Gated Contrastive Alignment Variational Autoencoder for Zero-Shot Skeleton Action Recognition

Jiang Liu[1], Rong Liu[2], Ming Liu[1,*]
[1]School of Computer Science, Central China Normal University, Wuhan, China
[2]School of Physical Science and Technology, Central China Normal University, Wuhan, China
[1,*]School of Computer Science, Central China Normal University, Wuhan, China
Emails: liujiang258@mails.ccnu.edu.cn, liurong@mail.ccnu.edu.cn, lium@ccnu.edu.cn

***Abstract*—Skeleton-based action recognition has attracted increasing attention due to its efficiency and robustness to visual variations, and plays an important role in applications such as human–computer interaction and healthcare. To alleviate the reliance on labeled data for rare or hazardous actions, Zero-Shot Skeleton Action Recognition (ZSSAR) exploits semantic knowledge to recognize unseen classes. However, skeleton sequences are inherently noisy and ambiguous, especially for fine-grained actions with similar motion patterns. Existing cross-modal methods adopt fixed latent space decompositions and perform coarse-grained skeleton–text alignment, often enforcing hard semantic–style disentanglement. Such rigid designs may discard useful correlations, amplify noise, and limit generalization when semantic and style factors are not strictly separable. To address these challenges, we propose Gated Contrastive Alignment Variational Autoencoder (GCA-VAE), a novel framework for robust ZSSAR. GCA-VAE introduces a learnable gating mechanism that softly modulates latent dimensions, enabling adaptive separation of semantic-relevant and style-related factors without relying on predefined partitions. In addition, we design a fine-grained contrastive alignment strategy to calibrate skeleton and text representations, which suppresses skeleton noise while preserving discriminative semantic cues. Extensive experiments on the NTU RGB+D 60 and NTU RGB+D 120 datasets demonstrate that GCA-VAE consistently outperforms state-of-the-art methods on both ZSSAR and generalized ZSSAR benchmarks.**

***Index Terms*—Zero-Shot Skeleton Action Recognition, Cross-Modal Representation Learning , Learnable Gating Mechanism, Variational Autoencoder**

I. Introduction

With the rapid growth of applications such as human–computer interaction, sports analytics, healthcare monitoring, and multimedia entertainment, recognizing human actions has emerged as an important research topic. Among various approaches, skeleton-based action recognition has attracted increasing interest because it offers high computational efficiency and shows strong robustness to changes in illumination, viewpoint, and complex backgrounds. Despite remarkable progress, most existing methods [1]–[7] heavily rely on fully supervised learning, which demands comprehensive labeled data for all action categories. This requirement becomes impractical when dealing with rare, unseen, or dangerous actions. To mitigate this limitation, Zero-Shot Learning (ZSL) has been explored for SAR, where unseen action classes can be recognized through auxiliary semantic information such as class names, attributes, or natural language descriptions.

Learning disentangled and semantically aligned representations remains a central challenge in multimodal action recognition. Skeleton-based representations are compact and structured, offering robustness against appearance variations and environmental noise. However, skeleton data alone is often ambiguous: actions with highly similar motion patterns (e.g., “drinking water” vs. “eating a meal”) are difficult to distinguish, and viewpoint changes can further amplify noise. In contrast, natural language provides clean, semantically rich cues that are critical for zero-shot generalization. A key challenge, therefore, is bridging noisy skeleton features with precise text semantics through effective representation learning.

Most existing methods rely on hard latent space disentanglement and coarse alignment strategies. Specifically, skeleton features are usually factorized into fixed semantic and style subspaces, and alignment with text representations is conducted at a global level. Such rigid partitioning risks introducing information loss when semantic and style dimensions are not strictly separable, while global alignment may overlook subtle semantic distinctions. More importantly, skeleton sequences are inherently noisy: viewpoint variations, intra-class motion diversity, and imperfect capture can blur the boundary between semantic-relevant and irrelevant factors, making hard disentanglement both brittle and suboptimal.

To overcome these limitations, we propose GCA-VAE (Gated Contrastive Alignment Variational Autoencoder), a novel framework designed to improve robustness and generalization in ZSSAR. GCA-VAE introduces two innovations. First, a learnable gating mechanism adaptively partitions latent dimensions into semantic and style-oriented subspaces. Unlike hard separation, this soft gating dynamically adjusts across datasets and actions, allowing flexible representation learning

* Corresponding author.

without relying on fixed priors (Fig. 1). Second, we design a calibrated contrastive alignment loss that performs fine-grained skeleton–text calibration, reducing the propagation of skeleton noise while preserving semantic fidelity. Together, these designs allow the model to robustly disentangle useful semantics and achieve stronger cross-modal alignment. Our contributions are summarized as follows:

- We introduce a learnable gating mechanism that replaces rigid latent partitioning with a soft, data-driven disentanglement strategy.
- We develop a calibrated contrastive alignment loss that mitigates skeleton noise while enhancing semantic alignment with text features.
- Extensive experiments on NTU RGB+D 60 and NTU RGB+D 120 datasets demonstrate that GCA-VAE achieves state-of-the-art performance on both ZSSAR and GZSSAR benchmarks.

