Beyond Alignment: A Generative Matching Paradigm via Flow Matching for Zero-Shot Skeleton-based Action Recognition

Xuan Liu[1[0009−0008−1749−2431]], Cong Wu[2[0000−0001−9555−9445]], Wei Fang[2[0000−0003−0750−4749]], and Zhenhua Feng*[1[0000−0002−4485−4249]]

[1] Multi-Modal Machine Intelligence Lab (M^{3}I), School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, Jiangsu, P.R. China
6243115014@stu.jiangnan.edu.cn, fengzhenhua@jiangnan.edu.cn
[2] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, Jiangsu, P.R. China
congwu@jiangnan.edu.cn, fangwei@jiangnan.edu.cn

Abstract. Zero-shot skeleton-based action recognition aims to identify unseen actions via semantic transfer. The existing methods often struggle with the topological mismatch between skeleton and text manifolds, leading to loose decision boundaries due to weak constraints and failing to adapt to unseen distributions during static inference. To mitigate this challenge, we propose ProtoFM, a novel generative framework based on conditional flow matching and prototype guidance. First, instead of heterogeneous skeleton-text matching, ProtoFM introduces a generative matching paradigm that performs homogeneous skeleton and pseudo-skeleton matching by synthesizing features via ODE-based flow integration. Second, we propose a prototype-guided center loss that anchors flow-trajectory states to learnable class prototypes, encouraging compact semantic clusters and improved class separation. Last, at the inference phase, we develop a training-free prototype-guided rectification strategy that exploits a global memory bank for progressive test-time refinement. Extensive experimental results obtained on multiple benchmarking datasets demonstrate the superiority of the proposed ProtoFM method over the state-of-the-art approaches.

Keywords: Skeleton-based Action Recognition · Zero-shot Learning · Generative Homogeneous Matching · Flow Matching

1 Introduction

Human action recognition underpins a wide spectrum of practical applications, including human behavior analysis [29,41], intelligent surveillance [15,17,28], etc. Among various modalities, skeleton sequences are particularly attractive due to their computational efficiency and relative robustness to appearance and environmental variations. While conventional methods [7, 26, 36] achieve promising

* Corresponding author: Zhenhua Feng (fengzhenhua@jiangnan.edu.cn).

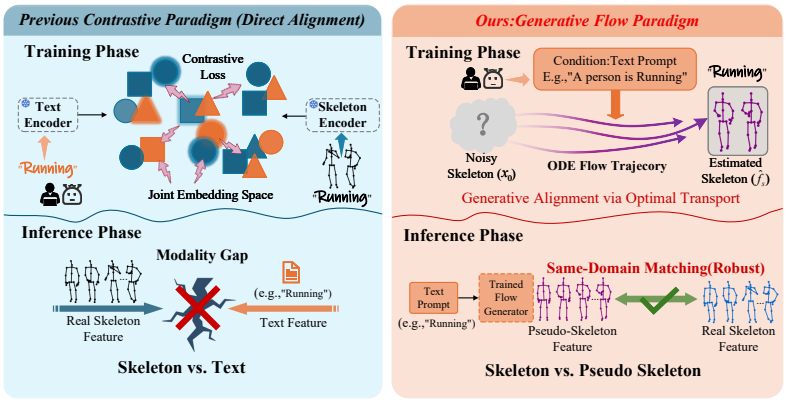


Fig. 1: A conceptual comparison between ProtoFM and previous methods.

results when large labeled data is available, their performance degrades significantly on unseen action categories that are absent from the training set. To mitigate this limitation, Zero-Shot Skeleton Action Recognition (ZSSAR) has been introduced, aiming to recognize unseen actions by transferring knowledge via a shared semantic space, typically bridging the heterogeneous gap between skeleton sequences and linguistic descriptions.

Early ZSSAR methods [12, 21, 43] predominantly rely on global alignment, mapping holistic skeleton features to class-level semantic representations. However, such coarse-grained matching may overlook fine-grained motion cues that are often critical to distinguish visually similar actions. Motivated by this limitation, recent studies have shifted to fine-grained modeling, ranging from the decomposition of skeleton sequences into local body parts [4, 44] to the integration of LLMs for semantic refinement [3, 45]. Notably, generative models, such as Variational Autoencoders (VAEs) and Diffusion Models [9, 22, 35], have gained increasing attention from the community. These approaches typically employ cross-modal generative frameworks in which the skeleton and semantic modalities are coupled through a shared latent prior (often Gaussian), thereby encouraging implicit alignment and reducing the modality gap.

Despite recent advances, existing ZSSAR paradigms, whether contrastive or generative, face two persistent challenges: limited feature discriminability and unstable latent geometry. Contrastive approaches typically optimize global- or local-level alignment objectives that pull skeleton representations toward semantic embeddings. However, such soft alignment constraints may not explicitly enforce intra-class compactness, leading to overlapping class distributions where fine-grained actions remain entangled. Generative methods aim to synthesize features for unseen classes, but they commonly rely on stochastic sampling. Without explicit geometric regularization, sampling-induced variance can yield dispersed feature embeddings that drift away from true class centers, further weakening decision boundaries. Beyond feature representations, both paradigms suffer from domain shift: models trained on seen categories generalize poorly to

unseen actions, motivating recent test-time refinement strategies that calibrate semantic anchors or caches with target-domain statistics.

As depicted in Fig. 1, to bridge the modality gap while mitigating the instability induced by stochastic generation, we propose a novel framework, namely ProtoFM, which reframes ZSSAR from heterogeneous cross-modal alignment to homogeneous generative matching. While conventional generative models (e.g., VAEs) often fail to establish stable mappings between static semantics and dynamic skeleton motions due to their stochastic nature, ProtoFM adopts Conditional Flow Matching (CFM) as the backbone and models cross-modal generation via Ordinary Differential Equation (ODE) based flow integration. This formulation constructs a stable transport path from a simple noise distribution to skeleton features, enabling deterministic feature synthesis conditioned on semantics and direct, efficient matching in a shared latent space. Nevertheless, feature generation alone does not guarantee discriminability or compactness. We observe that, without explicit geometric constraints, synthesized features for unseen classes tend to be loosely distributed, leading to ambiguous decision boundaries. To address this, we propose a Prototype-Guided Center (PGC) loss that serves as a geometric regularizer by explicitly anchoring the states of flow trajectories to dynamic and learnable class prototypes. In this way, the generated features are encouraged to concentrate into tight and separable clusters rather than forming dispersed clusters. Last, domain shift remains a practical obstacle: prototypes inferred purely from seen-class semantics may not align well with the true feature distribution of unseen samples. To reduce this discrepancy, we introduce a test-time prototype rectification strategy that leverages high-confidence test-time statistics to calibrate prototypes during inference. This rectification aligns prototypes with the actual unseen-domain distribution without requiring additional training.

