BALANCING CROSS-MODAL CLASS ALIGNMENT AND LOCAL SEPARATION FOR ZERO-SHOT SKELETON ACTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot skeleton action recognition requires visual–semantic relations learned from seen classes to generalize to unseen actions. Existing cross-modal methods mainly reduce instance- or distribution-level modality discrepancy, but such alignment does not directly guarantee a suitable class structure. Visual and semantic representations of the same action may still form inconsistent class centers, while stronger positive alignment may compress the local separation between similar actions. We address this alignment–separation problem with a hierarchical class-geometry objective. Cross-Modal Prototype Neighborhood Modeling (CPNM) maintains visual and semantic prototype memories and uses modality-swapped anchors to form positive cross-modal class neighborhoods. Hard-Neighbor Discrepancy Amplification (HNDA) then identifies the nearest confusing class prototypes and imposes a relative margin around these neighborhoods. Experiments on NTU RGB+D 60, NTU RGB+D 120, and PKU-MMD demonstrate competitive performance under ZSL and GZSL protocols. Controlled ablations and geometry analyses further verify the complementary roles of cross-modal neighborhood alignment and local hard-neighbor separation.

1 INTRODUCTION

Skeleton-based action recognition identifies human actions from body-joint sequences Yan et al. (2018). Compared with RGB videos, skeleton data is less sensitive to background clutter, illumination, and appearance variation Liu et al. (2019). However, supervised methods generally assume a fixed training label space Shi et al. (2019). Zero-shot skeleton action recognition instead uses action names and textual descriptions to recognize unseen classes without their visual samples during training Xian et al. (2017).

Recent methods map skeleton features and class semantics into a shared latent space and transfer visual–semantic relations from seen to unseen classes Gupta et al. (2021); Zhou et al. (2023). Generative models further reduce the modality gap through latent distribution modeling and cross-modal reconstruction Schonfeld et al. (2019); Li et al. (2024). Nevertheless, reducing instance- or distribution-level discrepancy does not necessarily produce a suitable class structure. Skeleton actions often differ only in subtle joint movements or trajectories Chen et al. (2021b). Their visual and semantic representations may therefore form inconsistent class centers, while stronger positive alignment does not guarantee sufficient separation between neighboring actions. The problem is not simply to align the two modalities more closely, but to improve cross-modal class consistency without losing local discriminability.

To this end, we introduce Cross-Modal Prototype Neighborhood Modeling (CPNM) and Hard-Neighbor Discrepancy Amplification (HNDA). CPNM maintains visual and semantic prototype memories and uses modality-swapped class anchors to form positive cross-modal neighborhoods, reducing modality-specific center inconsistency. Positive neighborhood modeling alone, however, does not determine how far nearby classes should remain separated. HNDA therefore builds on the CPNM neighborhood energy, selects the nearest confusing negative-class prototypes, and imposes a relative margin against them. In this hierarchy, CPNM determines what should remain cross-modally compact, whereas HNDA constrains what must remain locally separated.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

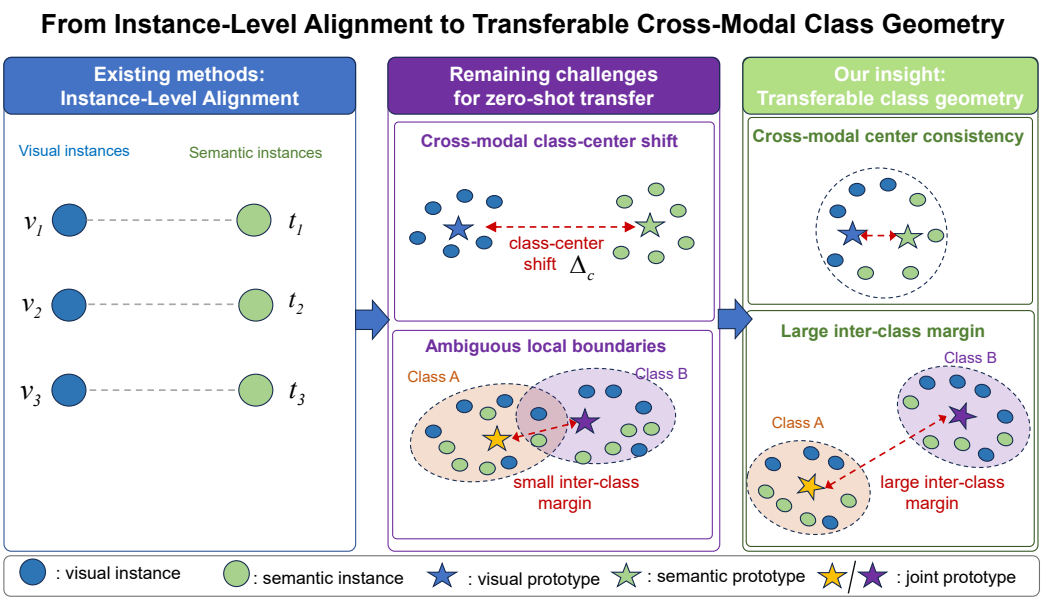


Figure 1: Motivation of class-level cross-modal geometry learning. Instance-level alignment may leave visual–semantic class-center inconsistency and insufficient local separation between similar actions. We jointly model cross-modal class consistency and hard-neighbor separation.

The contributions are summarized as follows:

- We formulate zero-shot skeleton representation learning from the joint perspective of cross-modal class alignment and local separation, showing that reducing modality discrepancy alone does not guarantee discriminative class neighborhoods.
- We develop a hierarchical class-geometry objective in which CPNM forms positive cross-modal neighborhoods through modality-swapped prototype anchoring, while HNDA preserves their local boundaries with a hard-neighbor relative margin.
- Controlled component ablations and class-geometry analyses show that CPNM reduces visual–semantic center discrepancy, while HNDA recovers local separation that is not guaranteed by positive neighborhood modeling alone. Experiments on three skeleton benchmarks further validate the resulting representation under ZSL and GZSL settings.

2 Related Work

2.1 Zero-Shot Skeleton Action Recognition

Zero-shot skeleton action recognition transfers visual–semantic relations learned from seen skeleton samples to unseen action classes represented by auxiliary semantics. Early approaches learn shared embeddings between visual observations and semantic descriptions Tsai et al. (2017); Wray et al. (2019), while later skeleton-specific methods exploit structured linguistic information or stronger semantic representations to improve cross-modal transfer Gupta et al. (2021); Zhou et al. (2022).

