and NTU-120. ZSL accuracy (%) is reported.

Components				NTU-60		NTU-120	
Motion Views	W	$\mathcal{L}_{\text{hard}}$	Target Calibration	55/5	48/12	110/10	96/24
✗	✗	✗	✗	79.3	58.2	74.4	56.8
✓	✗	✗	✗	81.9	62.0	78.6	60.1
✓	✗	✗	✓	82.2	70.5	80.3	63.2
✓	✗	✓	✓	81.6	73.0	81.9	64.5
✓	✓	✗	✓	83.9	**77.2**	82.4	66.7
✓	✓	✓	✓	**84.8**	75.9	**83.6**	**66.9**

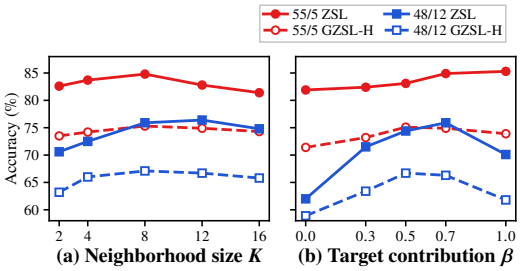


Figure 3: Sensitivity to the confusion neighborhood size K and target contribution β on NTU-60.

4.3 Ablation Studies

Component ablation. Table 4 isolates the main components. All variants share the frozen encoders, class anchor, adapter/projector, and $\mathcal{L}_{\text{cls}}$. Motion views improve all splits, confirming complementary cues, while target calibration is especially helpful on broader splits. The relation-guided transform further separates confusable classes, and $\mathcal{L}_{\text{hard}}$ adds local supervision. Their combination is best on three protocols; the 48/12 exception suggests that overly local supervision can hurt transfer when the unseen label space is broader.

Semantic interface. To isolate the semantic interface itself, Table 5 keeps the six semantic views, skeleton encoder, grounding objectives, relation-guided transform, and target calibration fixed, changing only how LLM knowledge is converted into semantic features. Direct Probe is consistently strongest across all four splits. Compared with generated descriptions re-encoded by either CLIP or the same LLM, direct access to prompt-conditioned hidden states avoids an explicit verbalization step and yields more effective semantics for downstream alignment and calibration. This controlled result supports our central premise that the interface to the LLM, rather than only later refinement modules, materially affects ZSSAR performance.

Table 5: Semantic-interface control. ZSL accuracy (%).

Semantic interface	Verbalization	Auxiliary encoder	55/5	48/12	110/10	96/24
Generate + CLIP	✓	✓	82.4	65.3	77.5	62.4
Generate + same LLM	✓	✗	82.1	69.7	74.8	63.6
Direct Probe	✗	✗	**84.8**	**75.9**	**83.6**	**66.9**

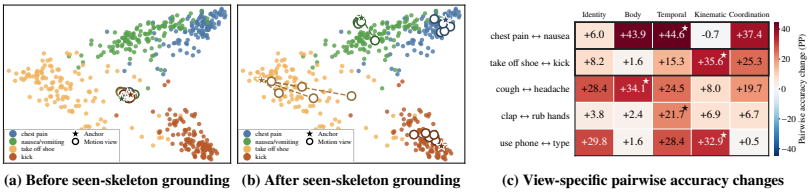
Parameter sensitivity. Figure 3(a) examines the confusion neighborhood size K. Performance peaks at $K = 8$: smaller neighborhoods omit useful confusions, whereas larger ones introduce weakly related classes that dilute the discriminative geometry.

Figure 3(b) varies β, which balances the semantic prototype and its target center. Nonzero values consistently improve over $\beta = 0$, while $\beta = 1$ degrades performance because the semantic prior is discarded. Performance is strongest around 0.5–0.7, and we use $\beta = 0.7$.

4.4 Qualitative Analysis

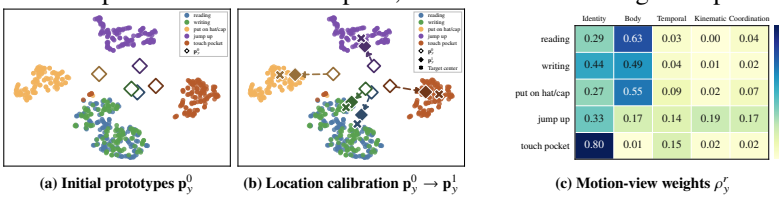
Grounding with seen skeletons. Figure 4 visualizes how seen skeletons shape semantic constellations. Before grounding, the anchors and motion views of the selected confusion pairs largely overlap despite their distinct skeleton distributions. After training, the constellation elements become better aligned with their corresponding skeleton clusters and more separated from the competing class. Panel (c) further shows that the largest pairwise gain arises from different motion views for different confusion pairs. Seen skeleton supervision therefore does not favor a single universally optimal view, but turns complementary views into carriers of different fine-grained motion distinctions.

