Saber: Anchoring Semantics to Scale-Aware Kinetic Salience for Zero-Shot Skeleton Action Recognition

Yongquan Zhu, Biru Ning, and Jingyu Zhang*

School of Computer Science and Technology, Changsha University of Science and Technology
{yongquanzhu,biruning,zhangzhang}@csust.edu.cn

Abstract. Zero-Shot Skeleton Action Recognition (ZSAR) aims to recognize unseen action categories by establishing a generalizable mapping between skeletal kinematics and semantic representations. However, existing methods frequently struggle with two fundamental issues: (i) the loss of high-frequency dynamics caused by recursive feature aggregation, and (ii) rigid semantic anchoring, which leaves cross-modal alignments highly vulnerable to semantically irrelevant background noise. To overcome these limitations, we propose the **S**cale-**A**ware **B**ipartite **E**nergy-guided **R**egistration (**Saber**) framework, a unified kinematics-driven paradigm for robust cross-modal alignment. Specifically: (i) Scale-Aware Bipartite Encoder actively reshapes the spatio-temporal topology and applies instance-adaptive filtering to effectively isolate high-frequency motion details, thereby preserving critical structural dynamics against "spectral smoothing". (ii) A Kinetically Driven Probabilistic Alignment strategy, comprising energy-guided dynamic focusing and uncertainty-aware probabilistic anchoring, maps part-level representations into variance-adaptive distributions. This mechanism dynamically resolves structural uncertainties and robustly anchors physical movements to explicit semantics. Extensive evaluations on various benchmark datasets validate that **Saber** achieves state-of-the-art performance, demonstrating exceptional robustness and generalization.

Keywords: Zero-Shot Skeleton Action Recognition · Cross-Modal Alignment · Spatio-Temporal Modeling

1 Introduction

Human Action Recognition serves as a cornerstone of human-centric computing, facilitating diverse applications from human-computer interaction to medical rehabilitation. Among various input modalities, skeletal data is widely adopted due to its compact structure and inherent robustness to environmental noise, such as background clutter and illumination changes. Nevertheless, the scalability of traditional supervised paradigms is hindered by the prohibitive cost of

* Corresponding author.

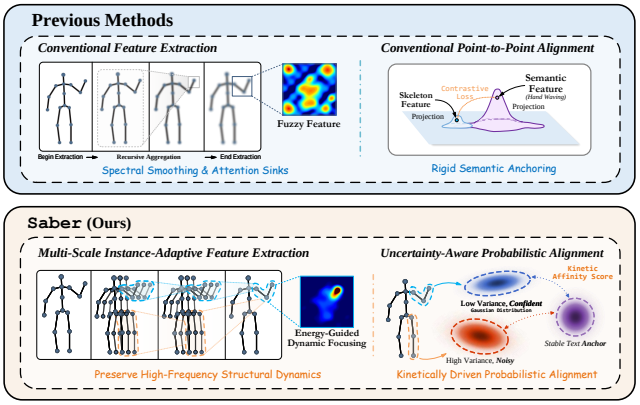

Fig. 1: Overview of our **Saber** *vs.* previous methods.

exhaustively annotating the unbounded action space in real-world scenarios. To overcome this reliance on labeled data, Zero-Shot Skeleton Action Recognition (ZSAR) emerges as a critical solution. It aims to generalize to unseen categories by establishing a transferable mapping between visual and semantic spaces.

Current mainstream ZSAR methods tackle the cross-modal semantic gap primarily by exploring ***embedding-based***, ***generative-based***, and ***transport-based*** paradigms. While these approaches have significantly improved recognition accuracy on unseen categories, they still suffer from two fundamental issues. (i) **Loss of High-Frequency Dynamics.** For effective cross-modal transfer, models demand highly discriminative motion representations. However, widely adopted pre-trained Graph Convolutional Networks (*e.g.*, ST-GCN [30], Shift-GCN [6]) suffer from "spectral smoothing" during recursive aggregation, severely diluting critical high-frequency details. Additionally, conventional implicit attention mechanisms frequently fall into "attention sinks", erroneously prioritizing background noise over the primary moving subject. (ii) **Rigid Semantic Anchoring.** Effective cross-modal alignment demands precise mapping between part-level motion details and semantic descriptions. However, existing approaches frequently rely on rigid, "point-to-point" static feature projections. These paradigms ignore that active limbs convey far richer semantic cues than static ones. Alternatively, while some methods employ generative models (*e.g.*, Diffusion [7], Flow Matching [5]) to fit global data distributions, these architectures might incur prohibitive computational overhead.

To overcome these dynamic and semantic limitations, we propose **S**cale-**A**ware **B**ipartite **E**nergy-guided **R**egistration (**Saber**) paradigm, a unified, kinematics-driven ZSAR framework (Fig. 1). By integrating multi-scale spatio-temporal reshaping with energy-guided probabilistic distribution alignment, our approach effectively anchors complex skeletal dynamics to explicit textual semantics. Specifically: (i) **To preserve high-frequency dynamics**, we introduce the Scale-Aware Bipartite Encoder. Instead of relying on recursive aggregation that causes spectral smoothing, this module actively reconstructs the spatio-temporal topology via a multi-scale temporal folding mechanism. Coupled with

a multi-head instance-adaptive high-pass filter, it effectively isolates and preserves critical motion details and interactive dynamics from background noise. (ii) **To overcome rigid semantic anchoring**, we propose a Kinetically Driven Probabilistic Alignment strategy. Comprising energy-guided dynamic focusing and uncertainty-aware probabilistic anchoring, this mechanism maps part-level representations into variance-adaptive Gaussian distributions based on their kinetic activity. By utilizing structural *energy* as a reliability indicator, it explicitly focuses the cross-modal alignment on semantically meaningful movements, effectively overcoming the limitations of static "point-to-point" matching.

The main contributions of this work are summarized as follows:

- We propose the Scale-Aware Bipartite Encoder. By applying an instance-adaptive high-pass filter to the multi-scale reshaped spatio-temporal topology, it effectively isolates high-frequency motion details and preserves structural dynamics against spectral smoothing.
- We formulate a Kinetically Driven Probabilistic Alignment strategy. By integrating energy-guided dynamic focusing with uncertainty-aware probabilistic anchoring, it dynamically resolves part-level uncertainties to robustly anchor physical movements to explicit semantic concepts.
- Extensive evaluations on the NTU RGB+D, NTU RGB+D 120, and PKU-MMD benchmarks demonstrate that **Saber** consistently surpasses state-of-the-art approaches, exhibiting superior performance.

2 Related Work

2.1 Zero-Shot Skeleton Action Recognition

Current ZSAR methods primarily bridge the cross-modal semantic gap through three paradigms: *embedding-based*, *generative-based*, and *transport-based*.

Embedding-Based Paradigm. Most existing frameworks project visual kinematics and textual semantics into a shared latent space to perform distance-based retrieval. Early attempts like ReViSE [22] and JPoSE [24] aligned global sequence features, while recent works such as PURLS [33] and STAR [4] explore fine-grained, part-level alignments to capture local correspondences.