Generative methods further synthesize unseen-class representations from semantic information. CADA-VAE aligns modality-specific latent distributions in a shared variational space Schonfeld et al. (2019); SA-DVAE introduces disentangled representations and domain calibration Li et al. (2024); and FS-VAE combines frequency-aware skeleton features with multi-level semantics Wu et al. (2025). These methods improve visual–semantic correspondence at the instance or distribution level. However, their objectives do not explicitly coordinate modality-specific class-center consistency with the local separation of neighboring action classes. Our work focuses on this class-level relation between cross-modal alignment and local discrimination.

2.2 Cross-Modal Generative and Prototype Learning

Cross-modal generative learning maps visual and semantic observations into a shared latent space and uses class semantics to construct representations for categories without visual training samples Schonfeld et al. (2019). Recent skeleton methods further improve cross-modal correspondence through direct or augmented visual–text alignment Kuang et al. (2025). Such approaches mainly regulate instance- or distribution-level relations, while the organization of modality-specific class centers is not their primary optimization target.

Prototype learning represents class structure using compact representative features and has been studied in metric, representation, and zero-shot learning Li et al. (2020); Chen et al. (2021a). Recent methods also enrich semantic prototypes through prompt-guided or multi-view semantic information Zhu et al. (2025b); Li et al. (2025). PGFA further introduces prototype-guided skeleton–text feature alignment and uses unseen test-skeleton distributions during inference Zhou et al. (2025). Its setting therefore differs from the inductive protocol considered here, where unseen-class skeleton samples are not used for representation learning, validation, or parameter selection; its reported results are thus not treated as a directly controlled comparison with ours.

CPNM serves a different role from semantic prototype construction or inference-time prototype adaptation. It maintains separate visual and semantic prototype memories for seen classes and uses each modality as the class anchor of the other. These modality-swapped anchors define positive cross-modal neighborhoods, which subsequently provide the class-level reference for local hard-neighbor discrimination.

2.3 Hard-Negative and Local Discriminative Learning

Contrastive, margin-based, and hard-negative objectives are commonly used to improve discrimination near class boundaries Ge et al. (2020). Hard negatives generally provide stronger supervision than already separated alternatives, but sample-level selection can vary with mini-batch composition, while applying the same negative constraint to all classes may place unnecessary emphasis on distant categories.

HNDA defines this relation at the class level. It selects the nearest valid negative-class prototypes rather than individual negative samples and restricts the margin constraint to locally confusing classes. More importantly, the margin is measured relative to the positive cross-modal neighborhood energy defined by CPNM. Hard-negative learning is therefore not introduced as an independent global discrimination objective; it is used to preserve the local boundary of the positive cross-modal neighborhood established by CPNM.

3 Methodology

3.1 Problem Formulation and Framework Overview

Let $\mathcal{Y}^s$ and $\mathcal{Y}^u$ denote the seen and unseen action classes, where $\mathcal{Y}^s \cap \mathcal{Y}^u = \emptyset$. The training set is

$$\mathcal{D}^s = \{(\mathbf{x}_i, y_i, \mathbf{a}_{y_i})\}_{i=1}^{N_s}, \qquad y_i \in \mathcal{Y}^s, \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^{d_v}$ is the skeleton feature, y_i is its class label, and $\mathbf{a}_{y_i} \in \mathbb{R}^{d_t}$ is the corresponding semantic feature. We follow the inductive ZSL setting Pourpanah et al. (2022): skeleton samples are available only for $\mathcal{Y}^s$, while semantic features are available for $\mathcal{Y}^s \cup \mathcal{Y}^u$. No unseen-class skeleton sample is used for representation learning, validation, or parameter selection.

As shown in Fig. 2, the base model establishes cross-modal compatibility but does not explicitly determine class-center consistency or local class separation. CPNM therefore forms positive visual–semantic class neighborhoods, while HNDA constrains their boundaries against the nearest confusing classes. The two components are hierarchically coupled because the CPNM neighborhood energy is directly embedded in the HNDA objective.

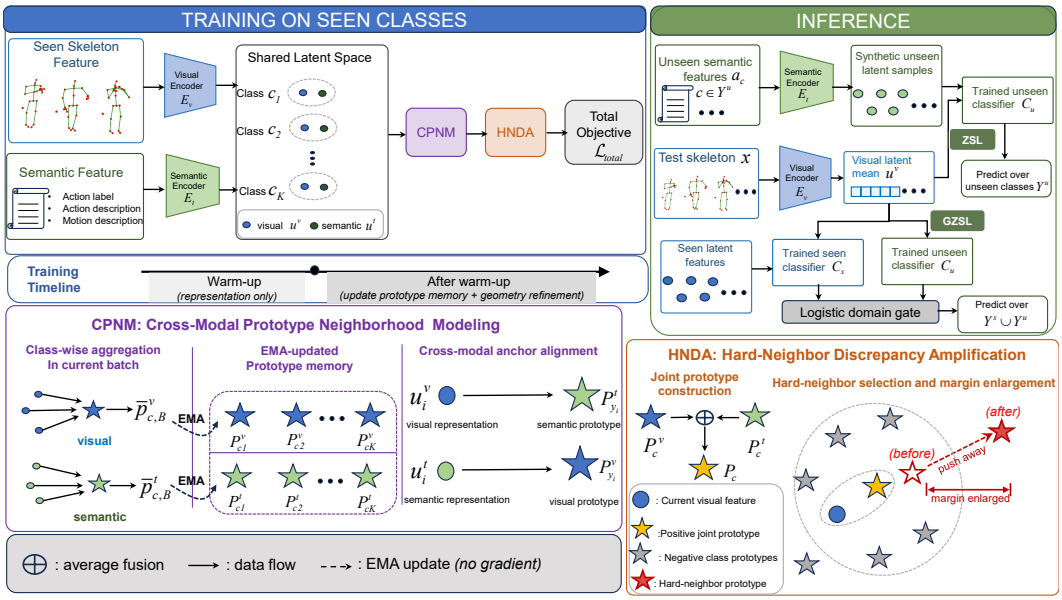


Figure 2: Overview of the proposed framework. The base model learns visual–semantic correspondence in a shared latent space. CPNM defines positive cross-modal class neighborhoods through modality-specific prototype memories, and HNDA uses the resulting neighborhood energy to impose a relative margin against locally confusing classes. During inference, unseen semantic features are used for ZSL and GZSL prediction.