Calibration with target skeletons. Figure 5 visualizes the two stages of target calibration. The initial prototypes in panel (a) do not always occupy regions well supported by the target skeletons. In panel (b), balanced target centers move them toward the observed target distributions to obtain $\mathbf{p}_y^1$, while $\beta = 0.7$ retains the semantic prior. Panel (c) shows that the five motion views receive different weights across classes. For example, reading and writing emphasize body motion, whereas touch pocket relies more strongly on action identity. Target skeletons therefore calibrate both the location of each constellation and the composition of its motion views.



(a) Before seen-skeleton grounding (b) After seen-skeleton grounding (c) View-specific pairwise accuracy changes

Figure 4: Seen-skeleton grounding of semantic constellations. Panels (a) and (b) show selected confusion pairs before and after grounding. Panel (c) reports view-specific pairwise accuracy changes. Outlined rows correspond to the visualized pairs, and stars mark the largest improvement.



(a) Initial prototypes $\mathbf{p}_y^0$ (b) Location calibration $\mathbf{p}_y^0 \rightarrow \mathbf{p}_y^1$ (c) Motion-view weights ρ_y^r

Figure 5: Calibration with target skeletons on NTU-60 under the 55/5 split. Panels (a) and (b) share the same t-SNE embedding, with arrows indicating the updates from $\mathbf{p}_y^0$ to $\mathbf{p}_y^1$. Panel (c) reports the weights assigned to the five motion views.

5 Conclusion

We revisited semantic prototype construction in zero-shot skeleton action recognition and showed that the prevailing generate-then-encode pipeline is not neutral, as its prototypes depend on stochastic LLM generation and auxiliary encoder bias. MotionProbe directly reads prompt-conditioned hidden states from a frozen LLM to form multi-view semantic constellations. Seen skeletons ground them in discriminative motion geometry, while target skeletons calibrate their locations and view compositions. Experiments across ZSL and GZSL benchmarks validate this design and reinforce semantic construction as a central problem in ZSSAR.

AI USE STATEMENT

Generative AI assistance in preparing this paper was limited to polishing English expression, assisting with LaTeX and reference formatting, and drafting conference submission statements from author-provided information and the existing manuscript. It was not used to develop research ideas or methods, design experiments, implement the methods, process experimental data, or analyze and interpret results. The use of pretrained language models as research components, including the generation-based comparison experiments, is described in the main text and supplementary material. The authors take responsibility for the final content of this paper.

ETHICS STATEMENT

This work evaluates skeleton action recognition using the existing public NTU RGB+D 60, NTU RGB+D 120, and PKU-MMD datasets. Skeleton representations reduce the reliance on appearance information but do not eliminate privacy or misuse risks. In particular, applications involving monitoring people should account for consent, privacy, and potential performance differences across populations and deployment settings. Dataset reuse and any future deployment should respect the applicable data access terms and appropriate privacy safeguards.

REPRODUCIBILITY STATEMENT

Section 3 specifies the semantic probing, constellation grounding, optimization objectives, and target calibration procedure. Section 4.1 describes the datasets, evaluation protocols, metrics, and core implementation settings. Section D of the supplementary material provides the exact prompts, model configuration, optimization and inference settings, and hardware used for the supplemental experiments. Sections B and C of the supplementary material describe the semantic perturbation and target class imbalance studies. These details document the experimental setup used in this work.

REFERENCES

Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Computer Vision – ECCV 2016*, volume 9906 of *Lecture Notes in Computer Science*, pp. 52–68. Springer, 2016. doi: 10.1007/978-3-319-46475-6_4.

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. Association for Computing Machinery, 2024. doi: 10.1145/3664647.3681196.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8721–8730, 2025.

Yang Chen, Miaoge Li, Zhijie Rao, Deze Zeng, Song Guo, and Jingcai Guo. Learning by neighbor-aware semantics, deciding by open-form flows: Towards robust zero-shot skeleton action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 3374–3383, 2026.

Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 183–192, 2020.

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.

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*, pp. 439–443. IEEE, 2021. doi: 10.1109/ICIP42928.2021.9506179.

Jidong Kuang, Hongsong Wang, Chaolei Han, Yang Zhang, and Jie Gui. Zero-shot skeleton-based action recognition with dual visual-text alignment. *Pattern Recognition*, 171:112342, 2026. doi: 10.1016/j.patcog.2025.112342.