Generative-Based Paradigm. To address the extreme lack of unseen visual samples, generative approaches synthesize pseudo-skeletal features from semantic prototypes. Models utilizing Variational Autoencoders (VAEs), such as CADA-VAE [20], SynSE [9], and SA-DVAE [15], successfully reformulate ZSAR into a fully supervised classification task by augmenting the training space.

Transport-Based Paradigm. Recently, transport-based methods have emerged to model cross-modal alignment as a continuous distribution matching problem driven by advanced dynamical systems. Specifically, TDSM [7] employs diffusion models to iteratively denoise semantic priors into visual feature spaces, while Flora [5] leverages open-form flow matching to construct optimal transport paths between multi-modal distributions.

Unlike prior paradigms, our **Saber** integrates scale-aware spatio-temporal reshaping with kinetic-driven probabilistic anchoring, avoiding the rigidity of embedding methods and the computational overhead of generative or transport models, while preserves high-frequency dynamics through discriminative spatio-temporal modeling.

2.2 Cross-Modal Alignment

Cross-modal alignment seeks to connect visual data, such as skeletal sequences or videos, with textual descriptions to support tasks like action recognition and retrieval. Common approaches include contrastive learning, prototype matching, and granularity- or event-based methods.

Contrastive paradigms often use relation-aware techniques to manage noise in alignments. For example, some studies build proxy features for ranking distributions in video-text retrieval [13], while others apply contrastive features for domain adaptation in action recognition [11]. Additionally, interaction-first strategies precede alignment in cross-domain settings [31], and dual alignments handle unsupervised adaptations [10]. Prototype methods focus on collaborative representations, such as virtual text synthesis with relation consistency [29], or multi-modal alignments to reduce domain shifts [27]. Granularity-aware approaches unify multi-level similarities [23], incorporate linguistic patch slimming [8], or model event logic [18].

However, these methods tend to rely on static projections, which may limit handling of motion uncertainties and dynamic details. In contrast, our **Saber** framework integrates scale-aware reshaping with energy-driven probabilistic strategies for robust semantic connections.

3 Methodology

3.1 Problem Definition

Let $\mathcal{Y}$ denote the action category space, partitioned into disjoint seen ($\mathcal{Y}_s$) and unseen ($\mathcal{Y}_u$) sets. During training, we utilize a labeled dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N_s}$, where each skeleton sequence $\mathbf{X}_i \in \mathbb{R}^{T\times N\times C}$ is annotated with a seen label $y_i \in \mathcal{Y}_s$. To bridge the supervision gap, we leverage text-derived semantic prototypes $\mathcal{Z}_{\text{sem}} = \{\mathbf{e}_c \mid c \in \mathcal{Y}\}$ (*e.g.*, via CLIP [19]) to learn a generalized cross-modal alignment mapping.

Inference is evaluated under two settings: (i) Zero-Shot Learning (ZSL), restricting the search space strictly to $\mathcal{Y}_u$ on a test set $\mathcal{D}^u_{\text{test}}$; and (ii) Generalized Zero-Shot Learning (GZSL), expanding the search space to the unified set $\mathcal{Y}_s \cup \mathcal{Y}_u$ to assess unbiased cross-modal transferability.

3.2 High-Frequency Structural Feature Extraction

Conventional graph convolutions often face the issue of *spectral smoothing* due to recursive feature aggregation, which progressively suppresses the high-frequency

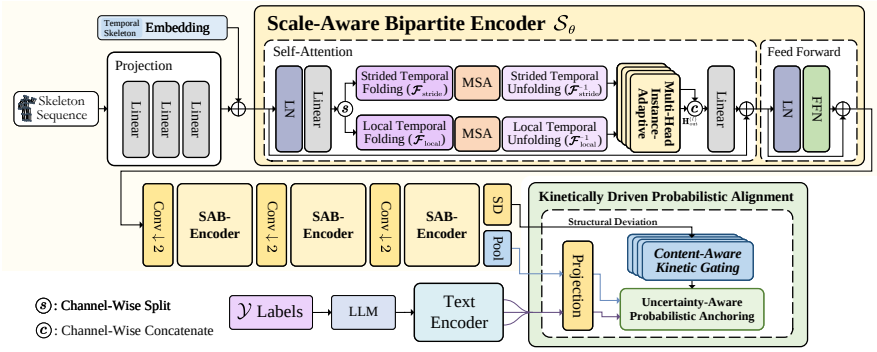

Fig. 2: Training framework of our proposed **Saber** for ZSAR.

motion details essential for accurate classification. As illustrated in Figure 2, we propose the Scale-Aware Bipartite Encoder $\mathcal{S}_\theta$ (**SAB-Encoder**). Instead of passive aggregation, it reshapes spatio-temporal topologies to preserve dynamics.

Semantic Anchoring and Skeletal Feature Initialization. To resolve the ambiguity of discrete action labels, we use GPT-4o to generate *fine-grained* descriptions for each category. A pre-trained text encoder (*e.g.*, CLIP [19]) then projects these detailed body movement descriptions into a shared semantic space, providing robust anchors for cross-modal alignment.

For the visual input, the raw skeleton sequence $\mathbf{X}_{\text{raw}} \in \mathbb{R}^{B\times T\times M\times N\times C}$ (M is the number of individuals, N is the number of skeletal joints) is temporally interpolated to a fixed length T' and linearly projected into a D-dimensional space, yielding $\mathbf{X}_p$. The initial representation $\mathbf{H}$ is then formulated by explicitly injecting learnable structural ($\mathbf{E}_{\text{skel}}$) and fixed temporal ($\mathbf{E}_{\text{temp}}$) embeddings:

$$\mathbf{H} = \mathbf{X}_p + \mathbf{E}_{\text{skel}} + \mathbf{E}_{\text{temp}} \quad \in \mathbb{R}^{B\times T'\times M\times N\times D} \tag{1}$$

To evenly divide the N joints into N_p partitions of size S_p (*e.g.*, $N = 25, S_p = 6$ [1]), we pad the spatial dimension to N' using virtual nodes. We then apply a mask $\mathbf{M}^{\text{pad}} \in \{0, -\infty\}^{N'\times N'}$ to ignore them during attention:

$$\mathbf{M}^{\text{pad}}_{ij} = \begin{cases} 0, & \text{if } j \le N \\ -\infty, & \text{if } j > N \end{cases} \tag{2}$$

Multi-Scale Temporal Folding. Effective motion modeling presents a temporal *trade-off*: high-frequency sampling captures rapid limb movements, while long-range stability models global coordination.