I will provide the data and code after the paper is accepted.

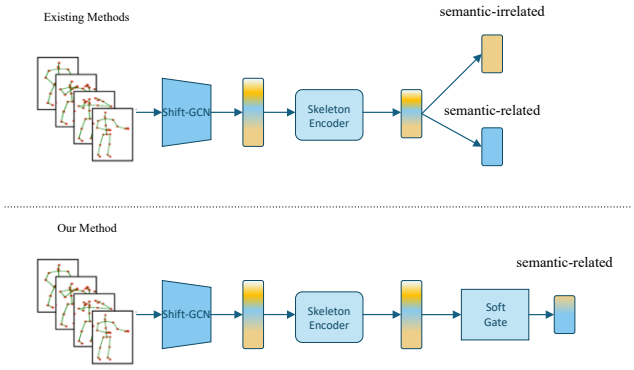


Fig. 1: Comparison with existing methods. Our method assign each latent dimension a learnable semantic relevance score via a gating mechanism. Instead of enforcing strict separability, soft gating adaptively modulates the contribution of each dimension to cross-modal alignment, preserving mixed but discriminative representations while suppressing noisy factors.

II. Related Work

A. Skeleton-Based Zero-Shot Action Recognition

A significant portion of current research focuses on the semantic side [8]–[10], under the premise that refined textual descriptions can better guide the discovery of latent action patterns. However, the reliance on LLM-generated content introduces a "generalization bottleneck" as the inconsistency of generated text can act as a form of semantic noise. Simultaneously, while advancements have been made in skeleton feature extraction [11]–[14], the issue of inherent data noise remains a critical obstacle. In standard skeleton recognition, the model can often learn to ignore non-essential noise through supervised training on a fixed label set. In contrast, ZSL requires the model to generalize knowledge to entirely novel categories via a cross-modal bridge. This makes the framework exceptionally sensitive to feature corruption. If the skeleton features are noisy, the resulting visual-textual alignment becomes fragile, leading to catastrophic interference when the model attempts to project unseen samples into a compromised semantic space.

B. Feature Decoupling in Generalized Zero-Shot Learning

To combat the persistent challenge of domain shift in GZSL, feature disentanglement has emerged as a critical strategy. By isolating class-discriminative information from non-semantic noise, models can achieve superior generalization across seen and unseen domains. Traditional frameworks, such as SDGZSL [15], implement this by partitioning visual representations into semantic-consistent and irrelevant subspaces, subsequently leveraging a relation network to align the former with class embeddings. Despite its efficacy in knowledge transfer, such an approach is essentially anchored to static, predefined attribute sets. This reliance creates a bottleneck when dealing with complex datasets where semantic structures are highly fluid or insufficiently captured by manual annotations.

In contrast, SA-DVAE [14] eliminates the reliance on intermediate semantic attributes and directly utilizes textual descriptions for cross-modal alignment. By jointly modeling semantic-consistent and semantic-irrelevant factors within a variational framework, SA-DVAE improves flexibility compared to attribute-based methods. However, SA-DVAE still adopts a hard partition of the latent space, where semantic and non-semantic factors are assumed to be strictly separable. Such an assumption may be overly restrictive, especially for complex visual modalities where semantic cues and nuisance variations are inherently entangled. Moreover, the alignment between modalities is typically performed at a global level, which may amplify the influence of noisy or unstable feature dimensions. Recent works in cross-modal representation learning have also explored disentanglement through adversarial objectives or dual-branch architectures. While effective in certain settings, these approaches often suffer from unstable training or rely on strong independence assumptions, limiting their robustness when semantic boundaries are ambiguous. These limitations motivate the need for more flexible decoupling strategies that avoid rigid latent partitioning and can adaptively regulate the contribution of different feature dimensions during cross-modal alignment.