3.2 Cross-Modal Latent Representation Learning

Following SA-DVAE Li et al. (2024) and FS-VAE Wu et al. (2025), we use the same pre-extracted 256-D Shift-GCN features Cheng et al. (2020), with the skeleton feature extractor fixed during cross-modal training. Given a skeleton feature $\mathbf{x}_i$ and its semantic feature $\mathbf{a}_{y_i}$, the visual and semantic encoders estimate

$$\begin{aligned}(\boldsymbol{\mu}_i^v, \boldsymbol{\ell}_i^v) &= E_v(\mathbf{x}_i), \quad q_v(\mathbf{z} \mid \mathbf{x}_i) = \mathcal{N}(\boldsymbol{\mu}_i^v, \mathrm{diag}(\exp(\boldsymbol{\ell}_i^v)))\,, \\ (\boldsymbol{\mu}_i^t, \boldsymbol{\ell}_i^t) &= E_t(\mathbf{a}_{y_i}), \quad q_t(\mathbf{z} \mid \mathbf{a}_{y_i}) = \mathcal{N}\big(\boldsymbol{\mu}_i^t, \mathrm{diag}(\exp(\boldsymbol{\ell}_i^t))\big)\,.\end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_i^v$ and $\boldsymbol{\mu}_i^t$ are the modality-specific latent means, and $\boldsymbol{\ell}_i^v$ and $\boldsymbol{\ell}_i^t$ are the corresponding log-variances. Reparameterized samples are used during training, while posterior means are used for prototype construction and test representation.

The semantic latent distribution enables sampling unseen-class representations, while KL regularization encourages compatibility between the modality-specific posteriors. The VAE is used as the generative backbone rather than the methodological contribution of this work. Its base objective is

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{rec}} + \beta_v \mathcal{L}_{\text{KL}}^v + \beta_t \mathcal{L}_{\text{KL}}^t + \lambda_{\text{ca}} \left(\mathcal{L}_{v\to t} + \mathcal{L}_{t\to v}\right), \tag{3}$$

where $\mathcal{L}_{\text{rec}}$ denotes within-modal reconstruction, $\mathcal{L}_{\text{KL}}^v$ and $\mathcal{L}_{\text{KL}}^t$ regularize the two latent distributions, and $\mathcal{L}_{v\to t}$ and $\mathcal{L}_{t\to v}$ establish cross-modal correspondence. Their detailed asymmetric definitions are provided in the supplementary material. This objective provides visual–semantic compatibility but leaves class-level geometry unspecified.

3.3 Cross-Modal Prototype Neighborhood Modeling

For each seen class c, let $\mathcal{B}_c = \{i \mid y_i = c\}$ denote its samples in the current mini-batch. The visual and semantic batch centers are

$$\overline{\mathbf{p}}_{c,B}^v = \frac{1}{|\mathcal{B}_c|} \sum_{i \in \mathcal{B}_c} \boldsymbol{\mu}_i^v, \qquad \overline{\mathbf{p}}_{c,B}^t = \frac{1}{|\mathcal{B}_c|} \sum_{i \in \mathcal{B}_c} \boldsymbol{\mu}_i^t. \tag{4}$$

CPNM maintains visual and semantic prototype memories using exponential moving average:

$$\mathbf{P}_c^m \leftarrow \rho \mathbf{P}_c^m + (1-\rho)\overline{\mathbf{p}}_{c,B}^m, \qquad m \in \{v, t\}, \tag{5}$$

where ρ is the momentum coefficient. Each prototype is initialized by its first available batch center and updated outside gradient back-propagation.

Let $\mathrm{sg}(\cdot)$ denote stop-gradient. CPNM uses modality-swapped anchors to define the positive cross-modal neighborhood energy:

$$e_i^{\mathrm{CPNM}} = \left\|\boldsymbol{\mu}_i^v - \mathrm{sg}(\mathbf{P}_{y_i}^t)\right\|_2^2 + \left\|\boldsymbol{\mu}_i^t - \mathrm{sg}(\mathbf{P}_{y_i}^v)\right\|_2^2 + \gamma \left\|\mathbf{P}_{y_i}^v - \mathbf{P}_{y_i}^t\right\|_2^2. \tag{6}$$

The first two terms pull each representation toward the same-class prototype of the opposite modality. Since the stored prototypes are detached, the last term does not directly align $\mathbf{P}_{y_i}^v$ and $\mathbf{P}_{y_i}^t$; instead, it calibrates the positive energy according to the current cross-modal prototype discrepancy. A larger discrepancy therefore produces a stronger relative constraint when e_i^{CPNM} is used by HNDA.

For the CPNM-only ablation, we use

$$\mathcal{L}_{\mathrm{CPNM}} = \frac{1}{|\mathcal{I}_{\mathrm{p}}|} \sum_{i \in \mathcal{I}_{\mathrm{p}}} e_i^{\mathrm{CPNM}}, \tag{7}$$

where $\mathcal{I}_{\mathrm{p}}$ contains samples with initialized ground-truth prototypes. Because the prototype memories are detached, the γ-weighted term does not contribute gradients in this standalone variant. $\mathcal{L}_{\mathrm{CPNM}}$ is used only for this diagnostic ablation and is not separately added to the full model.

The modality-specific memories are combined into a joint class prototype:

$$\mathbf{P}_c = \frac{1}{2}\left(\mathbf{P}_c^v + \mathbf{P}_c^t\right), \qquad c \in \mathcal{Y}^s. \tag{8}$$

Hard-Neighbor Discrepancy Amplification. Positive neighborhood modeling improves cross-modal center consistency but does not guarantee local class separation. HNDA therefore focuses the discriminative constraint on the nearest confusing classes. For a negative class $c \neq y_i$,

$$\Delta_{i,c}^v = \|\boldsymbol{\mu}_i^v - \mathrm{sg}(\mathbf{P}_c)\|_2^2, \qquad \Delta_{i,c}^t = \left\|\boldsymbol{\mu}_i^t - \mathrm{sg}(\mathbf{P}_c)\right\|_2^2, \tag{9}$$

and the joint negative-class distance is

$$\Delta_{i,c} = \frac{1}{2}\left(\Delta_{i,c}^v + \Delta_{i,c}^t\right). \tag{10}$$

Let $M_c \in \{0,1\}$ indicate whether prototype c has been initialized. The valid negative classes are

$$\mathcal{C}_i^- = \{c \in \mathcal{Y}^s \setminus \{y_i\} \mid M_c = 1\}, \qquad \Delta_{i,c_{i,1}} \leq \cdots \leq \Delta_{i,c_{i,|\mathcal{C}_i^-|}}. \tag{11}$$

The hard-neighbor set is

$$k_i = \min(k, |\mathcal{C}_i^-|), \qquad \mathcal{H}_i = \{c_{i,j}\}_{j=1}^{k_i}, \tag{12}$$

where k is the maximum number of selected hard neighbors. Samples without valid negative prototypes are omitted from HNDA.