To explicitly decouple multi-scale dynamics, the multi-scale temporal folding mechanism divides the initialized $\mathbf{H}$ channel-wise into $\mathbf{H}_{\text{long}}$ and $\mathbf{H}_{\text{short}}$. Governed by a temporal factor K_t, two complementary operators physically reconstruct the temporal topology: (i) **Local Temporal Folding** ($\boldsymbol{\mathcal{F}}_{\text{local}}$) reorganizes the sequence T' into a windowed view $(T'/K_t) \times K_t$. By applying Multi-Head Self-Attention (MSA) within these local windows ($K_t \times N'$), this branch aggregates rapid micro-dynamics without diluting high-frequency details. (ii) **Strided**

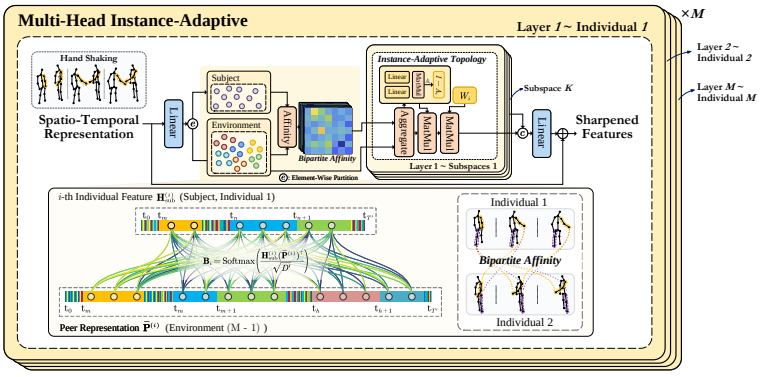

Fig. 3: Architecture of the Multi-Head Instance-Adaptive paradigm.

Temporal Folding ($\boldsymbol{\mathcal{F}}_{\text{stride}}$) permutes T' into a dilated view $K_t \times (T'/K_t)$. Consequently, MSA can efficiently capture the low-frequency macro-evolution of the action across the global timeline ($T'/K_t \times N'$).

To maintain structural integrity, the spatial mask $\mathbf{M}^{\text{pad}}$ is temporally expanded to strictly exclude virtual nodes within both branches. Subsequently, the processed representations are reverted to their original temporal resolution via their respective inverse operators ($\boldsymbol{\mathcal{F}}^{-1}_{\text{local/stride}}$). Comprehensive partitioning details are provided in the *Supplementary Material*.

$$\begin{cases} \hat{\mathbf{H}}_{\text{long}} \leftarrow \boldsymbol{\mathcal{F}}^{-1}_{\text{stride}}(\text{MSA}(\boldsymbol{\mathcal{F}}_{\text{stride}}(\mathbf{H}_{\text{long}}), \mathbf{M}^{\text{pad}})) \\ \hat{\mathbf{H}}_{\text{short}} \leftarrow \boldsymbol{\mathcal{F}}^{-1}_{\text{local}}(\text{MSA}(\boldsymbol{\mathcal{F}}_{\text{local}}(\mathbf{H}_{\text{short}}), \mathbf{M}^{\text{pad}})) \end{cases} \in \mathbb{R}^{B\times T'\times M\times N'\times (D/2)} \quad (3)$$

To synthesize the extracted multi-scale dynamics, we concatenate the restored features from both branches along the channel dimension, yielding the unified spatio-temporal representation $\mathbf{H}' = [\hat{\mathbf{H}}_{\text{long}}; \hat{\mathbf{H}}_{\text{short}}] \in \mathbb{R}^{B\times T'\times M\times N'\times D}$. where $[\cdot;\cdot]$ denotes the concatenation operation. This fused representation $\mathbf{H}'$ serves as the input for the subsequent partition-level interaction modeling.

Multi-Head Instance-Adaptive Sharpening. To capture structural dependencies beyond fixed topologies, we introduce the Multi-Head Instance-Adaptive (MI) paradigm (Fig. 3). Initially, to transition from joint-level to part-level semantics, we average the valid nodes within each of the N_p partitions in $\mathbf{H}'$, strictly excluding the padded virtual nodes. This yields the subject-specific partition representation $\mathbf{H}_{\text{sub}} \in \mathbb{R}^{B\times T'\times M\times N_p\times D}$.

Next, to explicitly model *subject-environment* interactions, we construct a peer representation $\bar{\mathbf{P}}^{(i)}$ for the i-th individual by concatenating the partition features of all other participants:

$$\bar{\mathbf{P}}^{(i)} = \begin{cases} \text{Concat}\left(\left\{\mathbf{H}^{(j)}_{\text{sub}}\right\}_{j\neq i}\right), & \text{if } M > 1 \\ \mathbf{0}, & \text{if } M = 1 \end{cases} \in \mathbb{R}^{B\times T'\times(\max(1,M-1)N_p)\times D} \quad (4)$$

where $\mathbf{H}^{(j)}_{\text{sub}}$ denotes the j-th individual's feature, and $\mathbf{0}$ ensures dimensional stability for single-person scenarios. Subsequently, we compute the semantic affinity

matrix $\mathbf{B}_i$ via a scaled dot-product:

$$\mathbf{B}_i = \text{Softmax}\left(\frac{\mathbf{H}_{\text{sub}}^{(i)}(\bar{\mathbf{P}}^{(i)})^\top}{\sqrt{D}}\right) \in \mathbb{R}^{B\times T'\times N_p\times(\max(1,M-1)N_p)} \tag{5}$$

Crucially, $\mathbf{B}_i$ functions as a bipartite graph, structurally mapping the individual's body parts to the collective peer context. We aggregate the raw interaction context using this bipartite affinity to obtain the base representation $\mathbf{Z}^{(i)} = \mathbf{B}_i\bar{\mathbf{P}}^{(i)} \in \mathbb{R}^{B\times T'\times N_p\times D}$.

The system splits $\mathbf{Z}^{(i)}$ channel-wise into K subspaces (*heads*), denoted as $\mathbf{Z}_k^{(i)} \in \mathbb{R}^{B\times T'\times N_p\times d_k}$ ($d_k = D/K$). This separation learns complementary topologies, whereas $K = 1$ forces all channels to share one topology and may mix distinct motion patterns.

$$\mathbf{A}_i^{(k)} = \text{Softmax}\left(\frac{\phi_k(\mathbf{Z}_k^{(i)})\cdot\psi_k(\mathbf{Z}_k^{(i)})^\top}{\sqrt{d_k}}\right) \quad \in \mathbb{R}^{B\times T'\times N_p\times N_p} \tag{6}$$

where ϕ_k and ψ_k are linear projections. The normalized affinities aggregate structurally related signals, forming a smoothed low-frequency reference. We then extract the residual:

$$\mathbf{H}_k^{(i)} = (\mathbf{I} - \mathbf{A}_i^{(k)})\mathbf{Z}_k^{(i)}\mathbf{W}_i^{(k)} \quad \in \mathbb{R}^{B\times T'\times N_p\times d_k} \tag{7}$$

where $\mathbf{I}$ is the identity matrix. Acting as a graph Laplacian variant, the operator $(\mathbf{I} - \mathbf{A}_i^{(k)})$ effectively functions as an adaptive structural high-pass filter. Moreover, $\mathbf{W}_i^{(k)} \in \mathbb{R}^{d_k\times d_k}$ represents the subspace-specific weight matrix.