III. Methodology

A. Problem Formulation

In the context of skeleton-based recognition, we define a label space divided into two disjoint subsets: seen categories $\mathcal{L}^s$ and unseen categories $\mathcal{L}^u$, where $\mathcal{L}^s \cap \mathcal{L}^u = \varnothing$. The available data $\mathcal{D}$ is categorized into training samples $\mathcal{D}^s_{tr}$ and test samples $\mathcal{D}_{te} = \mathcal{D}^s_{te} \cup \mathcal{D}^u_{te}$. Specifically, $\mathcal{D}^s_{tr}$ contains N_{tr} sequences $\{x^s_i, y^s_i\}$ exclusively from the seen classes. Our methodology is validated through two distinct evaluation paradigms. The ZSL paradigm requires the model to map input

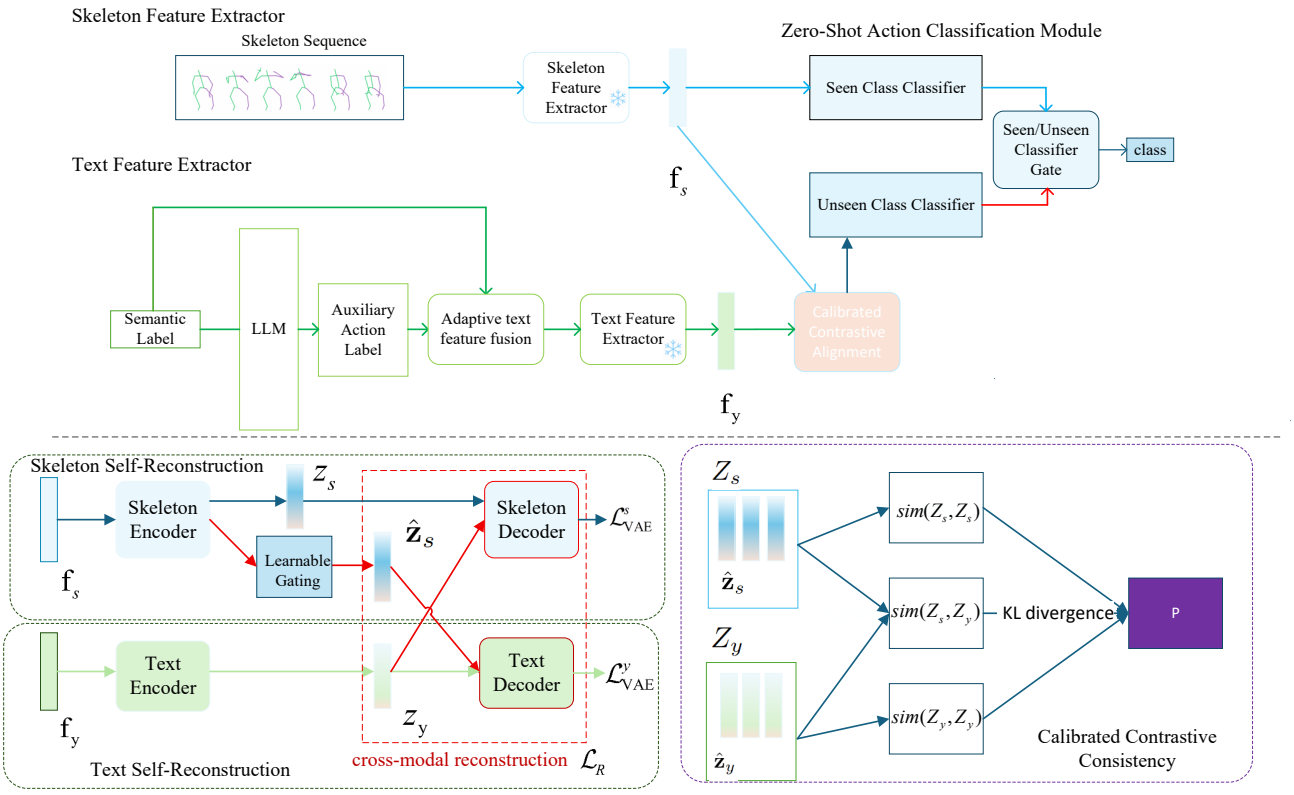


Fig. 2: The pipeline of the proposed method.

sequences x_i^u to the unseen space $\mathcal{L}^u$ without prior exposure to these classes during training. Conversely, the GZSL setting relaxes this restriction, demanding the model to categorize test instances within the unified space $\mathcal{L}^s \cup \mathcal{L}^u$. This shift from ZSL to GZSL evaluates the model's capacity to maintain performance on known classes while generalizing to novel ones. The overall architecture is outlined in Fig. 2.

B. Skeleton Modality with Learnable Gating

A central challenge in cross-modal representation learning lies in the intrinsic entanglement between semantic motion patterns and skeleton-specific variations, such as viewpoint changes, body scale differences, and capture noise. Enforcing hard semantic–style separation in the latent space may therefore be brittle and suboptimal, especially when such factors are not strictly separable.

To address this issue, we introduce a learnable gating mechanism that models the semantic relevance of each latent dimension, enabling soft and adaptive modulation rather than rigid partitioning. Given skeleton features $\mathbf{f}_s$ extracted by a pretrained backbone, the skeleton encoder produces latent variables via the reparameterization trick:

$$\mathbf{z}_s = \mu_s + \sigma_s \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\mathbf{z}_s \in \mathbb{R}^d$ denotes the latent representation of a single skeleton sample.