In summary, the main contributions of ProtoFM include:

- A generative paradigm based on conditional flow matching, reformulating heterogeneous alignment as homogeneous skeleton & pseudo-skeleton matching via deterministic ODE trajectories for robust semantic transfer.
- A prototype-guided center loss that improves latent-space compactness and separability by steering flow-trajectory states toward learnable class prototypes, thereby reducing intra-class variance and enhancing discriminability.
- A training-free memory-bank guided rectification strategy that accumulates high-confidence test predictions in an online memory bank and uses the resulting empirical evidence to calibrate prototypes during inference, mitigating domain shift without additional training.
- Extensive experiments obtained on NTU RGB+D 60, NTU RGB+D 120, and PKU-MMD demonstrate that ProtoFM achieves state-of-the-art performance under both zero-shot and generalized zero-shot settings.

2 Related Work

2.1 Zero-Shot Skeleton-Based Action Recognition

The existing ZSSAR approaches can be roughly categorized into *embedding-alignment-based* methods and *generative-alignment-based* methods, depending on how skeleton features are associated with semantic descriptions.

Embedding-Alignment-Based Methods usually learn a shared embedding or a compatibility function to directly align skeleton representations with language semantics. Typical techniques include contrastive learning, mutual information maximization, and prompt learning, which encourage skeleton features to be close to the corresponding class semantics. Early works mainly focus on global alignment. For instance, SMIE [43] aligns class-level skeleton and semantic distributions via mutual information maximization. Recent methods pursue richer semantics and finer-grained alignment by exploiting LLM-generated descriptions, part-aware modeling, dual prompts, evolving prototypes, and multi-level semantic alignment [3,4,20,37,44]. These approaches enrich semantic representations and strengthen cross-modal correspondence between skeleton and textual features, leading to improved knowledge transfer from seen to unseen classes. From the adaptation perspective, PGFA [42] rectifies unseen text features with high-confidence test-time prototypes to alleviate alignment bias, while Dyna-PURLS [46] dynamically refines textual features using a confidence-aware memory bank. Skeleton-Cache [47] further explores training-free test-time adaptation through a non-parametric cache of structured skeleton descriptors. In addition, PP-CDL [48] and SCoPLe [45] adapt vision-language models via LoRA [31] or semantic-guided prompts, respectively, further enhancing semantic transfer and generalization under limited supervision.

Generative-Alignment-Based Methods aim to bridge the modality gap by synthesizing features or learning robust latent distributions with generative models such as VAE [18] and diffusion models. SynSE [12] learns separate VAEs for verbs and nouns to enable compositional generalization. GZSSAR [21] incorporates a VAE-based module to fuse multiple semantic sources, including action and motion descriptions. SA-DVAE [22] improves robustness by disentangling skeleton features into semantic-related & semantic-irrelevant components and aligning only the informative parts. FS-VAE [35] further integrates frequency-domain analysis via discrete cosine transform into the VAE framework. Flora [5] adopts a dual-VAE architecture and replaces the standard KL divergence with a geometric consistency objective that directly aligns the latent statistics of skeletons and semantics. TDSM [9] introduces diffusion models for ZSSAR, where the reverse diffusion process implicitly aligns skeleton features with text prompts in a unified latent space.

2.2 Flow Matching

Flow Matching (FM) [23, 24] provides a stable and deterministic alternative to diffusion [13] by learning vector fields for ODE-based density transport. Recent

methods have scaled FM to multi-modal tasks like speech [6, 39], images [11], videos [30], motion [14], and robotics [40], while others linearize trajectories by incorporating geometric priors such as optimal transport [2, 19] and topology preservation [10]. These advances underscore FM's potential as a transparent geometric interface, enabling explicit regularization of latent structures for effective cross-modal alignment, which is particularly beneficial for aligning heterogeneous skeleton-semantic features.

Summary: To the best of our knowledge, ProtoFM is the first to leverage the generative capability of conditional flow matching for ZSSAR. Rather than direct *"Text-to-Skeleton"* alignment, which suffers from a large modality gap, we reformulate the task as *"Real-to-Pseudo Skeleton"* matching. Through an ODE-based generative trajectory, text prompts are converted into pseudo-skeleton features, transforming heterogeneous cross-modal matching into homogeneous same-domain matching and mitigating domain disparity during inference.

3 The Proposed ProtoFM Method

3.1 Problem Definition

ZSSAR aims to transfer knowledge from seen to unseen action categories. Let $\mathcal{D} = \{(s_i, y_i, \tau_{y_i})\}_{i=1}^{N}$ denote a dataset, $s_i \in \mathbb{R}^{J\times 3\times F\times P}$ is the skeleton sequence with J joints, 3 coordinates, F frames, and P subjects, $y_i \in \mathcal{C}$ is the action label, and τ_{y_i} is the corresponding text description. The label set $\mathcal{C}$ is partitioned into disjoint seen and unseen sets, denoted by $\mathcal{C}_s$ and $\mathcal{C}_u$, respectively. Accordingly, $\mathcal{D}$ is split into a training set $\mathcal{D}_{\text{tr}}$ with $y_i \in \mathcal{C}_s$, and testing sets $\mathcal{D}_{\text{te}}^{s}$ and $\mathcal{D}_{\text{te}}^{u}$. Standard Zero-Shot Learning (ZSL) and Generalized ZSL (GZSL) evaluate a method on $\mathcal{D}_{\text{te}}^{u}$ and $\mathcal{D}_{\text{te}}^{s} \cup \mathcal{D}_{\text{te}}^{u}$, respectively.

For ZSSAR, we extract skeleton and semantic representations and relate them in a shared latent space. Specifically, the model consists of a skeleton encoder $\mathcal{F}s$ pretrained on seen classes and a pretrained text encoder $\mathcal{F}\tau$. For an input instance s_i, the features are processed as follows:

$$f_s^i = \mathcal{F}_s(s_i), \quad f_\tau^i = \mathcal{F}_\tau(\tau_{y_i}), \tag{1}$$

where f_s^i and f_τ^i are skeleton and semantic features. The objective of ZSSAR is to learn a transferable association between these two modalities such that, given a test skeleton from an unseen class, the model can correctly predict its label using only the corresponding semantic descriptions.