Ultimately, we concatenate and linearly project the sharpened representations from all K heads to reconstruct the full feature space:

$$\mathbf{H}^{(i)} = \text{Concat}\left(\mathbf{H}_1^{(i)},\ldots,\mathbf{H}_K^{(i)}\right)\mathbf{W}_O \quad \in \mathbb{R}^{B\times T'\times N_p\times D} \tag{8}$$

where $\mathbf{W}_O \in \mathbb{R}^{D\times D}$ fuses the information across different subspaces. We perform this multi-head sharpening in parallel for all M individuals to capture scene-wide interactions. To restore the global topology, we aggregate the sharpened features $\{\mathbf{H}^{(i)}\}_{i=1}^{M}$ into a unified representation and integrate a residual connection:

$$\mathbf{H}_{\text{out}}^{(l)} = \boldsymbol{\mathcal{E}}\left(\text{Concat}\left(\mathbf{H}^{(1)},\ldots,\mathbf{H}^{(M)}\right)\right) + \mathbf{H}' \quad \in \mathbb{R}^{B\times T'\times M\times N'\times D} \tag{9}$$

where l indicates the current layer index, and $\boldsymbol{\mathcal{E}}(\cdot)$ is a linear projection restoring partition features back to the N' joint level. The reconstructed manifold $\mathbf{H}_{\text{out}}^{(l)}$ effectively preserves both the original spatial layout and the enhanced high-frequency cues, serving as the robust input for the subsequent layer.

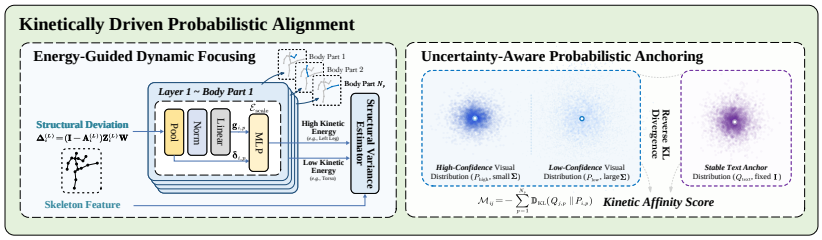

Fig. 4: Detailed structure of the kinetically driven probabilistic alignment.

3.3 Kinetically Driven Probabilistic Alignment

Physically, active limbs convey far richer semantic cues than static ones, as higher kinetic energy directly correlates with semantic significance. However, conventional methods frequently rely on rigid semantic anchoring, ignoring that an action's core meaning is fundamentally driven by explicit *kinetic activity*. To leverage this, we propose a kinetically driven probabilistic alignment strategy (Fig. 4). This mechanism dynamically resolves part-level uncertainties, robustly anchoring physical movements to explicit semantic concepts while naturally filtering out background noise.

Structural Deviation Response. At the final layer L, let $\mathbf{Z}_i^{(L)} = \mathbf{B}_i^{(L)}\bar{\mathbf{P}}^{(i,L)}$ denote the affinity-aggregated interaction context for the i-th individual. Following the sharpening operation in Eq. (7), we extract the Structural Deviation (SD) $\mathbf{\Delta}_i^{(L)}$ to highlight distinct motions:

$$\mathbf{\Delta}_i^{(L)} = (\mathbf{I} - \mathbf{A}_i^{(L)})\mathbf{Z}_i^{(L)}\mathbf{W}_L \in \mathbb{R}^{B\times T'\times N_p\times D} \tag{10}$$

Energy-Guided Dynamic Focusing. In skeleton sequences, high-frequency jitter or irrelevant background movements frequently display large structural deviations (high ℓ_2-norm) despite lacking semantic significance. Left unaddressed, these high-energy artifacts act as "attention sinks" [26], misleading the model to focus on noisy regions.

To overcome this limitation, we design a content-aware kinetic gating mechanism as the core component of our energy-guided dynamic focusing strategy to enforce kinetic sparsity. Initially, we aggregate the temporal dynamics of the structural deviation $\mathbf{\Delta}_{i,p}^{(L)}$ into a static descriptor $\delta_{i,p}$ via global average pooling across the temporal dimension T':

$$\delta_{i,p} = \frac{1}{T'}\sum_{t=1}^{T'} \mathbf{\Delta}_{i,p}^{(L)}[t] \quad \in \mathbb{R}^{B\times D} \tag{11}$$

Importantly, to decouple semantic validity from raw motion intensity, we apply Layer Normalization to this kinetic descriptor:

$$\hat{\delta}_{i,p} = \mathrm{LayerNorm}\left(\delta_{i,p}\right) \tag{12}$$

This normalization eliminates magnitude bias, directing the semantic validator to focus exclusively on the directional motion patterns. Next, we compute content-adaptive gate $\mathbf{g}_{i,p} \in (0,1)^{B\times D}$ based on this representation:

$$\mathbf{g}_{i,p} = \sigma(\mathbf{W}_g \hat{\delta}_{i,p} + \mathbf{b}_g) \tag{13}$$

This gate is then applied to the *original* unnormalized descriptor via element-wise multiplication ($\odot$) to filter out invalid motion channels. Ultimately, the energy-guided attention weight $w_{i,p} \in (0,1)^B$ is derived from the magnitude of the purified representation:

$$w_{i,p} = \sigma\left(\mathcal{E}_{\text{scale}}\left(\|\delta_{i,p} \odot \mathbf{g}_{i,p}\|_2\right)\right) \tag{14}$$

Here, $\|\cdot\|_2$ denotes the ℓ_2-norm, and $\mathcal{E}_{\text{scale}}$ is a parameterized linear mapping. This mechanism ensures that the cross-modal alignment is anchored strictly to semantically significant structural transformations.

Uncertainty-Aware Probabilistic Anchoring. Skeletal movements inherently contain uncertainty: active parts convey explicit semantics, while static regions introduce noise. To overcome the limitations of rigid point-to-point projections, we map feature embeddings into probability distributions, using kinetic salience to indicate certainty.

To formalize this probabilistic mapping, we introduce a structural variance estimator, which models the p-th partition of the i-th skeleton as a multivariate Gaussian $P_{i,p} = \mathcal{N}(\boldsymbol{\mu}_{i,p}, \boldsymbol{\Sigma}_{i,p})$. The mean vector $\boldsymbol{\mu}_{i,p} = \mathbf{z}_{i,p}$ is directly extracted from the interaction context $\mathbf{Z}_i^{(L)}$. Importantly, we utilize the energy gate $w_{i,p}$ as an inverse indicator of uncertainty. Therefore, the covariance matrix $\boldsymbol{\Sigma}_{i,p}$ exponentially scales the variance based on this kinetic energy:

$$\boldsymbol{\Sigma}_{i,p} = (\exp(\alpha \cdot (1 - w_{i,p})) + \epsilon)\,\mathbf{I} \tag{15}$$

where α controls distribution sharpness and ϵ ensures numerical stability. Naturally, highly active parts ($w_{i,p} \approx 1$) yield a low variance approaching $1 \cdot \mathbf{I}$, structurally matching the semantic anchor. Conversely, inactive background parts ($w_{i,p} \approx 0$) manifest as diffuse clouds with high variance (*e.g.*, $\exp(\alpha) \cdot \mathbf{I}$), explicitly reflecting low confidence.