A learnable gating vector $\mathbf{g} \in \mathbb{R}^d$ is introduced to estimate the semantic relevance of each latent dimension:

$$\mathbf{g}_{mask} = \sigma(\mathbf{g}), \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function. The gating mask softly modulates the latent representation:

$$\mathbf{z}_s^r = \mathbf{z}_s \odot \mathbf{g}_{mask}, \quad \mathbf{z}_s^v = \mathbf{z}_s \odot (1 - \mathbf{g}_{mask}), \tag{3}$$

where $\mathbf{z}_s^r$ emphasizes semantically relevant factors, while $\mathbf{z}_s^v$ captures complementary variations such as motion style and noise. Importantly, this decomposition is soft and dimension-wise, allowing partially semantic dimensions to be preserved rather than discarded.

The skeleton VAE objective regularizes the latent distribution as:

$$\begin{aligned}\mathcal{L}_{\text{VAE}}^s = \mathbb{E}_{q_\phi(\mathbf{z}_s|\mathbf{f}_s)}\left[\log p_\theta(\mathbf{f}_s|\mathbf{z}_s)\right] \\ - \beta_s^r \, \mathbb{E}\left[\mathbf{g}_{mask}\right] \cdot D_{\text{KL}}\left(q_\phi(\mathbf{z}_s|\mathbf{f}_s) \,\|\, p(\mathbf{z}_s)\right),\end{aligned} \tag{4}$$

where the expected gating activation controls the effective strength of the KL regularization on semantically relevant dimensions.

To prevent degenerate solutions where all gating values collapse to trivial extremes, we introduce a lightweight regularization on the gating vector during training. Specifically, the regularization consists of (i) a mean activation constraint that controls the overall proportion of semantic dimensions, and (ii)

an entropy-based term that encourages diversity across gating responses. This regularization promotes a balanced activation ratio while avoiding hard binary decisions. The corresponding regularization term is weighted by a coefficient λ_g in the overall training objective.

C. Adaptive text feature fusion

Compared with skeleton data, textual descriptions naturally encode high-level semantic information and are less affected by modality-specific variations. Recent works increasingly leverage LLMs to generate enriched textual descriptions for action categories, aiming to provide more fine-grained semantic supervision. However, such generated descriptions often exhibit heterogeneous quality and may introduce modality-irrelevant or stylistic factors that are not directly observable in skeleton sequences.

To address this issue, we adopt an adaptive text feature fusion strategy that calibrates the contribution of multiple textual sources in a class-aware manner. Specifically, we construct an expanded text representation by combining the original action label with multiple auxiliary descriptions generated by an LLM. Each textual source is encoded independently, and their contributions are modulated by learnable fusion weights, allowing the model to emphasize semantically reliable cues while suppressing noisy or skeleton-irrelevant information.

Formally, let $\{\mathbf{f}_y^{(k)}\}_{k=1}^{K}$ denote text features from K textual sources. We introduce a set of class-specific fusion weights $\mathbf{w}_c \in \mathbb{R}^K$, normalized via softmax, and obtain the fused text representation as:

$$\mathbf{f}_y = \bigoplus_{k=1}^{K} \alpha_{c,k}\, \mathbf{f}_y^{(k)}, \quad \alpha_{c,k} = \frac{\exp(w_{c,k})}{\sum_{j=1}^{K} \exp(w_{c,j})}, \tag{5}$$

where $\oplus$ denotes feature concatenation.

The fused text feature $\mathbf{f}_y$ is then fed into a standard VAE encoder to produce the latent representation $\mathbf{z}_y$. The corresponding VAE objective is defined as:

$$\mathcal{L}_{\text{VAE}}^{y} = \mathbb{E}_{q_\phi(\mathbf{z}_y|\mathbf{f}_y)}\left[\log p_\theta(\mathbf{f}_y|\mathbf{z}_y)\right] - \beta_y D_{\text{KL}}\big(q_\phi(\mathbf{z}_y|\mathbf{f}_y) \| p(\mathbf{z}_y)\big). \tag{6}$$

This adaptive fusion mechanism complements our calibrated contrastive alignment by ensuring that the textual modality itself provides a reliable and semantically consistent target for cross-modal learning.

D. Calibrated Contrastive Alignment

a) Calibrated Contrastive Consistency: we propose a *calibrated contrastive alignment* strategy that leverages the learned gating mask to perform dimension-wise semantic calibration during cross-modal alignment.