3.2 Generative Flow Matching Process

The objective of ProtoFM is to synthesize discriminative skeleton features for action categories by learning a deterministic transport from noise prior to skeleton distribution. This process is conditioned on enriched semantic representations that capture both global context and temporal evolution of actions.

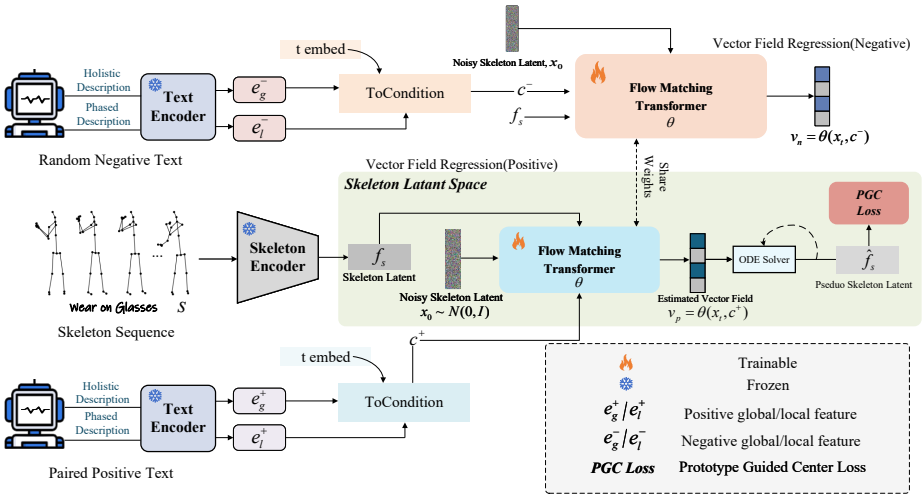


Fig. 2: An overview of the proposed ProtoFM framework.

To fully exploit the guiding capability of text, we employ GPT-4 [1] to generate action descriptions at two granularities: a *holistic description* providing a global summary, and a *phase-wise description* decomposing the action into initiation, execution, and completion stages, the specific details are provided in Appendix C. Given these descriptions, we utilize a frozen CLIP text encoder [31] to extract e_g and e_l, where $e_g \in \mathbb{R}^{1\times D_{txt}}$ denotes global text features, and $e_l \in \mathbb{R}^{M_l\times D_{txt}}$ denotes local text features, with M_l text tokens, D_{txt} is the size of text features. The final semantic condition c is formed by concatenating the global and local features: $c = [e_g|e_l]$. During training, we utilize both the ground-truth positive condition $c^+ = [e_g^+|e_l^+] + t$ and a randomly sampled negative condition $c^- = [e_g^-|e_l^-] + t$ to enhance the model's discriminative capability, where t is the timestep embedding.

As shown in Fig. 2, we model the generation as a Conditional Flow Matching (CFM) process. Let s denote a skeleton sequence, which is first mapped by a skeleton encoder to a latent feature $f_s = \mathcal{F}_s(s)$, where $f_s \in \mathbb{R}^{1\times D_{sk}}$ represents a sample from the target skeleton distribution P_s, D_{sk} is the size of skeleton features. Simultaneously, we sample noise x_0 from a standard Gaussian prior $P_n = \mathcal{N}(0, I)$. Following the theory of Optimal Transport, we construct a linear interpolation from the noise sample x_0 to the skeleton latent f_s. This interpolation induces a linear probability path p_t, where the intermediate state $x_t \sim p_t$ is defined as:

$$x_t = [1 - (1 - \sigma_m)t]x_0 + tf_s, \quad \sigma_m = 10^{-4}, t \in [0, 1]. \tag{2}$$

The target of our framework is to learn a flow matching Transformer, denoted as $\theta(x_t, t, c)$, which predicts a time-dependent velocity field that transport x_0 toward f_s. The primary optimization objective is to minimize the discrepancy between the predicted velocity and the constant target drift $v^* = \frac{\mathrm{d}x_t}{\mathrm{d}t} = f_s -$

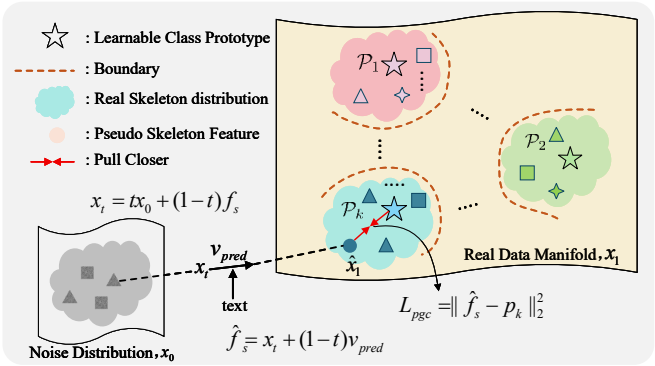


Fig. 3: An illustration of the proposed Prototype-Guided Center (PGC) loss.

$(1-\sigma_m)x_0$:

$$\mathcal{L}_{fm} = \mathbb{E}_{t,x_0,f_s}\left[\|v_\theta(x_t,t,c^+) - v^*\|_2^2\right], \tag{3}$$

where $v_\theta(x_t,t,c^+)$ denotes the predicted velocity. By minimizing $\mathcal{L}_{fm}$, the model learns to construct a straight and efficient trajectory in the latent space, guided by the positive semantic condition c^+.

Semantic Alignment via Triplet Loss. Beyond trajectory correctness, the learned vector field should be sensitive to semantic differences. Inspired by TDSM [9], we incorporate a Triplet Flow Matching loss ($\mathcal{L}_{tf}$) to enhance semantic discriminability. We define a flow-based distance metric $Dis(v_\theta, c) = \mathbb{E}_{t,x_0,f_s}\|v_\theta(x_t,t,c) - (f_s - x_0)\|_2^2$, which quantifies the alignment discrepancy between the predicted velocity and the target drift under condition c. The triplet loss enforces that the alignment error for a matched text-skeleton pair (c^+) remains smaller than that of a mismatched pair (c^-) by a margin β:

$$\mathcal{L}_{tf} = \max\left(0, Dis(v_\theta,c^+) - Dis(v_\theta,c^-) + \beta\right). \tag{4}$$

This constraint ensures that the evolution of the vector field is accurately guided by the intended semantic context, effectively refining the cross-modal alignment.