HNDA embeds the CPNM neighborhood energy into a relative margin objective:

$$\mathcal{L}_{\mathrm{HNDA}} = \frac{1}{|\mathcal{I}_{\mathrm{h}}|} \sum_{i \in \mathcal{I}_{\mathrm{h}}} \frac{1}{|\mathcal{H}_i|} \sum_{c \in \mathcal{H}_i} \left[m + e_i^{\mathrm{CPNM}} - \Delta_{i,c}\right]_+, \tag{13}$$

where m is the margin, $[u]_+ = \max(0, u)$, and $\mathcal{I}_{\mathrm{h}}$ contains samples with at least one valid hard neighbor. The constraint becomes inactive when

$$\Delta_{i,c} \geq e_i^{\mathrm{CPNM}} + m. \tag{14}$$

Thus, CPNM defines what should remain cross-modally compact, while HNDA preserves its local boundary against the nearest confusing classes.

Table 1: Main implementation settings used in our experiments.

Setting	Value
Visual/Semantic dimensions	256 / 1,536
Latent dimension	100
Training/Warm-up iterations	10,200 / 1,700
ρ / γ	0.98 / 0.5
λ_{ca} / λ_{geo}	50 / 0.012
HNDA m / k	0.18 / 1
Generated samples R	500
Classifier/Gate	MLP / validation-trained logistic gate

3.4 Optimization Objective and Prediction

We first optimize $\mathcal{L}_{\text{base}}$ during a warm-up stage because early latent representations and prototype statistics are unstable. After warm-up, prototype memories and HNDA are activated. The full objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda_{\text{geo}}\mathcal{L}_{\text{HNDA}}, \tag{15}$$

where λ_{geo} controls the class-level geometry constraint. Since $\mathcal{L}_{\text{HNDA}}$ already contains e_i^{CPNM}, $\mathcal{L}_{\text{CPNM}}$ is not added separately.

After representation learning, unseen semantic distributions are sampled to generate latent training examples for the unseen-class expert C_u. ZSL predicts over $\mathcal{Y}^u$ using the visual posterior mean of each test skeleton. For GZSL, a seen-class expert C_s and C_u are combined by a validation-trained logistic domain gate over $\mathcal{Y}^s \cup \mathcal{Y}^u$. Gate parameters, temperatures, and fusion biases are selected on a class-disjoint validation split without final-test labels. Detailed sampling and classifier-training procedures are provided in the supplementary material.

4 Experiments

4.1 Experimental Settings

We evaluate our method on NTU RGB+D 60 Shahroudy et al. (2016), NTU RGB+D 120 Liu et al. (2019), and PKU-MMD Liu et al. (2017). We use the 55/5 and 48/12 splits on NTU-60 and the 110/10 and 96/24 splits on NTU-120. Following SA-DVAE Li et al. (2024), PKU-MMD results are averaged over three random 46/5 seen–unseen partitions. Only seen-class skeleton samples are used for representation learning and parameter selection.

Following SA-DVAE Li et al. (2024) and FS-VAE Wu et al. (2025), we use the same pre-extracted 256-D Shift-GCN features Cheng et al. (2020). Semantic representations concatenate CLIP ViT-B/32 embeddings of the action name, action-level description, and motion-level description Radford et al. (2021), yielding a 1,536-D class representation. Main implementation settings are summarized in Table 1.

The model is optimized using Adam with a batch size of 64 and a learning rate of 10^{-4}. Under ZSL, prediction is restricted to unseen classes. Under GZSL, we report seen-class accuracy S, unseen-class accuracy U, and their harmonic mean $H = 2SU/(S+U)$. All gate parameters are selected on the class-disjoint validation split without using final-test labels.

Published methods may use different semantic constructions or inference designs; therefore, the comparisons in Sec. 4.2 should be interpreted as complete-system comparisons. In contrast, all component ablations use identical visual features, semantic representations, optimization settings, generated-sample counts, classifiers, and GZSL gates. Additional analyses of semantic composition and HNDA hyperparameters are provided in the supplementary material.

4.2 Comparison with State-of-the-Art Methods

We compare with embedding-based ReViSE and JPoSE Tsai et al. (2017); Wray et al. (2019); cross-modal representation and generative methods CADA-VAE, SynSE, SMIE, SA-DVAE, and

Table 2: Comparison with published methods under the ZSL and GZSL protocols. Best and second-best ZSL and GZSL H results are shown in bold and underlined, respectively. "–" denotes unavailable results.

	NTU-60 55/5				NTU-60 48/12				NTU-120 110/10				NTU-120 96/24			
Method	**ZSL**	S	U	H	**ZSL**	S	U	H	**ZSL**	S	U	H	**ZSL**	S	U	H
ReViSE (ICCV17)	53.91	74.22	34.73	47.32	17.49	62.36	20.77	31.16	55.04	48.69	44.84	46.68	32.38	49.66	25.06	33.31
JPoSE (ICCV19)	64.82	64.44	50.29	56.49	28.75	60.49	20.62	30.75	51.93	47.66	46.40	47.05	32.44	38.62	22.79	28.67
CADA-VAE (CVPR19)	76.84	69.38	61.79	65.37	28.96	51.32	27.03	35.41	59.53	47.16	49.78	48.44	35.77	41.11	34.14	37.31
SynSE (ICIP21)	75.81	61.27	56.93	59.02	33.30	52.21	27.85	36.33	62.69	52.51	57.60	54.94	38.70	56.39	32.25	41.04
SMIE (ACM MM23)	77.98	–	–	–	40.18	–	–	–	65.74	–	–	–	45.30	–	–	–
PURLS (CVPR24)	79.23	–	–	–	40.99	–	–	–	71.95	–	–	–	52.01	–	–	–
SA-DVAE (ECCV24)	82.37	62.28	70.80	66.27	41.38	50.20	36.39	42.56	68.77	61.10	59.75	60.42	46.12	58.82	35.79	44.50
STAR (ACM MM24)	81.40	69.00	69.90	69.40	45.10	62.70	37.00	46.60	63.30	59.90	52.70	56.10	44.30	51.20	36.90	42.90
FS-VAE (ICCV25)	<u>86.90</u>	77.00	74.50	**75.70**	<u>57.20</u>	56.20	48.60	52.10	74.40	59.20	67.90	<u>63.30</u>	<u>62.50</u>	57.80	51.90	<u>54.70</u>
SCoPLe (CVPR25)	84.10	69.60	71.94	70.75	52.96	54.49	61.83	**57.93**	74.53	63.51	60.18	62.27	52.17	53.33	51.18	52.23
InfoCPL (TMM25)	85.91	–	–	–	53.32	–	–	–	<u>74.81</u>	–	–	–	60.05	–	–	–
STAR++ (TCSVT26)	84.40	61.10	73.60	66.80	49.50	58.20	40.40	47.70	72.00	59.00	55.40	57.20	53.50	52.80	45.20	48.70
Ours	**86.95**	71.58	74.79	<u>73.15</u>	**62.23**	65.46	50.40	<u>56.95</u>	**75.37**	65.66	62.41	**64.00**	**62.93**	57.17	52.54	**54.76**