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 *Computer Vision – ECCV 2024*, volume 15074 of *Lecture Notes in Computer Science*, pp. 447–462. Springer, 2024. doi: 10.1007/978-3-031-72640-8_25.

Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding, 2017.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020. doi: 10.1109/TPAMI.2019.2916873.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8247–8255, 2019.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1010–1019, 2016.

Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3571–3580, 2017.

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.

Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 450–459, 2019.

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.

Haojun Xu, Yan Gao, Jie Li, and Xinbo Gao. An information compensation framework for zero-shot skeleton-based action recognition. *IEEE Transactions on Multimedia*, 27:4882–4894, 2025. doi: 10.1109/TMM.2025.3543004.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 7444–7452, 2018. doi: 10.1609/aaai.v32i1.12328.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger

Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

Kai Zhou, Shuhai Zhang, Zeng You, Jinwu Hu, Mingkui Tan, and Fei Liu. Zero-shot skeleton-based action recognition with prototype-guided feature alignment. *IEEE Transactions on Image Processing*, 34:4602–4617, 2025. doi: 10.1109/TIP.2025.3586487.

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. Association for Computing Machinery, 2023. doi: 10.1145/3581783.3611888.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025b.

SUPPLEMENTARY MATERIAL

A OVERVIEW

This supplement extends the main paper with three studies of semantic construction. First, we instantiate the conventional generation and encoding pipeline with a recent model, then vary its action descriptions to reveal how the semantic interface reshapes recognition and prototype geometry. Second, we rewrite MotionProbe's shared prompts to examine the control offered by direct probing. Third, we vary the target class distribution to evaluate calibration beyond a balanced setting. We also provide the exact prompts, core implementation settings, and experimental protocols.

B STABILITY OF SEMANTIC INTERFACES

B.1 GENERATION AND ENCODING PIPELINE

The main paper shows that meaning preserving changes to action descriptions can alter recognition. We examine this sensitivity more broadly using the released Flora design (Chen et al., 2026) as a representative implementation of the common generation and encoding pipeline. For each NTU RGB+D split, we retrain Flora for every semantic set. Its architecture, including the CLIP encoder (Radford et al., 2021), and all optimization settings remain unchanged, so the descriptions are the only experimental factor.

Repeated generation independently samples five description sets from the same instruction:

```
For the action class [CLASS], write exactly one concise
English sentence that preserves the ordinary action meaning
and describes the visible human motion.  Output only the
sentence.
```

Controlled rewriting applies lexical, syntactic, and concise rewrites that preserve the visible motion facts of each reference description.

Table 6 reports the resulting accuracy variation and geometry preservation, together with MotionProbe under the same semantic perturbations.

Table 6: Stability of semantic interfaces. Δ is the accuracy range and ρ is geometry correlation. $\star$ denotes our direct LLM probing method. The best result is bold.

	55/5		48/12	
Interface	$\Delta\downarrow$	$\rho\uparrow$	$\Delta\downarrow$	$\rho\uparrow$
Generation	8.70	0.810	11.61	0.752
Rewrite	5.26	0.957	3.34	0.937
MotionProbe*	**2.73**	**0.984**	**2.07**	**0.990**

Repeated generation yields ranges of 8.70 and 11.61 points on 55/5 and 48/12, compared with 5.26 and 3.34 under controlled rewriting. It also preserves less of the pairwise class geometry. On 55/5, the rewrites raise *reading* from 50.2% to 67.0–71.4%, while *writing* falls from 71.7% to 51.5–53.7%. Small linguistic changes can therefore reorganize competition among the complete set of unseen prototypes.

B.2 DIRECT PROBING WITH A SHARED PROMPT INTERFACE

We next examine the control offered by direct probing. We apply lexical, syntactic, and compact rewrites to the five shared motion queries while preserving their semantic roles. We keep the class anchor, semantic adapter, projector, and skeleton features unchanged. The semantic confusion graph and relation-guided discriminative transform are recomputed from the revised constellations, and performance is measured before target calibration.

The MotionProbe row in Table 6 completes this comparison. Across the three rewrites, MotionProbe spans only 2.73 points on 55/5 and 2.07 points on 48/12. Its geometry correlations reach 0.984 and 0.990. The conventional pipeline generates a separate description for every class, whereas MotionProbe applies one shared change across the entire label space. This global interface turns many independent linguistic choices into one controllable design choice and preserves the semantic geometry more consistently.