Let $Z_s = \{\mathbf{z}_s^{(i)}\}_{i=1}^{N}$ and $Z_y = \{\mathbf{z}_y^{(i)}\}_{i=1}^{N}$ denote the sets of skeleton and semantic latent representations within a mini-batch of size N. For alignment, we apply the learned gating mask to calibrate latent dimensions:

$$\tilde{\mathbf{z}}_s = \mathbf{z}_s \odot \mathbf{g}_{mask}, \quad \tilde{\mathbf{z}}_y = \mathbf{z}_y \odot \mathbf{g}_{mask}^{\gamma}, \tag{7}$$

where $\tilde{\mathbf{z}}_s$ corresponds to the semantic-aware latent $\mathbf{z}_s^r$ defined in the previous subsection. The exponent $\gamma \in \{0, 1\}$ controls whether gradients from text alignment are propagated to the gating parameters. Since semantic relevance is primarily inferred from the skeleton modality, the gating mask is shared to calibrate semantic dimensions across modalities. Accordingly, the gating mechanism is primarily driven by skeleton-based alignment signals, while the text modality serves as a stable semantic reference.

The calibrated features are ℓ_2-normalized prior to similarity computation:

$$\hat{\mathbf{z}}_s = \frac{\tilde{\mathbf{z}}_s}{\|\tilde{\mathbf{z}}_s\|_2}, \quad \hat{\mathbf{z}}_y = \frac{\tilde{\mathbf{z}}_y}{\|\tilde{\mathbf{z}}_y\|_2}. \tag{8}$$

Rather than enforcing instance-level hard matching, we define a semantic-targeted similarity distribution $\mathbf{P}$ based on class relations and a temperature parameter τ, this distribution is derived from the class-level relational prior. Specifically, let $\mathbf{Y}$ be the set of textual class embeddings; the target distribution $\mathbf{P}$ for a mini-batch is formulated as:

$$P_{i,j} = \frac{\exp(\text{sim}(y_i, y_j)/\tau)}{\sum_{k=1}^{B} \exp(\text{sim}(y_i, y_k)/\tau)} \tag{9}$$

where $\text{sim}(\cdot)$ denotes the cosine similarity, B is the batch size, and τ is a temperature parameter that controls the smoothness of the distribution. By utilizing $\mathbf{P}$, we encode the latent relatedness between actions into the optimization objective, preventing the model from over-penalizing semantically similar samples. The calibrated contrastive objective aligns skeleton–skeleton, text–text, and cross-modal similarity distributions via KL divergence:

$$\begin{aligned} \mathcal{L}_c = \frac{1}{3}\Big(& D_{\text{KL}}(\text{sim}(Z_s, Z_s) \,\|\, \mathbf{P}) \\ & + D_{\text{KL}}(\text{sim}(Z_y, Z_y) \,\|\, \mathbf{P}) \\ & + D_{\text{KL}}(\text{sim}(Z_s, Z_y) \,\|\, \mathbf{P})\Big), \end{aligned} \tag{10}$$

Admittedly, the latent prior $\mathbf{P}$ derived from textual embeddings may not be inherently flawless; however, it provides a structured semantic consensus that is far more representative of the physical world than the simplistic one-hot encoding. By supervising the triple similarity distributions (SS, YY, and SY) with a unified $\mathbf{P}$, we construct a multi-view geometric constraint. This shared target forces the skeleton and text branches to converge towards a consistent relational topology, where the mutual reinforcement between intra-modal cohesion and cross-modal parity compensates for the potential bias in individual modalities. Consequently, this triplet alignment encourages the Soft Gate to ignore instance-specific artifacts and focus on the universal semantic features that are robust across both skeletal and textual domains. An illustrative diagram of the expected effect is provided in Fig. 3.

b) Cross-Modal Reconstruction Consistency: To prevent semantic collapse and preserve modality-specific information, a cross-modal reconstruction objective is employed to regularize the latent space. Specifically, the skeleton-related latent $\mathbf{z}_s^r$ is required to reconstruct the corresponding text feature

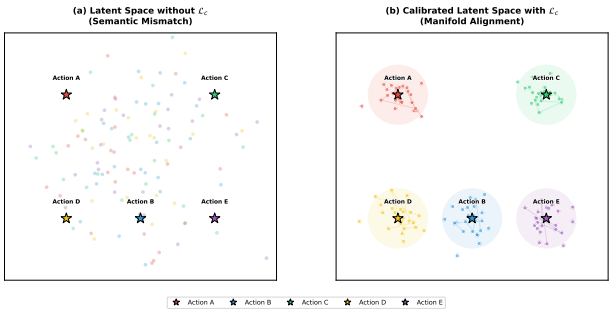


Fig. 3: Constrained loss $\mathcal{L}_c$ effectively guides the sample features to cluster around their corresponding semantic anchors, transforming a disordered latent space (a) into a well-aligned semantic manifold (b).