Table 3: Comparison on PKU-MMD under the random 46/5 split setting. Results are averaged over three random seen–unseen class partitions. Best and second-best values are shown in bold and underlined, respectively.

Method	**ZSL**	S	U	H
ReViSE	59.34	60.89	42.16	49.82
JPoSE†	57.17	60.26	45.18	51.64
CADA-VAE	60.74	<u>63.17</u>	35.86	45.75
SynSE†	53.85	63.09	40.69	49.47
SMIE	60.83	–	–	–
SA-DVAE	<u>66.54</u>	58.49	**51.40**	<u>54.72</u>
Ours	**70.31**	**66.13**	<u>51.06</u>	**57.63**

†PoS tags for PKU-MMD are obtained using spaCy, following the evaluation protocol of SA-DVAE Li et al. (2024).

FS-VAE Schonfeld et al. (2019); Gupta et al. (2021); Zhou et al. (2023); Li et al. (2024); Wu et al. (2025); and recent fine-grained semantic or prompt-based approaches PURLS, STAR, SCoPLe, InfoCPL, and STAR++ Zhu et al. (2024); Chen et al. (2024); Zhu et al. (2025a); Xu et al. (2025); Chen et al. (2026). Existing results are taken from the corresponding papers or officially reported evaluation settings. PGFA Zhou et al. (2025) additionally exploits unseen test-skeleton distributions during inference; its protocol therefore differs from the inductive setting considered here and is not treated as a directly controlled numerical comparison.

As shown in Table 2, our method remains competitive across all four NTU splits. Under ZSL, it obtains 86.95% and 62.23% on NTU-60 55/5 and 48/12, respectively. The advantage is most evident on the more challenging 48/12 split, where it exceeds FS-VAE by 5.03 percentage points. On NTU-120, it reaches 75.37% on 110/10 and 62.93% on 96/24; the gains over the strongest published results on these splits are more modest.

Under GZSL, our method obtains harmonic means of 73.15%, 56.95%, 64.00%, and 54.76% on the four splits. On NTU-60 48/12, it improves over FS-VAE by 4.85 points in H, although SCoPLe reports a slightly higher result. On NTU-120, the proposed method reaches 64.00% and 54.76% on the 110/10 and 96/24 settings, respectively. Overall, these results show that the proposed class-level constraints maintain competitive recognition performance across different seen–unseen partitions.

Table 3 provides an additional evaluation on PKU-MMD under the random-split protocol of SA-DVAE Li et al. (2024). Our method obtains an average ZSL accuracy of 70.31%, improving over SA-DVAE by 3.77 points. Under GZSL, it reaches an average H of 57.63%. The seen-class accuracy increases from 58.49% to 66.13%, while the unseen-class accuracy remains comparable at 51.06%

Table 4: Component ablation results on NTU-60 55/5 and NTU-120 110/10. Best ZSL and GZSL H results are shown in bold.

			NTU-60 55/5				NTU-120 110/10			
Variant	**CPNM**	**Negative Strategy**	**ZSL**	S	U	H	**ZSL**	S	U	H
A0		None	83.73	65.20	73.33	69.03	73.86	46.23	70.94	55.98
A1	✓	None	84.73	71.92	69.51	70.69	75.11	55.23	68.86	61.30
A2	✓	Random	85.62	66.82	78.24	72.08	74.49	52.68	69.69	60.00
A3	✓	Hard neighbor	**86.95**	71.58	74.79	**73.15**	**75.37**	65.66	62.41	**64.00**

Table 5: Ablation of training strategies on the NTU-120 110/10 split.

Configuration	**ZSL**	S	U	H
CPNM + HNDA	**75.37**	65.66	62.41	**64.00**
w/o EMA ($\rho = 0$)	75.32	**65.99**	61.66	63.75
w/o prototype-discrepancy calibration ($\gamma = 0$)	75.34	64.25	62.20	63.20
w/o warm-up	71.55	55.23	**66.23**	60.23

versus 51.40%, resulting in a 2.91-point improvement in H. These results provide complementary evidence on a different skeleton benchmark and across multiple random seen–unseen partitions.

4.3 Component Ablation Study

We construct four controlled variants to separate positive cross-modal neighborhood modeling from negative-class selection. A0 is the dual-VAE baseline with $\lambda_{\text{geo}} = 0$. A1 replaces HNDA with the standalone CPNM objective in Eq. (7), retaining positive neighborhood modeling without negative-class discrimination. A2 retains the relative-margin objective but replaces hard neighbors with randomly sampled negative classes. A3 is the complete CPNM+HNDA model. All variants use identical features, semantic representations, optimization settings, generated-sample counts, classifiers, and GZSL gates.

As shown in Table 4, A1 consistently improves over A0. On NTU-60, ZSL accuracy increases from 83.73% to 84.73% and H increases from 69.03% to 70.69%. On NTU-120, ZSL improves from 73.86% to 75.11%, while H increases from 55.98% to 61.30%. These results show that positive cross-modal neighborhood modeling provides useful class-level supervision beyond the base latent objective. Its geometric effect is examined further in Sec. 4.4.

A2 introduces randomly sampled negative classes on top of CPNM. Although it improves over A1 on NTU-60, its H decreases from 61.30% to 60.00% on NTU-120, showing that simply adding a negative-class margin does not consistently improve transfer. A3 instead targets the nearest confusing classes and achieves the highest ZSL accuracy and H on both splits. Relative to A2, HNDA improves H from 72.08% to 73.15% on NTU-60 and from 60.00% to 64.00% on NTU-120. This comparison isolates the effect of hard-neighbor selection from the presence of the negative margin itself.