Furthermore, we represent text prototypes as standard unit Gaussian $Q_{j,p} = \mathcal{N}(\mathbf{t}_{j,p}, \mathbf{I})$, serving as reliable semantic anchors, where $\mathbf{t}_{j,p}$ is the text embedding for the j-th category. To measure the cross-modal mismatch, we apply the reverse Kullback-Leibler (KL) divergence $\mathbb{D}_{\text{KL}}(Q_{j,p} \| P_{i,p})$. Ultimately, the kinetic affinity score $\mathcal{M}_{ij}$ is defined as the negative sum of these divergences across all partitions:

$$\mathcal{M}_{ij} = -\sum_{p=1}^{N_p} \mathbb{D}_{\text{KL}}(Q_{j,p} \| P_{i,p}) \tag{16}$$

Omitting constant terms, the effective KL divergence is simplified to:

$$\mathbb{D}_{\mathrm{KL}}(Q_{j,p} \| P_{i,p}) = \frac{1}{2}\Bigg[\frac{\ln|\boldsymbol{\Sigma}_{i,p}|}{D} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{i,p}^{-1}\boldsymbol{\Sigma}_q)}{D} - 1 + (\boldsymbol{\mu}_{i,p} - \mathbf{t}_{j,p})^\top \boldsymbol{\Sigma}_{i,p}^{-1}(\boldsymbol{\mu}_{i,p} - \mathbf{t}_{j,p})\Bigg] \tag{17}$$

To maintain dimension-invariance and sensitivity to mean differences in high-dimensional embeddings, we average the log-determinant and trace over D.

This adaptive formulation fundamentally optimizes cross-modal alignment. For kinetically significant parts ($\boldsymbol{\Sigma}_{i,p} \approx 1 \cdot \mathbf{I}$), the distance term dominates, ensuring precise matching. In contrast, for noisy parts with inflated variance, the inverse covariance $\boldsymbol{\Sigma}_{i,p}^{-1}$ naturally suppresses the distance penalty, providing a "soft" mechanism to filter out ambiguity.

Kinetic-Driven Optimization Objective. To optimize the cross-modal semantic matching, we employ the InfoNCE loss formulated over the established kinetic affinity scores. Specifically, given a batch of B skeleton sequences, the contrastive alignment loss $\mathcal{L}_{\mathrm{align}}$ aims to maximize the similarity between the visual sequence and its ground-truth seen category $y_i \in \mathcal{Y}_s$:

$$\mathcal{L}_{\mathrm{align}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\mathcal{M}_{i,y_i}/\tau)}{\sum_{j\in\mathcal{Y}_s}\exp(\mathcal{M}_{i,j}/\tau)} \tag{18}$$

where τ acts as a temperature hyperparameter. Furthermore, we explicitly enforce kinetic sparsity to prevent the underlying gating mechanism (Eq. 14) from degenerating into a trivial solution that outputs uniform values to artificially eliminate variance. We apply an ℓ_1 regularization penalty over the energy gates $w_{i,p}$ across all N_p partitions:

$$\mathcal{L}_{\mathrm{sparse}} = \frac{1}{B \cdot N_p}\sum_{i=1}^{B}\sum_{p=1}^{N_p} w_{i,p} \tag{19}$$

Consequently, the overall objective function for training the **Saber** framework is formulated as the weighted sum of these two components:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{align}} + \lambda\mathcal{L}_{\mathrm{sparse}} \tag{20}$$

where λ is a balancing coefficient regulating the strictness of the kinetic sparsity.

3.4 ZSL/GZSL Prediction

During inference, classification relies on the kinetic affinity scores. (i) For the standard ZSL setting, the search space is restricted to unseen categories ($\mathcal{Y}_u$). Given a test skeleton sequence x, the predicted label $\hat{y}_{\mathrm{ZSL}}$ is obtained by maximizing the affinity score $\mathcal{M}(x, c)$ derived from Eq. 17:

$$\hat{y}_{\mathrm{ZSL}} = \arg\max_{c\in\mathcal{Y}_u} \mathcal{M}(x, c) \tag{21}$$

(ii) For the GZSL setting, the search space expands to encompass all categories ($\mathcal{Y}_{\text{all}} = \mathcal{Y}_s \cup \mathcal{Y}_u$). To counteract the inherent prediction bias toward seen categories caused by training exclusively on $\mathcal{Y}_s$, we employ the Calibrated Stacking (CS) strategy [2]. This approach penalizes the scores of seen categories using a calibration factor γ:

$$\hat{y}_{\text{GZSL}} = \arg\max_{c \in \mathcal{Y}_{\text{all}}} \left(\mathcal{M}(x, c) - \gamma \cdot \mathbb{I}[c \in \mathcal{Y}_s] \right) \tag{22}$$

where $\mathbb{I}[\cdot]$ is the indicator function, and $\gamma \geq 0$ is a hyperparameter empirically tuned to balance the recognition performance across both domains.

4 Experiments

4.1 Datasets

NTU RGB+D [21]. Captured via multi-view Kinect sensors, this benchmark contains 56,880 skeleton sequences distributed across 60 action categories and performed by 40 distinct subjects. We adhere to two standard evaluation protocols: Cross-Subject (X-Sub), which splits the data based on subject identities, and Cross-View (X-View), which divides samples according to camera angles.
NTU RGB+D 120 [17]. Building upon its predecessor, this large-scale extension features 114,480 sequences covering 120 action categories performed by 106 subjects. Performance is evaluated under the official Cross-Subject (X-Sub) and Cross-Setup (X-Set) criteria.
PKU-MMD [16]. Comprising approximately 20,000 continuous multi-modal sequences spanning 51 action categories, this dataset is evaluated using the official Cross-Subject (X-Sub) protocol, allocating 57 subjects for training and the remaining 9 for testing.

4.2 Implementation Details

Implemented in PyTorch on an NVIDIA L20 GPU, we use CLIP-ViT-B/32 for semantic anchoring. Skeletons are interpolated to $T' = 120$ frames, projected to $D = 256$ dimensions, and grouped into $N_p = 5$ partitions. The 5-layer SAB-Encoder ($K = 4$) is trained for 200 epochs (batch size 256, base learning rate 10^{-4} with a 10-epoch warmup). The network is optimized using InfoNCE loss ($\tau = 0.07$) and kinetic sparsity regularization ($\lambda = 0.1$). For probabilistic alignment, distribution sharpness is set to $\alpha = 4.0$. During GZSL inference, the calibration factor is dynamically tuned $\gamma = 1.5$.