3.3 Prototype-Guided Geometric Regularization

Standard conditional flow matching emphasizes matching the vector field along the path, but it does not explicitly encourage compact class-wise structure in the generated feature space. As a result, the generated skeleton features are loosely distributed, which harms discriminability. To address this issue, we introduce a Prototype-Guided Center (PGC) loss, as illustrated in Fig. 3. First, we maintain a set of learnable prototypes $\mathcal{P} = \{\mathcal{P}_k\}_{k=1}^K$, where each $\mathcal{P}_k$ represents a class center in the latent space and is randomly initialized. Second, since the intermediate state x_t is noisy and unsuitable for direct clustering, we leverage the

linearity of the ODE formulation to project the current state x_t onto the clean data manifold, yielding a one-step estimation of $\hat{f}_s$:

$$\hat{f}_s = x_t + (1-t) \cdot v_\theta(x_t, t, c^+), \tag{5}$$

where $\hat{f}_s$ represents the predicted skeleton features at $t = 1$ given the current state at time t. We then formulate the PGC loss to minimize the intra-class distance between the estimated features and its corresponding class prototype $\mathcal{P}_y$. Crucially, to prevent gradient instability during the early training phase, we adopt a **Soft-Start Strategy** with a warm-up threshold E_{warm}:

$$\mathcal{L}_{pgc} = \begin{cases} \frac{1}{2}\sum \|\text{sg}(\hat{f}_s) - \mathcal{P}_y\|_2^2 & \text{if epoch} < E_{warm} \\ \frac{1}{2}\sum \|\hat{f}_s - \mathcal{P}_y\|_2^2 & \text{otherwise} \end{cases}, \tag{6}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. This strategy ensures that initially, only prototypes are updated to adapt to the feature distribution. Then, the flow matching Transformer is actively guided to generate compact clusters around these prototypes.

Overall Training Objective. The overall training loss is defined as:

$$\mathcal{L}_{total} = \lambda_{fm}\mathcal{L}_{fm} + \lambda_{pgc}\mathcal{L}_{pgc} + \lambda_{tf}\mathcal{L}_{tf}, \tag{7}$$

where λ_{fm}, λ_{pgc}, and λ_{tf} are balancing parameters of the flow matching loss, the geometric regularization, and the semantic alignment term.

3.4 Generative Inference with Online Rectification

To mitigate the domain shift issue, we propose a unified online inference paradigm for ZSL and GZSL, which sequentially adapts to the target domain. This is realized via the Prototype Memory bank Guided Rectification (PMGR) strategy. Specifically, we employ class-level prototypes as stable anchors for feature matching and domain-specific adjustment. As illustrated in Fig. 4, for each incoming feature $f = \mathcal{F}_s(s)$ ($s \in \mathcal{D}_{\text{te}}$, where $\mathcal{D}^u_{\text{te}}$ for ZSL and $\mathcal{D}^s_{\text{te}} \cup \mathcal{D}^u_{\text{te}}$ for GZSL), PMGR performs incremental refinement through three core steps: generative matching, memory update, and prototype guided rectification.

Generative Maximum Matching. Let $\mathcal{C}_{\text{tgt}}$ denote the candidate label space ($\mathcal{C}_u$ for ZSL and $\mathcal{C}_s \cup \mathcal{C}_u$ for GZSL). For each class $y \in \mathcal{C}_{\text{tgt}}$, conditioned on its textual description τ_y, the flow matching Transformer synthesizes a support set $\mathcal{G}_y = \{\hat{f}_m\}_{m=1}^M$ via ODE sampling with n_s integration steps. To evaluate the similarity between the features f of a test sample and the class label y, we define the class-wise generative score as the maximum cosine similarity between f and the generated samples in $\mathcal{G}_y$:

$$s_{\text{gen}}(f, y) = \max_{\hat{f} \in \mathcal{G}_y} \cos(f, \hat{f}). \tag{8}$$

By calculating this for all classes, we obtain the generative score vector $\mathbf{s}_{\text{gen}} = [s_{\text{gen}}(f, y)]_{y \in \mathcal{C}_{\text{tgt}}}$, which serves as the primary confidence criterion for subsequent online adaptation.

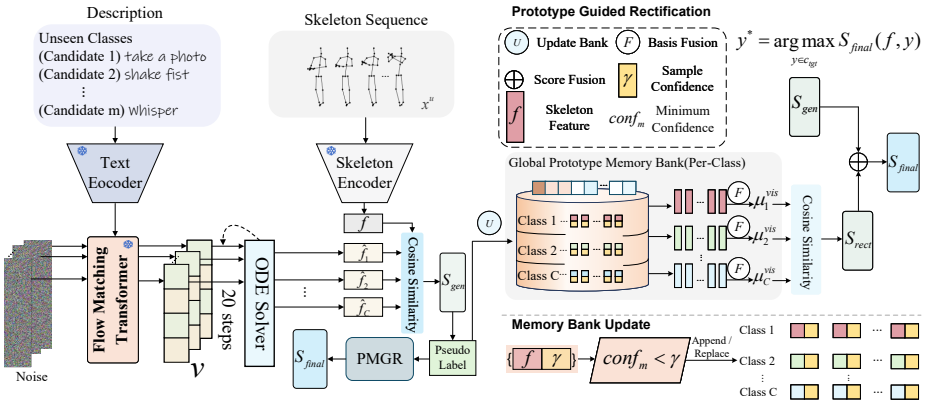


Fig. 4: The inference pipeline of ProtoFM for ZSSAR, where the right part illustrates the Prototype Memory bank Guided Rectification (PMGR) strategy.