Training strategies. We further examine prototype-memory updating, prototype-discrepancy calibration, and warm-up activation. Each variant modifies only the corresponding strategy while retaining the remaining components of the full model.

As shown in Table 5, replacing EMA memories with current mini-batch centers causes only a small reduction in H, suggesting that EMA mainly stabilizes cross-batch prototype statistics. Setting $\gamma = 0$ removes the prototype-discrepancy calibration from e_i^{CPNM} while retaining modality-swapped anchoring and hard-neighbor discrimination. Since the stored prototypes are detached, this term does not directly align the prototype memories; instead, it modulates the HNDA constraint according to the current cross-modal prototype discrepancy. Removing it reduces H from 64.00% to 63.20%.

Table 6: Class-level geometry analysis of the seen classes on NTU-120 110/10.

Variant	Cross-modal Dist. ↓	Nearest-neg. Dist. ↑	Inter/Intra ↑	ZSL	S	U	H
A0	0.572	0.182	1.474	73.86	46.23	70.94	55.98
A1	0.145	0.116	1.416	75.11	55.23	68.86	61.30
A3	0.204	0.146	1.539	**75.37**	**65.66**	62.41	**64.00**

Removing warm-up has a larger effect, reducing ZSL accuracy from 75.37% to 71.55% and H from 64.00% to 60.23%. This supports delaying prototype-based geometry learning until the latent representations and class statistics become more stable. Sensitivity to the HNDA margin m and hard-neighbor count k is reported in the supplementary material.

4.4 Cross-Modal Alignment and Local Separation Analysis

We further examine how CPNM and HNDA reshape the class-level geometry of the seen classes on NTU-120 110/10. Cross-modal distance measures the discrepancy between visual and semantic prototypes of the same class. Nearest-negative distance measures the distance to the closest negative-class prototype, while the inter-to-intra-class ratio measures class separation relative to within-class dispersion. Smaller cross-modal distance indicates stronger center consistency, whereas larger nearest-negative distance and inter/intra ratio indicate stronger separation.

The training objectives use squared Euclidean distance. For this post-training diagnostic analysis, representations are L_2-normalized and the geometric statistics are computed using cosine distance. These quantities are used only to diagnose the learned seen-class geometry and are not additional training objectives.

As shown in Table 6, moving from A0 to A1 sharply reduces the cross-modal distance from 0.572 to 0.145, confirming that CPNM improves visual–semantic class-center consistency. However, the nearest-negative distance decreases from 0.182 to 0.116 and the inter/intra ratio decreases from 1.474 to 1.416. Thus, improving positive cross-modal alignment alone does not guarantee stronger local separation. After introducing HNDA, A3 increases the nearest-negative distance from 0.116 to 0.146 and the inter/intra ratio from 1.416 to 1.539 relative to A1. Although its cross-modal distance increases from 0.145 to 0.204, it remains substantially below the A0 value of 0.572. HNDA therefore recovers local separation relative to CPNM-only while retaining the improved center consistency established by CPNM. This behavior agrees with the relative-margin design in Eq. (13), which protects the boundary of the positive neighborhood rather than maximizing all inter-class distances. These diagnostics characterize the geometry learned on seen classes during training. Their transfer to unseen actions is evaluated separately through the ZSL and GZSL recognition results in Sec. 4.2. Additional prototype-distance distributions are provided in the supplementary material.

5 Conclusion

This work studies zero-shot skeleton action recognition from the joint perspective of cross-modal class alignment and local class separation. We introduce a hierarchical class-geometry objective in which CPNM forms positive visual–semantic class neighborhoods through modality-specific prototype memories and modality-swapped anchoring, while HNDA preserves their local boundaries by imposing a relative margin against the nearest confusing classes. Experiments on NTU RGB+D 60, NTU RGB+D 120, and PKU-MMD show competitive performance under both ZSL and GZSL settings. Controlled ablations further separate the contributions of positive neighborhood modeling and hard-neighbor selection. In addition, seen-class geometry analyses show that CPNM substantially reduces cross-modal center discrepancy, whereas HNDA recovers local separation that is not guaranteed by positive alignment alone. These results support jointly modeling cross-modal consistency and local discrimination for transferring semantic knowledge to unseen skeleton actions.

Reproducibility Statement

The main text specifies the task formulation, dataset splits, visual and semantic feature construction, model architecture, training objectives, key hyperparameters, and ZSL/GZSL evaluation protocols. All model variants use identical visual features, semantic representations, generated-sample counts, classifiers, and GZSL inference procedures, with gate parameters selected only on a class-disjoint validation split and without using final-test labels. The supplementary material further provides the complete definitions of the base cross-modal objective, unseen-class sample generation and ZSL/GZSL classification procedures, as well as additional analyses of HNDA hyperparameters, semantic inputs, and prototype-distance distributions. Together, these details provide the information required to reproduce the training, inference, and main experimental results reported in this paper.

AI Use Statement

The authors used generative AI tools solely for language polishing, improving manuscript organization and clarity, and assisting with LaTeX formatting. These tools were not used to design the methodology, conduct experiments, generate or modify experimental data, analyze experimental results, or formulate scientific conclusions. All manuscript content was carefully reviewed and verified by the authors, who take full responsibility for the final version and its accuracy.

Ethics Statement

This work uses existing public benchmark datasets and pre-extracted skeleton features, and does not involve the collection of new human-subject data. Although skeleton representations remove much of the appearance information contained in RGB videos, human motion data may still encode identity-related or demographic cues and may reflect biases present in the source datasets. The proposed method is intended for research on action recognition and zero-shot transfer, and its performance should not be interpreted as evidence of fairness, safety, or suitability for high-stakes applications such as surveillance or automated decision making. Any reuse of the datasets and derived features should follow the corresponding dataset licenses and usage restrictions.

References

Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. FREE: Feature refinement for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 122–131, 2021a.

Yang Chen, Jingcai Guo, Tian He, and Ling Wang. Fine-grained side information guided dual-prompts for zero-shot skeleton action recognition. *arXiv preprint arXiv:2404.07487*, 2024.

Yang Chen, Jingcai Guo, Miaoge Li, Zhijie Rao, and Song Guo. Star++: Region-aware conditional semantics via interpretable side information for zero-shot skeleton action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. 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, 2021b.

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

Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object Re-ID. *Advances in neural information processing systems*, 33:11309–11321, 2020.

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

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

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

Ming-Zhe Li, Zhen Jia, Zhang Zhang, Yaoning Li, Zhanyu Ma, and Liang Wang. Multi-view knowledge guided semantic prototype learning for generalized zero-shot action recognition. *IEEE Transactions on Multimedia*, 2025.