4.3 Evaluation Settings

Following established protocols, we evaluate **Saber** under two standard settings. For ZSL, we report the Top-1 accuracy $\text{Acc} = \frac{1}{N_u} \sum_{j=1}^{N_u} \mathbb{I}[\hat{y}_j = y_j]$ strictly within the unseen domain $\mathcal{Y}_u$. For GZSL, the recognition space extends to the joint label set $\mathcal{Y}_s \cup \mathcal{Y}_u$. To comprehensively evaluate cross-modal generalization and mitigate training-domain bias, we report the Top-1 accuracy on seen categories (S), unseen categories (U), and their harmonic mean $H = (2 \times S \times U)/(S + U)$.

Table 1: Top-1 accuracy results of various ZSAR methods on NTU RGB+D and NTU RGB+D 120 datasets under the ZSL and GZSL settings. The **red** and **blue** markers represent the best and second-best values. †: Methods utilizing CLIP as the text encoder. For fair comparisons, the reported results are the average accuracy over ***5*** independent test runs. Results on the X-View and X-Set benchmarks are reported in the *Supplementary Material*.

Methods	Publications	NTU RGB+D (X-Sub)								NTU RGB+D 120 (X-Sub)							
		55/5 Split				48/12 Split				110/10 Split				96/24 Split			
		ZSL	S	U	H	ZSL	S	U	H	ZSL	S	U	H	ZSL	S	U	H
ReViSE [22]	ICCV 2017	53.9	74.2	34.7	47.3	17.5	62.4	20.8	31.2	55.0	48.7	44.8	46.7	32.4	49.7	25.1	33.3
JPoSE [24]	ICCV 2019	64.8	64.4	50.3	56.5	28.8	60.5	20.6	30.8	51.9	47.7	46.4	47.0	32.4	38.6	22.8	28.7
CADA-VAE [20]	CVPR 2019	76.8	69.4	61.8	65.4	29.0	51.3	27.0	35.4	59.5	47.2	49.8	48.4	35.8	41.1	34.1	37.3
SynSE [9]	ICIP 2021	75.8	61.3	56.9	59.0	33.3	52.2	27.9	36.3	62.7	52.5	57.6	54.9	38.7	56.4	32.3	41.0
GZSSAR† [14]	ICIG 2023	83.6	71.7	66.2	68.8	49.2	58.8	40.0	47.6	71.2	46.8	**68.3**	55.5	59.7	56.8	48.6	52.4
SMIE [32]	ACM MM 2023	78.0	-	-	-	40.2	-	-	-	65.7	-	-	-	45.3	-	-	-
PURLS† [33]	CVPR 2024	79.2	-	-	-	41.0	-	-	-	72.0	-	-	-	52.0	-	-	-
SA-DVAE [15]	ECCV 2024	82.4	62.3	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
STAR† [4]	ACM MM 2024	81.4	69.0	69.9	69.4	45.1	**62.7**	37.0	46.5	63.3	59.9	52.7	56.1	44.3	51.2	36.9	42.9
DVTA† [12]	PR 2025	79.6	-	-	-	47.5	-	-	-	64.1	-	-	-	48.2	-	-	-
InfoCPL [28]	TMM 2025	85.9	-	-	-	53.3	-	-	-	**74.8**	-	-	-	60.1	-	-	-
SCoPLe† [34]	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† [3]	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
FS-VAE† [25]	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**
TDSM† [7]	ICCV 2025	86.5	-	-	-	56.0	-	-	-	74.2	-	-	-	**65.1**	-	-	-
Saber †**(Ours)**	This work	**87.5**	**78.7**	73.7	**76.1**	**61.5**	**62.9**	**60.6**	**61.7**	**76.8**	**64.5**	67.4	**65.9**	**65.3**	**59.8**	**55.2**	**57.4**

Table 2: Performance comparisons on PKU-MMD dataset under X-Sub and X-View benchmark.

Methods	Publications	PKU-MMD (X-Sub)								PKU-MMD (X-View)							
		46/5 Split				39/12 Split				46/5 Split				39/12 Split			
		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	39.0	19.3	35.7	13.0	19.1	54.1	50.7	39.9	44.7	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	38.9
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**
Saber (Ours)	This work	**79.3**	**78.8**	**60.8**	**68.6**	**51.6**	**71.9**	**48.3**	**57.8**	73.5	**74.3**	**72.8**	**73.5**	**57.6**	**73.6**	**49.2**	**59.0**

4.4 Performance Comparisons

Standard Benchmark Evaluation. Table 1 and Table 2 present a comprehensive evaluation of **Saber** against state-of-the-art (SOTA) methods across the NTU-60, NTU-120, and PKU-MMD datasets. Our method consistently establishes a clear performance margin under both ZSL and GZSL settings. By averting spectral smoothing and leveraging probabilistic anchoring, our method preserves high-frequency details and filters background noise, achieving robust zero-shot generalization.

Random Split Evaluation. To investigate the framework's resilience against split selection bias, we benchmark our approach under entirely randomized category divisions, as detailed in Table 3. The results demonstrate stability across diverse target configurations. This confirms that our approach captures generalizable structural dynamics rather than overfitting to the training distribution.

Extreme Split Evaluation. Table 6 evaluates the model's structural efficiency and generalization limits under conditions of severe data scarcity. Despite severely reduced seen categories, our approach avoids the rapid degradation of

Table 3: Performance evaluation on random splits. † denotes methods utilizing CLIP as the text encoder.

Methods	NTU RGB+D 55/5 Split		NTU RGB+D 120 110/10 Split		PKU-MMD 46/5 Split	
	ZSL	GZSL	ZSL	GZSL	ZSL	GZSL
ReViSE	60.9	60.3	44.9	40.3	59.3	49.8
JPoSE	59.4	60.1	46.7	43.7	57.2	51.6
CADA-VAE	61.8	66.4	45.2	45.6	60.7	45.8
SynSE	64.2	67.5	47.3	43.5	60.8	49.5
SMIE	65.1	-	46.4	-	60.8	-
SA-DVAE	84.2	75.3	50.7	47.5	66.5	54.7
SCoPLe†	83.7	**77.7**	53.3	**54.1**	**71.4**	54.9
TDSM†	**88.9**	-	**69.5**	-	70.8	-
FS-VAE†	-	-	-	-	71.2	**59.0**
Saber †**(Ours)**	**84.9**	**79.3**	**71.5**	**57.2**	**73.2**	**58.7**

Table 4: Ablation study on NTU-60 and NTU-120 datasets.