to enforce semantic consistency. Conversely, to ensure the alignment of the shared latent manifold, the text latent $\mathbf{z}_y$ is used to reconstruct the skeleton feature. Since the textual modality provides relatively clean semantic supervision, we do not apply gating to text latents, allowing full semantic information to guide skeleton reconstruction. Accordingly, the cross-modal reconstruction loss is formulated as:

$$\mathcal{L}_R = ||D_y(\mathbf{z}_s^r) - \mathbf{f}_y||_2^2 + ||D_x(\mathbf{z}_y) - \mathbf{f}_x||_2^2 \tag{11}$$

where D_y and D_x denote the text and skeleton decoders, respectively.

c) Overall Loss: The overall training objective consists of three components: the variational reconstruction losses for skeleton and text modalities, the proposed calibrated contrastive alignment loss, and the reconstruction-based regularization:

$$\mathcal{L} = \mathcal{L}_{\text{VAE}}^s + \mathcal{L}_{\text{VAE}}^y + \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_R, \tag{12}$$

where $\mathcal{L}_c$ plays a central role in cross-modal calibration, while $\mathcal{L}_R$ serves as an auxiliary constraint to stabilize training and preserve semantic fidelity.

IV. Experiments

A. Datasets

NTU-RGB+D 60 [16]: This dataset contains 56,880 skeleton sequences spanning 60 action categories, performed by 40 subjects and captured from 80 viewpoints using Microsoft Kinect v2 sensors. Each frame records the 3D spatial coordinates of 25 body joints, and the sequences can include up to two performers. When only a single person is present, the skeleton of the second performer is zero-padded. We adopt the two widely used experimental splits: Cross-Subject (XSub), where the training set includes half of the subjects while the remaining are used for testing; and Cross-View (XView), where training and testing samples are captured from different camera viewpoints.

NTU-RGB+D 120 [17]: As an extension of NTU RGB+D 60, this dataset includes 114,480 skeleton sequences across 120 action categories, performed by 106 subjects and captured from 155 camera viewpoints. The dataset provides two standard evaluation protocols: Cross-Subject (XSub), where sequences performed by 53 subjects are used for training and the rest for testing; and Cross-Setup (XSet), where sequences are split based on the camera setup—those with even IDs are used for training and odd IDs for testing.

B. Comparisons with State-of-the-Art Methods

We benchmark our approach against a range of recent zero-shot action recognition methods, with the full comparisons summarized in Tables I and II. To keep the evaluation on solid ground, we follow the same class-split protocol adopted by SynSE [18]. On the skeleton extractor side, we rely on the stronger Shift-GCN skeleton encoder [19], paired with the pre-trained ViT-B/32 text encoder from CLIP [20]. For context, FS-VAE [13] using GPT-4 to enrich class descriptions with more detailed action semantics. Rather than sidestepping that setup, we adopt their augmented descriptions directly, Evaluation follows standard practice. For ZSL, we report classification accuracy on unseen classes only. GZSL is a bit less forgiving, so we include accuracy on seen classes (Acc_s), unseen classes (Acc_u), along with their harmonic mean (H), which tends to reveal how well a model balances the two rather than excelling at just one.

Tables 1 and 2 respectively show the performance of the model on the ZSL and GZSL tasks. It can be seen that the model achieves excellent performance on the ZSL task, especially in complex split scenarios, such as the 48s/12u split under NTU60 and the split under NTU120. Furthermore, on the GZSL task, the model performs comparably to or surpasses the latest models on multiple datasets, demonstrating the good generalization ability of our model.

TABLE I: State-of-the-art comparisons on NTU RGB+D 60 and 120 datasets under the ZSL setting (Top-1 Accuracy %).

Method	Venue	NTU-60		NTU-120	
		55(s)/5(u)	48(s)/12(u)	110(s)/10(u)	96(s)/24(u)
ReViSE [21]	ICCV2017	53.91	17.49	55.04	32.38
JPoSE [22]	ICCV2019	64.82	28.75	51.93	32.44
CADA-VAE [23]	CVPR2019	76.84	28.96	59.53	35.77
SynSE [18]	ICIP2021	75.81	33.30	62.69	38.70
SA-DVAE [14]	ECCV2024	82.37	41.38	68.77	46.12
STAR [24]	ACMM2024	81.40	45.10	63.30	44.30
PURLS [25]	CVPR2024	79.23	40.99	71.95	52.01
FS-VAE [13]	ICCV2025	**86.90**	57.20	74.40	62.50
SCoPLe [9]	CVPR2025	84.10	52.96	74.53	52.17
GCA-VAE (Ours)	–	86.28	**57.60**	**76.27**	**64.68**

C. Ablation Studies

Influence of Components. From Table III, showing that the learnable gating (LG) module yields the most significant performance gain, indicating that skeleton representations indeed contain substantial noise and entangled variations. Unlike hard partitioning strategies adopted in prior works, LG performs a soft and adaptive modulation of latent dimensions, which effectively suppresses noisy factors while preserving mixed but discriminative features. The calibrated contrastive alignment

TABLE II: State-of-the-art comparisons on NTU RGB+D 60 and NTU RGB+D 120 datasets under the GZSL setting.