Online Memory Bank Update. For the features f of each test sample, we assign an initial pseudo-label $\hat{y} = \arg\max_{y \in \mathcal{C}_{\text{tgt}}} s_{\text{gen}}(f, y)$ and record its corresponding confidence $\gamma = s_{\text{gen}}(f, \hat{y})$. To capture the underlying distribution of the target domain, we maintain a class-wise memory bank $\mathcal{B} = \{\mathcal{B}_y\}_{y \in \mathcal{C}_{\text{tgt}}}$ with a maximum capacity of M for each class. Specifically, each $\mathcal{B}_y$ stores M skeleton features and their corresponding confidence scores, denoted as (f, γ) for each, which are utilized for the subsequent update. We adopt a priority-based update strategy: a new pair (f, γ) replaces the element in $\mathcal{B}_{\hat{y}}$ with the lowest confidence only if γ is higher. This mechanism ensures that the memory bank consistently retains the most representative skeleton features as target-domain evidence.

Prototype Guided Rectification. Based on the updated memory bank, we compute a visual prototype μ_y^{vis} for each class by taking the normalized mean features in $\mathcal{B}_y$, $\mu_y^{\text{vis}} = Normalize(\frac{1}{|\mathcal{B}_y|} \sum_{(f', \gamma') \in \mathcal{B}_y} f')$. The alignment with these real samples is captured by the rectification score $s_{\text{rect}}(f, y) = \cos(f, \mu_y^{\text{vis}})$, then we get $\mathbf{s}_{\text{rect}} = [s_{\text{rect}}(f, y)]_{y \in \mathcal{C}_{\text{tgt}}}$. The final prediction y^* is determined via a weighted fusion of the generative and rectified score vectors:

$$\mathbf{s}_{\text{final}} = \alpha \mathbf{s}_{\text{gen}} + (1 - \alpha)\mathbf{s}_{\text{rect}}, \quad y^* = \arg\max_{y \in \mathcal{C}_{\text{tgt}}} s_{\text{final}}(f, y), \tag{9}$$

where α is a balancing coefficient that weights the prior generative knowledge against the observed target-domain evidence.

4 Experiments

This section presents a comprehensive evaluation of the proposed ProtoFM method on widely used benchmarks for skeleton-based action recognition. We conduct experiments on three public datasets, including NTU RGB+D 60 [33],

Table 1: A comparison with the state-of-the-art methods on NTU RGB+D 60 (XSub) and NTU RGB+D 120 (XSub) dataset under the ZSL and GZSL settings. The best and the second-best results are marked in Red and Blue, respectively.

Method	Venue	NTU RGB+D 60 (Xsub)								NTU RGB+D 120 (Xsub)							
		55/5 Split				48/12 Split				110/10 Split				96/24 Split			
		ZSL	GZSL			ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
		$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$
ReViSE [16]	ICCV-17	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 [34]	ICCV-19	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 [32]	CVPR-19	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 [12]	ICIP-21	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
GZSSAR [21]	ICIG-23	83.6	71.7	66.2	68.8	49.2	58.8	40.0	47.6	71.2	46.8	**68.3**	55.6	59.7	56.8	48.6	52.4
SMIE [43]	ACMMM-23	77.9	-	-	-	41.5	-	-	-	61.3	-	-	-	42.3	-	-	-
PURALS [44]	CVPR-24	79.2	-	-	-	41.0	-	-	-	72.0	-	-	-	52.0	-	-	-
SA-DVAE [22]	ECCV-24	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
STAR [4]	ACMMM-24	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
DVTA [20]	PR-25	79.3	-	-	-	44.1	-	-	-	**74.9**	-	-	-	51.8	-	-	-
InfoCPL [37]	TMM-25	85.9	-	-	-	53.3	-	-	-	74.8	-	-	-	60.1	-	-	-
ScoPLe [45]	CVPR-25	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-25	**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 [35]	ICCV-25	**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 [9]	ICCV-25	86.5	-	-	-	56.0	-	-	-	74.2	-	-	-	**65.1**	-	-	-
ProtoFM	-	**87.3**	**76.0**	**76.2**	**76.1**	**73.2**	**70.5**	54.6	**61.5**	**82.9**	**68.3**	60.6	**64.2**	**69.5**	54.6	**57.4**	**56.0**

NTU RGB+D 120 [27], and PKU-MMD I [25]. In addition, ablation studies are conducted to verify the impact of the proposed innovative components for zero-shot action recognition.

4.1 Experimental Settings

Following previous studies [12,22], we adopt the standard seen and unseen class partition protocols for skeleton-based zero-shot action recognition. Specifically, we consider standard splits [4,12], random splits [4,22,43], and more stringent evaluation settings [44]. The full specifications of these datasets and partitions are provided in Appendix B of the supplementary material. In the ZSL setting, we report Top-1 accuracy $\mathcal{A}cc = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i == \hat{y})$ on the unseen test set $\mathcal{D}_{\text{te}}^{u}$. For GZSL, we measure accuracy on seen categories ($\mathcal{S}$) using $\mathcal{D}_{\text{te}}^{s}$, unseen categories ($\mathcal{U}$) using $\mathcal{D}_{\text{te}}^{u}$, and report their harmonic mean $\mathcal{H} = (2 \times \mathcal{S} \times \mathcal{U})/(\mathcal{S} + \mathcal{U})$. Additional implementation details are provided in Appendix D.

4.2 Comparison with State-of-the-Art

Evaluation on the Basic Split Benchmark. We first evaluate ProtoFM on the Xsub benchmarks of NTU-60 and NTU-120 under the basic split setting. The results are reported in Table 1. Following prior protocols, we conduct experiments using skeleton features extracted by SynSE [12]-based 4s-Shift-GCN [8] and STAR [4]-based 1s-Shift-GCN. Across all configurations, ProtoFM demonstrates strong performance in both ZSL and GZSL scenarios, achieving the best results on the two commonly used splits of NTU-60 and NTU-120.

Table 2: Average performance on three random seen–unseen splits.

Method	Venue	NTU-60		NTU-120		PKU-MMD I	
		ZSL	GZSL	ZSL	GZSL	ZSL	GZSL
ReViSE [16]	ICCV-17	60.9	60.3	44.9	40.3	59.3	49.8
JPoSE [34]	ICCV-19	59.4	60.1	46.7	43.7	57.2	51.6
CADA-VAE [32]	CVPR-19	61.8	66.4	45.2	45.6	60.7	45.8
SynSE [12]	ICIP-21	64.2	67.5	47.3	43.5	60.8	49.5
SMIE [43]	ACMMM-23	65.1	–	46.4	–	60.8	–
SA-DVAE [22]	ECCV-24	84.2	75.3	50.7	47.5	66.5	54.7
SCoPLe [45]	CVPR-25	83.7	**77.7**	53.3	**54.1**	**71.4**	54.9
TDSM [9]	ICCV-25	**88.9**	–	**69.5**	–	70.8	–
FS-VAE [35]	ICCV-25	–	–	–	–	71.2	**59.0**
ProtoFM	-	**94.1**	**78.2**	**72.1**	**62.1**	**71.5**	**55.6**

Table 3: Influence of each innovative component in our ProtoFM method.