Model Variants	NTU RGB+D 55/5 Split		NTU RGB+D 120 110/10 Split	
	ZSL	GZSL	ZSL	GZSL
A. SAB-Encoder Architecture				
Baseline: ST-GCN [30]	76.9	66.4	67.3	55.6
Baseline: Shift-GCN [6]	78.3	67.1	68.0	58.1
w/o Local Branch ($\boldsymbol{\mathcal{F}}_{\text{local}}$)	79.5	67.6	69.4	58.5
w/o Strided Branch ($\boldsymbol{\mathcal{F}}_{\text{stride}}$)	82.8	70.2	72.6	60.8
w/o High-Pass Filter ($(\mathbf{I}-\mathbf{A})$)	79.2	69.2	70.3	59.6
B. Cross-Modal Alignment Strategy				
Rigid Semantic Anchoring (Static)	80.2	68.4	70.3	59.2
w/o Content-Aware Kinetic Gating	82.6	69.7	70.7	59.4
Replace KL div. with ℓ_2 distance	81.4	69.5	71.6	60.2
C. Objective Function				
w/o Sparsity Regularization ($\mathcal{L}_{\text{sparse}}$)	82.4	69.8	71.3	59.3
Saber (Full Model)	**87.5**	**76.1**	**76.8**	**65.9**

conventional methods. This confirms that our dynamic focusing provides highly effective guidance for cross-modal alignment in data-scarce scenarios.

4.5 Ablation Study

Component Efficacy Analysis. Table 4 details our ablation results. In the SAB-Encoder, removing the local branch causes a severe performance drop, proving the critical need to capture short-term dynamics. Omitting the strided branch or high-pass filter similarly harms accuracy, validating their roles in global stability and high-frequency detail preservation. Furthermore, our probabilistic alignment clearly outperforms rigid static anchoring. Removing content-aware gating, replacing KL divergence with ℓ_2 distance, or dropping sparsity regularization consistently degrades results. This confirms their collective effectiveness in noise suppression, uncertainty resolution, and kinetic focusing.

Interaction Modeling Capability. Table 5 details performance across various interactions. The baseline struggles with complex two-person interactions

Table 5: Performance comparisons across different interaction types on NTU subsets.

Components	Interaction Type (NTU Subset)	NTU RGB+D		NTU RGB+D 120	
		ZSL	GZSL	ZSL	GZSL
Saber w/o MI	Single Person	86.6	73.5	76.0	63.7
	Two-Person Independent	84.7	70.2	74.3	60.8
	Two-Person Interaction	69.7	54.7	54.6	42.2
Saber (Full)	Single Person	89.2	81.4	79.2	70.5
	Two-Person Independent	87.8	75.9	77.4	65.4
	Two-Person Interaction	77.9	63.5	68.4	52.1

Table 6: Performance and efficiency on extreme split protocols.

Methods	Params (M)	FLOPs (G)	NTU RGB+D 40/20 Splits ($\lvert\mathcal{Y}_u\rvert=20$)		NTU RGB+D 30/30 Splits ($\lvert\mathcal{Y}_u\rvert=30$)		NTU RGB+D 120 80/40 Splits ($\lvert\mathcal{Y}_u\rvert=40$)		NTU RGB+D 120 60/60 Splits ($\lvert\mathcal{Y}_u\rvert=60$)	
			ZSL	FPS	ZSL	FPS	ZSL	FPS	ZSL	FPS
A. Embedding-Based Paradigm										
ReViSE [22]	3.45	0.85	24.3	2153	14.8	2148	19.5	2151	8.3	2145
JPoSE [24]	4.62	0.96	20.1	1421	12.4	1417	13.7	1423	7.7	1412
PURLS [33]	11.45	2.38	31.1	497	23.5	494	28.4	498	19.6	491
SCoPLe [34]	9.28	1.87	32.0	923	18.2	918	25.3	921	15.7	915
B. Generative-Based Paradigm										
CADA-VAE [20]	2.41	1.12	16.2	1684	11.5	1679	10.6	1681	5.7	1675
SynSE [9]	6.78	1.28	19.9	1157	12.0	1152	13.6	1149	7.7	1144
C. Transport-Based Paradigm										
TDSM [7]	74.25	5.95	**36.1**	341	25.9	228	**37.0**	173	27.2	115
Flora [5]	53.70	4.82	31.1	345	**35.7**	297	**40.1**	236	**29.0**	173
Saber (Ours)	10.36	2.14	**39.5**	752	**37.1**	744	36.8	741	**29.6**	735

due to severe interference, but **Saber** significantly bridges this gap. By leveraging the Multi-Head Instance-Adaptive (MI) paradigm to treat peer movements as dynamic references, it filters out uninvolved participants and isolates genuine dynamics, achieving substantial recoveries in challenging multi-person scenarios.

Efficiency and Complexity Analysis. Table 6 evaluates computational efficiency. While transport-based paradigms achieve competitive accuracy, they suffer from heavy computational overhead and slow inference. Conversely, embedding and generative models are lightweight but experience severe performance degradation under extreme splits. Our approach strikes an optimal balance, delivering robust performance alongside a compact real-time inference.

4.6 Qualitative Analysis

Visualization of Probabilistic Anchoring. Figure 5 illustrates the part-level Gaussian distributions generated by our model. Kinetically active body parts explicitly form dense, low-variance peaks, reflecting high semantic certainty and strong alignment with the text anchor. Conversely, inactive or background regions naturally manifest as diffuse, high-variance curves. This adaptive variance mechanism effectively down-weights ambiguous noise, ensuring the model strictly anchors textual concepts to the most informative physical movements.

Analysis of Energy-Guided Dynamic Focusing. Figure 6 visualizes the instance-adaptive topology matrix ($\mathbf{A}$) alongside part-level kinetic energy (w).

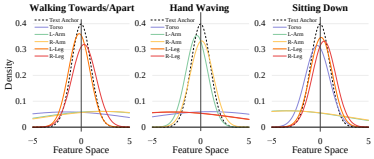

Fig. 5: Visualization of the learned variance-adaptive part-level Gaussian distributions.

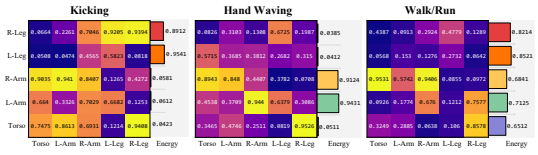

Fig. 6: Visualization of the Instance-Adaptive Topology Matrix (*heatmaps*) and part-level Kinetic Energy (*bars*) across different actions.

For localized actions, w sharply peaks at the semantically critical limbs, such as the legs in "Kicking" and the arms in "Hand Waving". Concurrently, the topology matrix $\mathbf{A}$ distinctly captures low-frequency structural redundancies. By filtering out these highly correlated static regions via $(\mathbf{I} - \mathbf{A})$, this visual evidence confirms that our gating mechanism effectively isolates explicit motion cues while dynamically suppressing irrelevant background parts.

5 Conclusion

In this work, we introduce **Saber**, a unified kinematics-driven framework designed to robustly anchor explicit semantic concepts to scale-aware kinetic salience for ZSAR. By actively reshaping spatio-temporal topologies, the proposed SAB-Encoder effectively preserves high-frequency motion details against spectral smoothing. Furthermore, by mapping part-level features into kinetic-driven variance-adaptive distributions, our uncertainty-aware probabilistic alignment strategy gracefully resolves structural uncertainties and avoids "attention sinks". Extensive experiments across the NTU RGB+D, NTU RGB+D 120, and PKU-MMD benchmarks demonstrate that **Saber** consistently achieves state-of-the-art performance, showcasing exceptional robustness and generalization.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. U23B2004 and 62402064), the Science and Technology Program of Hunan Province (Nos. 2024JK2005 and 2026AQ2032), and the Natural Science Foundation of Hunan Province, China (Nos. 2023JJ30054 and 2026JJ81155).