Repeated extraction with the standard prompts produces identical hidden states and predictions. Direct probing therefore removes sampling variation once the shared interface is fixed.

C Robustness of Target Calibration to Class Imbalance

The main paper uses balanced target assignments to calibrate each candidate constellation. Since Sinkhorn calibration assumes balanced target assignments, we examine whether its benefit persists when the available target samples are substantially imbalanced on the more challenging NTU-60 48/12 split.

Table 7 summarizes the uncalibrated baseline and the calibrated results under four class frequency ratios.

Table 7: Target calibration under class imbalance on NTU-60 with the 48/12 split. Each ratio denotes the maximum to minimum class frequency and averages 60 calibration subsets of 1,252 skeletons.

Setting	Accuracy	Gain
Before calibration	62.84	–
1:1	75.89	+13.05
2:1	74.29	+11.45
5:1	72.18	+9.34
10:1	70.03	+7.19

We construct calibration subsets with maximum to minimum class frequency ratios of 1, 2, 5, and 10. Every subset contains 1,252 skeletons. Each frequency profile is rotated over all 12 unseen classes, with five subsets sampled for every rotation. The resulting prototypes are evaluated on the same test set. We report macro accuracy to factor out the altered class prior.

Table 7 shows a clear and gradual response to target imbalance. Calibration raises macro ZSL accuracy from 62.84% to 75.89% under the balanced ratio. Its gain remains 11.45 points at 2:1 and 9.34 points at 5:1. Even at 10:1, it retains a 7.19 point improvement. The target skeletons therefore provide useful support even under marked imbalance. In particular, the 70.03% accuracy at the extreme 10:1 ratio remains strong, showing that the calibration remains effective under severe class imbalance.

D Implementation and Reproducibility Details

D.1 Direct probing and prompt interface

We use the frozen Qwen3-4B model (Yang et al., 2025). For each prompt, the semantic evidence is the hidden state of the final input token from the penultimate layer. The reference prompt is

```
Skeleton action: [CLASS]. Key visible body motion:
```

and the five standard motion prompts are

```
Skeleton action: [CLASS]. Core action identity:
Skeleton action: [CLASS]. Directly observable joint motion
and moving body regions:
Skeleton action: [CLASS]. Temporal order, duration, and
repetition pattern:
Skeleton action: [CLASS]. Motion trajectory, direction,
speed, and amplitude:
Skeleton action: [CLASS]. Coordination and relative
movement among body regions:
```

Here, `[CLASS]` is the dataset class label. No description is generated and no auxiliary text encoder is invoked. Table 8 lists the exact alternative suffixes used in the shared prompt study. The class name, anchor, and prompt order remain unchanged.

Table 8: Prompt suffixes used in the shared prompt study. Each suffix follows a class name.

View	Standard	Lexical	Syntactic	Compact
Identity	Core action identity	Essential action meaning	Considering the action, its central action identity is	Identity
Body	Directly observable joint motion and moving body regions	Visible joint movements and participating body parts	Considering the action, its directly visible joint movements and active body regions are	Visible joints and body parts
Temporal	Temporal order, duration, and repetition pattern	Motion sequence, duration, and recurring pattern	Considering the action, its temporal order, duration, and repeated motions are	Temporal pattern
Kinematic	Motion trajectory, direction, speed, and amplitude	Movement path, orientation, velocity, and extent	Considering the action, its movement path, direction, speed, and amplitude are	Trajectory and dynamics
Coordination	Coordination and relative movement among body regions	Coordinated and relative motions between body regions	Considering the action, the coordination and relative motion between body regions are	Body coordination

D.2 Optimization and inference

Skeleton sequences are sampled to 64 frames and encoded by a Shift-GCN (Cheng et al., 2020) pretrained only on seen classes. The skeleton encoder and LLM remain frozen. The semantic adapter maps the LLM hidden dimension to 768, and the projector maps 768 to the 256-dimensional skeleton space. The trainable modules are optimized with AdamW using a batch size of 64, a learning rate of 10^{-3}, and a weight decay of 10^{-4}. We use five motion views, confusion neighborhood size $K = 8$, and $\lambda_h = 2$.

At inference, Sinkhorn uses 50 iterations. The top 30% assignments per class are averaged to estimate its target center. We use one refinement round and target contribution $\beta = 0.7$.

D.3 Representative pipeline implementation

We instantiate the generation and encoding pipeline with the released Flora configuration for each NTU-60 split. Flora is retrained for every reference, generated, and rewritten semantic set. The training schedule, latent noise, and inference conditions remain identical across comparisons.

All supplemental experiments were conducted on one NVIDIA GeForce RTX 4090.