		NTU RGB+D 60						NTU RGB+D 120					
		55(s)/5(u)			48(s)/12(u)			110(s)/10(u)			96(s)/24(u)		
Method	Venue	Acc_s	Acc_u	H	Acc_s	Acc_u	H	Acc_s	Acc_u	H	Acc_s	Acc_u	H
ReViSE [21]	ICCV2017	74.22	34.73	29.22	62.36	20.77	31.16	48.69	44.84	46.68	49.66	25.06	33.31
JPoSE [22]	ICCV2019	64.44	50.29	56.49	60.49	20.62	30.75	47.66	46.40	47.05	38.62	22.79	28.67
CADA-VAE [23]	CVPR2019	69.38	61.79	65.37	51.32	27.03	35.41	47.16	19.78	48.44	41.11	34.14	37.31
SynSE [18]	ICIP2021	61.27	56.93	59.02	52.21	27.85	36.33	52.51	57.60	54.94	56.39	32.25	41.04
SA-DVAE [14]	ECCV2024	62.28	70.80	66.27	50.20	36.94	42.56	61.10	59.75	60.42	58.82	35.79	44.50
STAR [24]	ACMM2024	69.02	69.91	69.40	62.70	37.00	46.60	59.90	52.70	56.10	51.20	36.90	36.90
FS-VAE [13]	ICCV2025	**77.00**	74.50	**75.70**	56.20	48.60	52.10	59.20	**67.90**	63.30	57.80	**51.90**	54.70
SCoPLe [9]	CVPR2025	69.60	71.94	70.75	54.49	**61.83**	**57.93**	**63.51**	61.08	62.27	53.33	51.18	52.23
GCA-VAE (Ours)	–	66.86	**83.55**	74.27	**55.48**	50.76	53.02	60.48	67.40	**63.75**	**59.35**	51.27	**55.02**

(CA) module shows a moderate improvement when applied alone, suggesting that refining the alignment objective itself can partially alleviate misalignment caused by noisy or ambiguous skeleton features. More importantly, when combined with LG, CA brings a substantially larger gain. This indicates that CA benefits from cleaner semantic representations produced by LG, as the calibrated similarity distribution can be more accurately estimated when noisy dimensions are attenuated. These results demonstrate that LG and CA are complementary: LG improves feature quality at the representation level, while CA further refines cross-modal alignment at the structural level.

TABLE III: Ablation study of different components on NTU RGB+D 60 and 120 datasets.

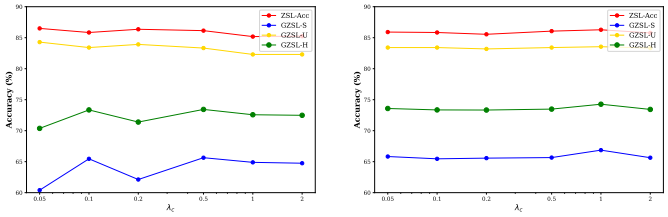
Modules		NTU-60		NTU-120	
LG	CA	55s/5u	48s/12u	110s/10u	96s/24u
✗	✗	84.12	51.83	69.46	61.26
✓	✗	85.62	53.00	74.68	63.36
✗	✓	84.63	51.96	70.37	61.42
✓	✓	**86.28**	**57.60**	**76.27**	**64.68**

Influence of Text Feature. We compare the original action labels (AL) with fine-grained descriptions generated by LLM. As shown in Table IV, enriched textual descriptions consistently improve performance, with larger gains observed under more challenging seen/unseen splits, highlighting the importance of detailed semantics for generalization. Moreover, the proposed Adaptive Text Feature Fusion further boosts performance by adaptively integrating multiple text sources, which helps mitigate noise and redundancy introduced by LLM-generated labels.

Influence of Loss Weighting Factors. We investigate the sensitivity of GCA-VAE to the loss weighting factors by varying the contrastive alignment weight λ_c and the gating regularization weight λ_g. Specifically, λ_c controls the strength of cross-modal contrastive calibration, while λ_g regulates the behavior of the learnable gating mechanism. For each experiment, one parameter is varied while the other is fixed to its default value. The results are illustrated in Fig. 4a and Fig. 4b.

TABLE IV: Influence of different text features on NTU RGB+D.

Text Modules	NTU-60		NTU-120	
(Feature Source)	55s/5u	48s/12u	110s/10u	96s/24u
AL (Action Labels)	84.96	38.84	69.71	49.07
LLM (Refined)	85.91	55.66	74.35	63.31
LLM + AF (Ours)	**86.28**	**57.60**	**76.27**	**64.68**



(a) Influence of λ_c (b) Influence of λ_g

Fig. 4: The influence of hyper-parameters on NTU-60.