References

1. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2021)
2. Chao, W.L., Changpinyo, S., Gong, B., Sha, F.: An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild. In: ECCV. pp. 52–68 (2016)

3. Chen, Y., Guo, J., Guo, S., Tao, D.: Neuron: Learning Context-Aware Evolving Representations for Zero-Shot Skeleton Action Recognition. In: CVPR. pp. 8721–8730 (2025)
4. Chen, Y., Guo, J., He, T., Lu, X., Wang, L.: Fine-Grained Side Information Guided Dual-Prompts for Zero-Shot Skeleton Action Recognition. In: ACM Multimedia. pp. 778–786 (2024)
5. Chen, Y., Li, M., Rao, Z., Zeng, D., Guo, S., Guo, J.: Learning by Neighbor-Aware Semantics, Deciding by Open-form Flows: Towards Robust Zero-Shot Skeleton Action Recognition. In: CVPR Finding (2026)
6. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-Based Action Recognition With Shift Graph Convolutional Network. In: CVPR. pp. 180–189 (2020)
7. Do, J., Kim, M.: Bridging the Skeleton-Text Modality Gap: Diffusion-Powered Modality Alignment for Zero-shot Skeleton-based Action Recognition. In: ICCV (2025)
8. Fu, Z., Zhang, L., Xia, H., Mao, Z.: Linguistic-Aware Patch Slimming Framework for Fine-Grained Cross-Modal Alignment. In: CVPR. pp. 26297–26306 (2024)
9. Gupta, P., Sharma, D., Sarvadevabhatla, R.K.: Syntactically Guided Generative Embeddings for Zero-Shot Skeleton Action Recognition. In: ICIP. pp. 439–443 (2021)
10. Hao, X., Zhang, W., Wu, D., Zhu, F., Li, B.: Dual Alignment Unsupervised Domain Adaptation for Video-Text Retrieval. In: CVPR. pp. 18962–18972 (2023)
11. Kim, D., Tsai, Y.H., Zhuang, B., Yu, X., Sclaroff, S., Saenko, K., Chandraker, M.: Learning Cross-Modal Contrastive Features for Video Domain Adaptation. In: ICCV. pp. 13598–13607 (2021)
12. Kuang, J., Wang, H., Han, C., Zhang, Y., Gui, J.: Zero-shot skeleton-based action recognition with dual visual-text alignment. Pattern Recognition **171**, 112342 (2026)
13. Lai, H., Xiong, G., Mai, H., Liu, X., Zhang, T.: Rethinking Noisy Video-Text Retrieval via Relation-aware Alignment. In: CVPR. pp. 9231–9241 (2025)
14. Li, M.Z., Jia, Z., Zhang, Z., Ma, Z., Wang, L.: Multi-semantic Fusion Model For Generalized Zero-Shot Skeleton-Based Action Recognition. In: ICIG. pp. 68–80 (2023)
15. Li, S.W., Wei, Z.X., Chen, W.J., Yu, Y.H., Yang, C.Y., Hsu, J.Y.j.: SA-DVAE: Improving Zero-Shot Skeleton-Based Action Recognition by Disentangled Variational Autoencoders. In: ECCV. pp. 447–462 (2024)
16. Liu, C., Hu, Y., Li, Y., Song, S., Liu, J.: PKU-MMD: A Large Scale Benchmark for Skeleton-Based Human Action Understanding. In: VSCC@MM. pp. 1–8 (2017)
17. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. IEEE Trans. Pattern Anal. Mach. Intell. **42**(10), 2684–2701 (2020)
18. Ma, Z., Zhang, Z., Chen, Y., Qi, Z., Yuan, C., Li, B., Luo, Y., Li, X., Qi, X., Shan, Y., Hu, W.: EA-VTR: Event-Aware Video-Text Retrieval. In: ECCV. pp. 76–94 (2024)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML. pp. 8748–8763 (2021)
20. Schönfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z.: Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders. In: CVPR. pp. 8247–8255 (2019)

21. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In: CVPR. pp. 1010–1019 (2016)
22. Tsai, Y.H.H., Huang, L.K., Salakhutdinov, R.: Learning Robust Visual-Semantic Embeddings. In: ICCV. pp. 3591–3600 (2017)
23. Wang, Z., Sung, Y.L., Cheng, F., Bertasius, G., Bansal, M.: Unified Coarse-to-Fine Alignment for Video-Text Retrieval. In: ICCV. pp. 2804–2815 (2023)
24. Wray, M., Csurka, G., Larlus, D., Damen, D.: Fine-Grained Action Retrieval Through Multiple Parts-of-Speech Embeddings. In: ICCV. pp. 450–459 (2019)
25. Wu, W., Guo, Z., Chen, C., Xue, H., Lu, A.: Frequency-Semantic Enhanced Variational Autoencoder for Zero-Shot Skeleton-based Action Recognition. In: ICCV. pp. 11122–11129 (2025)
26. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient Streaming Language Models with Attention Sinks. In: ICLR (2024)
27. Xiong, B., Yang, X., Song, Y., Wang, Y., sheng Xu, C.: Modality-Collaborative Test-Time Adaptation for Action Recognition. In: CVPR. pp. 26722–26731 (2024)
28. Xu, H., Gao, Y., Li, J., Gao, X.: An Information Compensation Framework for Zero-Shot Skeleton-Based Action Recognition. IEEE Trans. Multim. **27**, 4882–4894 (2025)
29. Yan, S., Liu, J., Dong, N., Zhang, L., Tang, J.: Cross-modal Collaborative Representation Learning for Text-to-Image Person Retrieval. In: IJCAI. pp. 2152–2160 (2025)
30. Yan, S., Xiong, Y., Lin, D.: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In: AAAI. pp. 7444–7452 (2018)
31. Yang, L., Huang, Y., Sugano, Y., Sato, Y.: Interact before Align: Leveraging Cross-Modal Knowledge for Domain Adaptive Action Recognition. In: CVPR. pp. 14702–14712 (2022)
32. Zhou, Y., Qiang, W., Rao, A., Lin, N., Su, B., Wang, J.: Zero-shot Skeleton-based Action Recognition via Mutual Information Estimation and Maximization. In: ACM Multimedia. pp. 5302–5310 (2023)
33. Zhu, A., Ke, Q., Gong, M., Bailey, J.: Part-Aware Unified Representation of Language and Skeleton for Zero-Shot Action Recognition. In: CVPR. pp. 18761–18770 (2024)
34. Zhu, A., Zhu, J., Bailey, J., Gong, M., Ke, Q.: Semantic-guided Cross-Modal Prompt Learning for Skeleton-based Zero-shot Action Recognition. In: CVPR. pp. 13876–13885 (2025)