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

Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.

Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4051–4070, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8247–8255, 2019.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12026–12035, 2019.

Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. *arXiv preprint arXiv:1703.05908*, 2017.

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

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

Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4582–4591, 2017.

Haojun Xu, Yan Gao, Jie Li, and Xinbo Gao. An information compensation framework for zero-shot skeleton-based action recognition. *IEEE Transactions on Multimedia*, 2025.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022.

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

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

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

Xingyu Zhu, Xiangbo Shu, Peng Huang, and Jinhui Tang. Prompt-guided prototype-aware commonality and discrimination learning for zero-shot skeleton-based action recognition. *IEEE Transactions on Multimedia*, 2025b.

A Additional Method Details

A.1 Detailed Cross-Modal Latent Learning

This section expands the base cross-modal objective summarized in Eq. (3) of the main paper. Given the modality-specific distributions defined in Eq. (2), stochastic latent representations are obtained through reparameterized sampling:

$$\mathbf{z}_i^m = \boldsymbol{\mu}_i^m + \exp\left(\frac{1}{2}\boldsymbol{\ell}_i^m\right) \odot \boldsymbol{\epsilon}_i^m, \qquad \boldsymbol{\epsilon}_i^m \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad m \in \{v, t\}, \tag{16}$$

where $m = v$ and $m = t$ denote the visual and semantic modalities, respectively, and $\odot$ denotes element-wise multiplication.

The visual and semantic latent codes are decoded within their own modalities to preserve the information contained in the corresponding inputs. The resulting variational objective is

$$\mathcal{L}_{\text{vae}} = \mathcal{L}_{\text{rec}} + \beta_v \mathcal{L}_{\text{KL}}^v + \beta_t \mathcal{L}_{\text{KL}}^t, \tag{17}$$

where $\mathcal{L}_{\text{rec}}$ contains the visual and semantic within-modal reconstruction errors. The terms $\mathcal{L}_{\text{KL}}^v$ and $\mathcal{L}_{\text{KL}}^t$ are defined as non-negative KL divergences from the corresponding posterior distributions to a standard Gaussian prior, while β_v and β_t are their weighting coefficients.

Within-modal reconstruction alone does not establish direct correspondence between skeleton and semantic representations. The visual latent code is therefore decoded by the semantic decoder, while the semantic latent code is decoded by the visual decoder. Let

$$\widehat{\mathbf{a}}_i^{v\to t} = D_t(\mathbf{z}_i^v), \qquad \widehat{\mathbf{x}}_i^{t\to v} = D_v(\mathbf{z}_i^t), \tag{18}$$

where $\widehat{\mathbf{a}}_i^{v\to t}$ and $\widehat{\mathbf{x}}_i^{t\to v}$ denote the outputs of the semantic decoder D_t and visual decoder D_v, respectively.

For the visual-to-semantic direction, a random permutation π is sampled within the current mini-batch. The discrepancy between the permuted and paired semantic reconstructions is

$$\Delta_i^{v\to t} = \left\|\mathbf{a}_{y_i} - \widehat{\mathbf{a}}_{\pi(i)}^{v\to t}\right\|_2^2 - \left\|\mathbf{a}_{y_i} - \widehat{\mathbf{a}}_i^{v\to t}\right\|_2^2, \tag{19}$$

where $(\mathbf{a}_{y_i}, \widehat{\mathbf{a}}_i^{v\to t})$ is the paired cross-modal reconstruction and $(\mathbf{a}_{y_i}, \widehat{\mathbf{a}}_{\pi(i)}^{v\to t})$ is its randomly permuted counterpart. The corresponding calibrated term is

$$\mathcal{L}_{v\to t} = \frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \frac{1}{1 + \exp\left(\Delta_i^{v\to t}/\tau\right)}, \tag{20}$$

where $\mathcal{B}$ is the current mini-batch and τ is the temperature parameter.

For the semantic-to-visual direction, the model directly measures the paired visual reconstruction discrepancy:

$$d_i^{t\to v} = \left\|\mathbf{x}_i - \widehat{\mathbf{x}}_i^{t\to v}\right\|_2^2, \tag{21}$$

and the corresponding loss is

$$\mathcal{L}_{t\to v} = \frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \frac{1}{1 + \exp\left(-d_i^{t\to v}/\tau\right)}. \tag{22}$$

where $\mathbf{x}_i$ is the pre-extracted Shift-GCN feature entering the visual branch. The random permutation is used only in the visual-to-semantic direction, whereas the semantic-to-visual direction directly reconstructs the paired visual feature.

The two cross-modal terms are summarized as

$$\mathcal{L}_{\text{ca}} = \mathcal{L}_{v\to t} + \mathcal{L}_{t\to v}. \tag{23}$$

Combining Eqs. (17) and (23) gives

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{rec}} + \beta_v \mathcal{L}_{\text{KL}}^v + \beta_t \mathcal{L}_{\text{KL}}^t + \lambda_{\text{ca}} \left(\mathcal{L}_{v\to t} + \mathcal{L}_{t\to v}\right), \tag{24}$$

which is the expanded form of the base objective used in the main paper.

A.2 ZSL AND GZSL PREDICTION

This section provides the classifier-training and prediction details summarized in Sec. 3.4 of the main paper. After representation learning, the semantic feature of each unseen class defines a latent distribution through the semantic encoder. Repeated sampling produces a synthetic unseen-class training set:

$$\mathbf{z}_{c,r}^{t} \sim q_t\left(\mathbf{z} \mid \mathbf{a}_c\right), \qquad \mathcal{D}_{\mathrm{syn}}^{u} = \left\{\left(\mathbf{z}_{c,r}^{t}, c\right) \mid c \in \mathcal{Y}^u,\ 1 \le r \le R\right\}, \tag{25}$$

where R is the number of latent samples generated for each unseen class.

The unseen-class expert C_u is trained on $\mathcal{D}_{\mathrm{syn}}^{u}$ using

$$\mathcal{L}_{\mathrm{cls}}^{u} = -\frac{1}{R|\mathcal{Y}^u|} \sum_{c \in \mathcal{Y}^u} \sum_{r=1}^{R} \log P_{C_u}\left(c \mid \mathbf{z}_{c,r}^{t}\right), \tag{26}$$

where $P_{C_u}(c \mid \mathbf{z})$ is the probability assigned to class c by C_u.