$\mathcal{L}_{tf}$	$\mathcal{L}_{pgc}$	PMGR	NTU-60 (48/12)		NTU-120 (110/10)	
			ZSL	GZSL	ZSL	GZSL
✗	✗	✗	55.1	54.0	69.2	60.1
✓	✗	✗	57.5	55.6	72.1	60.3
✓	✓	✗	69.2	58.1	76.2	63.2
✓	✓	✓	**73.2**	**61.5**	**82.9**	**66.6**

Specifically, for ZSL on the NTU-60 Xsub benchmark (under 55/5 and 48/12 splits), ProtoFM outperforms the current state-of-the-art method Neuron [3], by 0.4 and 10.5 in $\mathcal{A}cc$, respectively. Similarly, on the NTU-120 Xsub dataset, ProtoFM outperforms the second-best methods, DVTA [20] and TDSM [9], by 8.0 (110/10 split) and 4.4 (96/24 split) in $\mathcal{A}cc$, respectively. Meanwhile, ProtoFM achieves significantly higher accuracy than competing methods in the GZSL setting. Appendix E further demonstrates the classification stability and robustness of our method under extreme split settings (extended from PURLS [44]), and the results obtained on the Xview and Xset benchmarks.

Evaluation on the Random Split Benchmark. We further assess our approach under the random split benchmark following the evaluation protocol of [43]. For each dataset, we construct three random seen and unseen partitions and extract skeleton features using ST-GCN [38]. The results in Table 2 report the average performance over these splits. ProtoFM consistently outperforms all competing approaches across the three datasets, with especially clear gains in the ZSL setting.

Table 4: Ablation study on different types of descriptions under the ZSL setting. The results compare models trained with only holistic descriptions, only phased descriptions, and the combination of both.

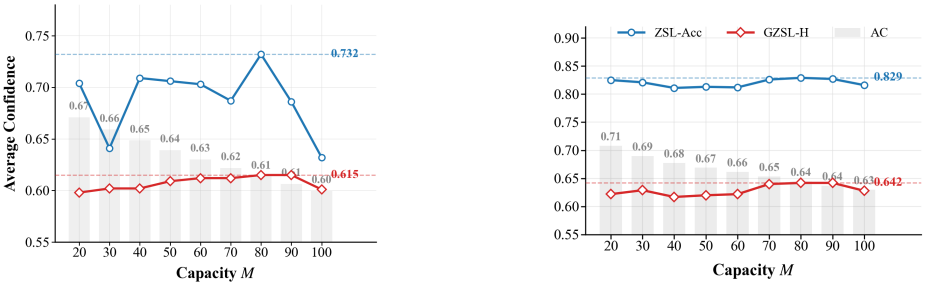
Holistic	Phased	NTU-60 (Acc, %)		NTU-120 (Acc, %)	
		55/5 split	48/12 split	110/10 split	96/24 split
✓	✗	82.7	59.3	79.8	59.1
✗	✓	80.9	65.8	80.8	56.7
✓	✓	**87.3**	**73.2**	**82.9**	**69.5**

4.3 Ablation Study

Influence of the Innovative Components. The ablation study in Table 3 evaluates the contribution of the triplet flow matching loss $\mathcal{L}_{tf}$, the prototype-guided center loss $\mathcal{L}_{pgc}$, and the Prototype Memory Bank Guided Rectification strategy (PMGR). Unless otherwise specified, all settings in Table 3 include the flow matching loss $\mathcal{L}_{fm}$ as the default objective. Specifically, employing $\mathcal{L}_{fm}$ alone ensures geometric path correctness but lacks explicit semantic constraints, leading to ambiguous decision boundaries. Conversely, adding $\mathcal{L}_{tf}$ enforces inter-class separability, while $\mathcal{L}_{pgc}$ enhances intra-class compactness by anchoring trajectories to learnable prototypes. Introducing the PMGR strategy during the inference stage further improves the accuracy. Ultimately, the full combination strikes an optimal balance between generative fidelity, semantic separability, and structural compactness, achieving state-of-the-art performance.

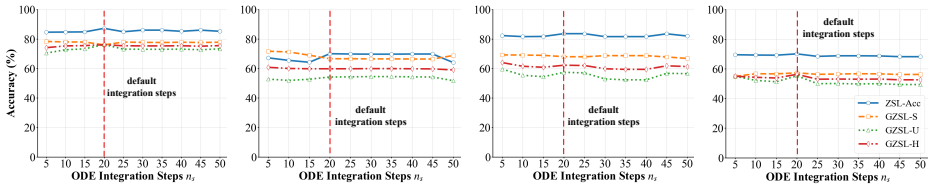
Contribution of the Holistic and Phased Descriptions: As shown in Table 4, using both description types consistently outperforms single-granularity baselines across all benchmarks. Holistic descriptions act as a global semantic anchor, capturing high-level context for robust matching. Complementarily, phased descriptions provide temporally structured supervision, enabling the distinction of actions with similar global contexts but different execution dynamics. The substantial accuracy boost (e.g., from 82.7 to 87.3 on the NTU-60 55/5 split) empirically confirms that global semantic consistency and fine-grained temporal discriminability are mutually complementary.

Impact of the Per-Class Memory Capacity. As illustrated in Fig. 5, we vary the memory capacity M of each class from 20 to 100 in increments of 10 and observe a rise-then-fall trend, with the best performance at M=80. When M is small (e.g., 20–40), each class memory bank contains too few reliable samples to cover intra-class variations, yielding unstable and noisy prototypes that are sensitive to early predictions. As M increases, the bank better captures class diversity, producing smoother prototypes and more effective rectification. However, a large M tends to accumulate imperfect pseudo-labeled instances over time, making prototypes over-smoothed and less discriminative, and amplifying confirmation bias during online updates. Overall, M=80 provides the best balance between representativeness and purity, and is used by default.



(a) NTU-60 (48/12) **(b)** NTU120 (110/10)

Fig. 5: Analysis of prototype memory bank capacity (M) of each class and the corresponding average confidence on NTU60 and NTU120.