V. Conclusion

This paper proposes GCA-VAE, a novel framework for zero-shot skeleton action recognition. We revisit skeleton–text alignment in zero-shot action recognition and show that hard semantic disentanglement is often suboptimal in the presence of noisy and entangled skeleton features. To address this issue, we introduce a soft gating strategy that suppresses unreliable skeleton components while preserving mixed semantic cues, and a calibrated contrastive alignment mechanism that stabilizes cross-modal matching and reduces the bias between seen and unseen classes. Extensive experiments demonstrate consistent performance improvements across benchmarks, especially when combined with adaptive text feature fusion.

References

[1] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7912–7921.

[2] Y. Cao, C. Liu, Z. Huang, Y. Sheng, and Y. Ju, "Skeleton-based action recognition with temporal action graph and temporal adaptive graph convolution structure," *Multimedia Tools and Applications*, vol. 80, no. 19, pp. 29 139–29 162, 2021.

[3] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[4] J. Li, Y. Zhou, H. Chu, H. Wang, Z. Zhao, F. Li, and Q. Li, "Dual multi-scale gcn with deformable temporal kernel for skeleton-based action recognition," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[5] Q. Liu, Y. Wu, B. Li, Y. Ma, H. Li, and Y. Yu, "Shotgcn: Spatial high-order temporal gcn for skeleton-based action recognition," *Neurocomputing*, vol. 632, p. 129697, 2025.

[6] W. Qian, Q. Huang, C. Li, Z. Chen, and Y. Mao, "Echogcn: An echo graph convolutional network for skeleton-based action recognition," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 245–261.

[7] H. Cui, R. Huang, R. Zhang, and T. Hayama, "Dstsa-gcn: Advancing skeleton-based gesture recognition with semantic-aware spatio-temporal topology modeling," *Neurocomputing*, vol. 637, p. 130066, 2025.

[8] Y. Chen, J. Guo, S. Guo, and D. Tao, "Neuron: Learning context-aware evolving representations for zero-shot skeleton action recognition," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8721–8730.

[9] A. Zhu, J. Zhu, J. Bailey, M. Gong, and Q. Ke, "Semantic-guided cross-modal prompt learning for skeleton-based zero-shot action recognition," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 13 876–13 885.

[10] M.-Z. Li, Z. Jia, Z. Zhang, Z. Ma, and L. Wang, "Multi-semantic fusion model for generalized zero-shot skeleton-based action recognition," in *International Conference on Image and Graphics*. Springer, 2023, pp. 68–80.

[11] H. Xu, Y. Gao, J. Li, and X. Gao, "An information compensation framework for zero-shot skeleton-based action recognition," *IEEE Transactions on Multimedia*, 2025.

[12] Y. Zhou, W. Qiang, A. Rao, N. Lin, B. Su, and J. Wang, "Zero-shot skeleton-based action recognition via mutual information estimation and maximization," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 5302–5310.

[13] W. Wu, Z. Guo, C. Chen, H. Xue, and A. Lu, "Frequency-semantic enhanced variational autoencoder for zero-shot skeleton-based action recognition," *arXiv preprint arXiv:2506.22179*, 2025.

[14] S.-W. Li, Z.-X. Wei, W.-J. Chen, Y.-H. Yu, C.-Y. Yang, and J. Y.-j. Hsu, "Sa-dvae: Improving zero-shot skeleton-based action recognition by disentangled variational autoencoders," in *European Conference on Computer Vision*. Springer, 2025, pp. 447–462.

[15] Z. Chen, Y. Luo, R. Qiu, S. Wang, Z. Huang, J. Li, and Z. Zhang, "Semantics disentangling for generalized zero-shot learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8712–8720.

[16] A. Shahroudy, J. Liu, T.-T. Ng, and G. 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*, 2016, pp. 1010–1019.

[17] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[18] P. Gupta, D. Sharma, and R. K. Sarvadevabhatla, "Syntactically guided generative embeddings for zero-shot skeleton action recognition," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 439–443.

[19] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 183–192.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[21] Y.-H. Hubert Tsai, L.-K. Huang, and R. Salakhutdinov, "Learning robust visual-semantic embeddings," in *Proceedings of the IEEE International conference on Computer Vision*, 2017, pp. 3571–3580.

[22] M. Wray, D. Larlus, G. Csurka, and D. Damen, "Fine-grained action retrieval through multiple parts-of-speech embeddings," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 450–459.

[23] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero-and few-shot learning via aligned variational autoencoders," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8247–8255.

[24] Y. Chen, J. Guo, T. He, X. Lu, and L. Wang, "Fine-grained side information guided dual-prompts for zero-shot skeleton action recognition," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 778–786.

[25] A. Zhu, Q. Ke, M. Gong, and J. 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*, 2024, pp. 18 761–18 770.