During ZSL inference, a test skeleton feature $\mathbf{x}$ is passed through the visual encoder and represented by its visual posterior mean $\boldsymbol{\mu}^v$. The prediction is

$$\widehat{y}_{\mathrm{ZSL}} = \arg\max_{c \in \mathcal{Y}^u} P_{C_u}\left(c \mid \boldsymbol{\mu}^v\right), \tag{27}$$

where the candidate label space contains only unseen classes.

For GZSL, a seen-class expert C_s is trained from the available seen-class visual representations and their labels, while C_u is trained from the synthetic unseen-class set in Eq. (25). At test time, the visual posterior mean of each sample is evaluated by both experts. A logistic domain gate uses confidence statistics derived from their outputs to select or fuse predictions over $\mathcal{Y}^s \cup \mathcal{Y}^u$.

The domain gate is applied independently to each test sample. Its parameters, together with the expert temperatures and fusion biases, are selected on a class-disjoint validation split. No final-test label is used for gate training or model selection, and the same dual-expert protocol is applied to all variants of our method.

B ADDITIONAL ABLATION STUDIES

B.1 HYPERPARAMETER SENSITIVITY

As shown in Table 7, we examine the HNDA margin m and the maximum number of hard neighbors k on the NTU-120 110/10 split. Changing m from 0.10 to 0.30 produces ZSL accuracies between 75.27% and 75.37% and GZSL harmonic means between 63.49% and 64.00%, indicating relatively stable performance within the evaluated range. Among the three settings, $m = 0.18$ gives the highest GZSL H, while $m = 0.30$ yields a slightly higher seen-class accuracy but a lower unseen-class accuracy. We therefore use $m = 0.18$ in the main experiments.

Using $k = 1$ produces the highest ZSL accuracy and GZSL harmonic mean. Increasing k to three or five reduces the overall harmonic mean, which is consistent with the motivation for concentrating HNDA on the most locally confusing negative classes rather than expanding the constraint to additional, less-confusing neighbors. We therefore use $k = 1$ as the default setting.

B.2 SEMANTIC COMPONENT ANALYSIS

As shown in Table 8, we evaluate different combinations of the semantic inputs while keeping the visual features, architecture, training strategy, and inference protocol unchanged. LB, AD, and MD denote the action label, action-level description, and motion-level description, respectively.

The action label provides a direct category-level semantic anchor, while the action-level description introduces broader information about the action and the motion-level description emphasizes motion-related cues. Evaluating these components separately and in combination allows us to examine how different forms of semantic information contribute under the same visual and inference settings.

Table 7: Sensitivity to the HNDA margin m and the number of hard neighbors k on the NTU-120 110/10 split.

Parameter	Setting	ZSL	S	U	H
Margin m	0.10	75.27	64.50	**62.86**	63.67
	0.18	**75.37**	65.66	62.41	**64.00**
	0.30	**75.37**	**65.99**	61.16	63.49
Hard neighbors k	1	**75.37**	**65.66**	62.41	**64.00**
	3	75.08	64.96	60.76	62.79
	5	75.13	64.57	**62.93**	63.74

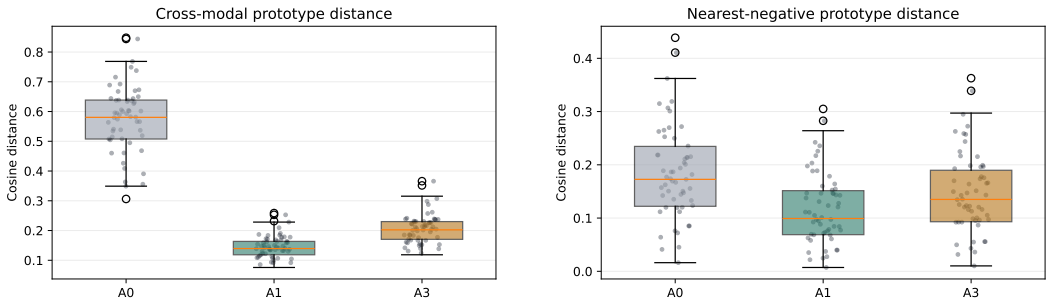
Table 8: Effect of different semantic components on the NTU-120 110/10 split.

Semantic Input	ZSL	S	U	H
LB	72.80	61.40	59.20	60.28
AD	69.60	58.10	55.40	56.72
MD	70.90	59.30	57.80	58.54
AD+MD	72.10	60.00	60.70	60.35
LB+AD+MD	**75.37**	**65.66**	**62.41**	**64.00**

The complete LB+AD+MD representation achieves the highest ZSL accuracy and GZSL harmonic mean among the evaluated semantic configurations. Compared with the individual components and the AD+MD combination, jointly using the label, action-level description, and motion-level description produces the strongest overall result. This indicates that the three semantic components provide complementary information under our setting.

C Additional Geometry Analysis

C.1 Prototype-Distance Distributions



(a) Cross-modal prototype distance.

(b) Nearest-negative prototype distance.

Figure 3: Prototype-distance distributions of seen classes on the NTU-120 110/10 split. CPNM shifts the cross-modal distance toward smaller values, while HNDA increases the nearest-negative distance relative to the CPNM-only variant.

Figure 3 provides a distribution-level view of the seen-class geometry corresponding to the averaged statistics reported in Table 6 of the main paper. The analysis is performed on the seen classes of the NTU-120 110/10 split using the same L_2-normalized representations and cosine-distance metric as in the main-text geometry analysis.

For the cross-modal prototype distance, both A1 and A3 shift toward smaller values than A0, showing that the reduction in visual–semantic prototype discrepancy occurs across classes rather than being determined by a few isolated categories. A1 exhibits the strongest shift toward smaller cross-modal distances, consistent with the center-consistency effect of CPNM. A3 retains a substantially smaller cross-modal distance than A0 after the hard-neighbor constraint is introduced.

For the nearest-negative prototype distance, A3 shifts toward larger values relative to A1. This trend is consistent with the averaged results in Table 6, where the nearest-negative distance increases from 0.116 for A1 to 0.146 for A3. The distribution therefore provides complementary evidence that HNDA recovers local separation relative to the CPNM-only variant. At the same time, this result should not be interpreted as A3 producing a larger nearest-negative distance than the base model A0; rather, HNDA recovers part of the separation reduced by positive neighborhood modeling while preserving substantially improved cross-modal center consistency over A0.

These distributions are computed only on seen classes and are used to diagnose the geometry learned during training. They do not directly measure the geometry of unseen classes. Transfer to unseen actions is evaluated separately by the ZSL and GZSL recognition results reported in the main paper.