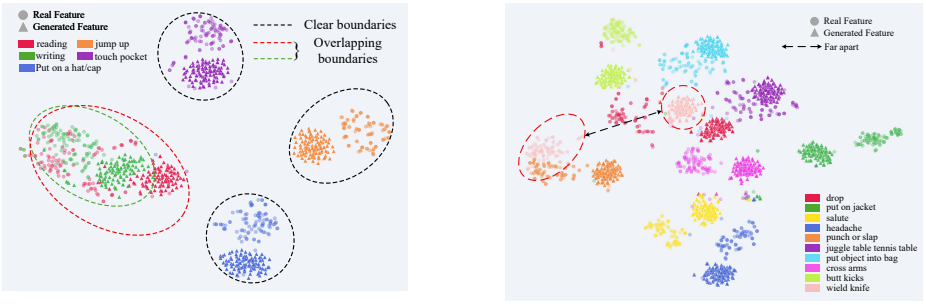
(a) NTU-60 (55/5) **(b)** NTU-60 (48/12) **(c)** NTU-120 (110/10) **(d)** NTU-120 (96/24)

Fig. 6: Performance comparison on NTU-60 and NTU-120 with different ODE integration steps n_s during the inference phase.

Influence of Inference ODE Steps. As illustrated in Fig. 6, the model's performance fluctuates across varying ODE integration steps (n_s). A consistent performance degradation is observed at both insufficient and excessive step counts. For smaller step counts, the trajectory estimation lacks adequate refinement and fails to fully capture the guidance from textual conditions. In contrast, as the step count increases beyond an optimal threshold, cumulative numerical errors and over-reliance on fragmented velocity directions lead to trajectory deviations. This leads to distribution shifts and a subsequent decline in classification accuracy.

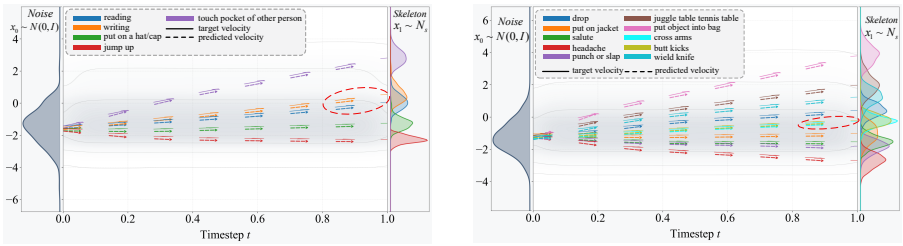
4.4 Qualitative Analysis

Visualization of Generation Quality. Fig. 7 visualizes the t-SNE distributions of real unseen-class skeleton features and the corresponding pseudo-features generated by the flow matching Transformer on NTU60 (55/5) and NTU120 (110/10). Overall, for most unseen categories, the generated samples form compact clusters that closely align with their real counterparts while remaining well-separated from other classes. This demonstrates the ability of flow matching Transformer in effectively bridging the domain gap between semantics and visual features. However, as highlighted in Fig. 7, we observe a challenging case induced by semantic similarity: classes like "*reading*" and "*writing*" appear



(a) NTU-60 (55/5) **(b)** NTU-120 (110/10)

Fig. 7: The t-SNE visualization of real (circles) and synthesized (triangles) skeleton features on NTU60 (55/5) and NTU120 (110/10).



(a) NTU-60 (55/5) **(b)** NTU-120 (110/10)

Fig. 8: Flow velocity visualization from shared noise to class-specific modes. Subfigures (a) and (b) illustrate the distributions of transport trajectories obtained on NTU-60 and NTU-120, respectively.

partially entangled in the feature space. Since these actions share highly similar motion patterns, they are embedded nearby in the pretrained space, making perfect disentanglement difficult even for the generative model. Nevertheless, the general trend remains consistent across different splits, suggesting that the learned trajectories generalize well to more complex unseen scenarios.

Visualization of Multi-Class Flow Velocity. Fig. 8 illustrates the trajectories of the same initial noise under different class-text conditions. Starting from an identical state at t=0, the paths progressively diverge into distinct, class-specific trajectories, indicating that flow matching Transformer learns a conditional vector field capable of steering a shared prior into unique modal regions. Notably, for semantically proximal actions such as "*reading*" and "*writing*", the predicted flow velocities exhibit some overlap (indicated by red dashed circles in Fig. 8a and Fig. 8b). From a manifold perspective, this suggests that the vector fields for such actions are closely aligned, posing a challenge for fine-grained separation during the transport process. To mitigate this, our Prototype-Guided Center Loss anchors intermediate states to class prototypes, effectively reducing trajectory drift and guiding the noise distribution toward the real skeleton man-

ifold while maintaining as much inter-class separability as possible as the value of t approaches 1.

5 Conclusion

In this paper, we presented ProtoFM, a novel framework that reframes the ZSSAR paradigm from heterogeneous cross-modal alignment to homogeneous generative matching. Leveraging conditional flow matching, ProtoFM modeled a deterministic transport process from noise to skeleton distributions conditioned on semantics, which effectively bridges the modality gap and enables efficient matching in a shared latent space. To improve the discriminability of the generated manifold, we introduced a prototype-guided center loss that explicitly encourages intra-class compactness by steering generated features toward learnable class prototypes along the flow trajectories. To further mitigate the domain shift in unseen scenarios, we proposed a training-free, memory bank-based rectification strategy that calibrated prototypes using high-confidence target-domain observations. Together, these components provide a robust generative baseline for ZSSAR and highlight the potential of flow-based modeling for cross-modal transfer under limited supervision.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Grant No. 61902153, Grant No. 62473176).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Cao, Y., Song, Z., Yang, C.: Video latent flow matching: Optimal polynomial projections for video interpolation and extrapolation. arXiv preprint arXiv:2502.00500 (2025)
3. Chen, Y., Guo, J., Guo, S., Tao, D.: Neuron: Learning context-aware evolving representations for zero-shot skeleton action recognition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8721–8730 (2025)
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: Proceedings of the 32nd ACM International Conference on 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3374–3383 (2026)
6. Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., JianZhao, J., Yu, K., Chen, X.: F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6255–6271 (2025)

7. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13359–13368 (2021)
8. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: 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)
9. Do, J., Kim, M.: 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)
10. Enciso, A., Peñafiel-Tomás, J., Peralta-Salas, D.: Steady 3d euler flows via a topology-preserving convex integration scheme. arXiv preprint arXiv:2501.13632 (2025)
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)
12. Gupta, P., Sharma, D., Sarvadevabhatla, R.K.: Syntactically guided generative embeddings for zero-shot skeleton action recognition. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 439–443. IEEE (2021)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
14. Hu, V.T., Yin, W., Ma, P., Chen, Y., Fernando, B., Asano, Y.M., Gavves, E., Mettes, P., Ommer, B., Snoek, C.G.: Motion flow matching for human motion synthesis and editing. arXiv preprint arXiv:2312.08895 (2023)
15. Huang, H., Lin, L., Zhao, L., Huang, H., Ding, S.: Tshnn: Temporal-spatial hybrid neural network for cognitive wireless human activity recognition. IEEE Transactions on Cognitive Communications and Networking **10**(6), 2088–2101 (2024)
16. Hubert Tsai, Y.H., Huang, L.K., Salakhutdinov, R.: Learning robust visual-semantic embeddings. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3571–3580 (2017)
17. Khan, Z.A., Sohn, W.: Abnormal human activity recognition system based on r-transform and kernel discriminant technique for elderly home care. IEEE Transactions on Consumer Electronics **57**(4), 1843–1850 (2012)
18. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
19. Kornilov, N., Mokrov, P., Gasnikov, A., Korotin, A.: Optimal flow matching: learning straight trajectories in just one step. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. NIPS '24, Curran Associates Inc., Red Hook, NY, USA (2024)
20. Kuang, J., Wang, H., Han, C., Zhang, Y., Gui, J.: Zero-shot skeleton-based action recognition with dual visual-text alignment. Pattern Recognition p. 112342 (2025)
21. Li, M.Z., Jia, Z., Zhang, Z., Ma, Z., Wang, L.: Multi-semantic fusion model for generalized zero-shot skeleton-based action recognition. In: International Conference on Image and Graphics. pp. 68–80. Springer (2023)

22. 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: European Conference on Computer Vision. pp. 447–462. Springer (2024)
23. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
24. Lipman, Y., Havasi, M., Holderrieth, P., Shaul, N., Le, M., Karrer, B., Chen, R.T., Lopez-Paz, D., Ben-Hamu, H., Gat, I.: Flow matching guide and code. arXiv preprint arXiv:2412.06264 (2024)
25. Liu, C., Hu, Y., Li, Y., Song, S., Liu, J.: Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. arXiv preprint arXiv:1703.07475 (2017)
26. Liu, H., Liu, Y., Ren, M., Wang, H., Wang, Y., Sun, Z.: Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29248–29257 (2025)
27. 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 Transactions on Pattern Analysis and Machine Intelligence **42**(10), 2684–2701 (2019)
28. Naimi, S., Bouachir, W., Bilodeau, G.A., Mishara, B.: Sstar: Skeleton-based spatio-temporal action recognition for intelligent video surveillance and suicide prevention in metro stations. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 813–823 (2025)
29. Pang, S.M., Cao, J.X., Jian, M.Y., Lai, J., Yan, Z.Y.: Br-gan: A pedestrian trajectory prediction model combined with behavior recognition. IEEE Transactions on Intelligent Transportation Systems **23**(12), 24609–24620 (2022)
30. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PmLR (2021)
32. Schonfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z.: Generalized zero-shot learning via aligned variational autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 54–57 (2019)
33. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: 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)
34. Wray, M., Larlus, D., Csurka, G., Damen, D.: 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)
35. Wu, W., Guo, Z., Chen, C., Xue, H., Lu, A.: Frequency-semantic enhanced variational autoencoder for zero-shot skeleton-based action recognition. arXiv preprint arXiv:2506.22179 (2025)
36. Wu, Z., Sun, P., Chen, X., Tang, K., Xu, T., Zou, L., Wang, X., Tan, M., Cheng, F., Weise, T.: Selfgcn: Graph convolution network with self-attention for skeleton-based action recognition. IEEE Transactions on Image Processing (2024)

37. Xu, H., Gao, Y., Li, J., Gao, X.: An information compensation framework for zero-shot skeleton-based action recognition. IEEE Transactions on Multimedia **27**, 4882–4894 (2025)
38. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
39. Yao, J., Yuguang, Y., Pan, Y., Ning, Z., Ye, J., Zhou, H., Xie, L.: Stablevc: Style controllable zero-shot voice conversion with conditional flow matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 25669–25677 (2025)
40. Zhang, Q., Liu, Z., Fan, H., Liu, G., Zeng, B., Liu, S.: Flowpolicy: Enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 14754–14762 (2025)
41. Zhou, E., Leather, O., Zhuang, A., Budhwani, A., Dempster, R., Li, Q., Al-Sharman, M., Rayside, D., Melek, W.: Ralacs: Action recognition in autonomous vehicles using interaction encoding and optical flow. IEEE Transactions on Cybernetics **55**(2), 512–525 (2024)
42. Zhou, K., Zhang, S., You, Z., Hu, J., Tan, M., Liu, F.: Zero-shot skeleton-based action recognition with prototype-guided feature alignment. arXiv preprint arXiv:2507.00566 (2025)
43. 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: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5302–5310 (2023)
44. Zhu, A., Ke, Q., Gong, M., Bailey, J.: 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)
45. Zhu, A., Zhu, J., Bailey, J., Gong, M., Ke, Q.: Semantic-guided cross-modal prompt learning for skeleton-based zero-shot action recognition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13876–13885 (2025)
46. Zhu, J., Zhu, A., Bailey, J., Liu, J., Rahmani, H., Bennamoun, M., Boussaid, F., Ke, Q.: Dynapurls: Dynamic refinement of part-aware representations for skeleton-based zero-shot action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
47. Zhu, J., Zhu, A., Rahmani, H., Liu, J., Bennamoun, M., Ke, Q.: Boosting skeleton-based zero-shot action recognition with training-free test-time adaptation. Advances in Neural Information Processing Systems **38**, 103521–103551 (2026)
48. Zhu, X., Shu, X., Huang, P., Tang, J.: Prompt-guided prototype-aware commonality and discrimination learning for zero-shot skeleton-based action recognition. IEEE Transactions on Multimedia **27**, 7053–7066 (2025)