SIGMA: SKELETON INJECTION WITH GENERATIVE MOTION AUGMENTATION FOR ZERO-SHOT ACTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot skeleton-based action recognition asks a model to classify actions it has never seen, from a textual description alone. The difficulty is the distance between the two modalities: a sentence and a sequence of joint coordinates share no common ground, and the sentence is the only evidence available for the categories to be predicted. Text-to-motion generation changes that. From the same description, these models now synthesize plausible movement, a second source of evidence, and project it into the modality the recognizer consumes. The raw generated skeletons cannot serve directly as supervision: the joint layout differs, body proportions and timing drift, two-person interactions cannot be produced at all, and quality varies sharply from one sample to the next. SIGMA turns that output into training material, retargeting and rescaling it onto the recorded format, composing two-body interactions from the action verb, and obtaining a well-structured, balanced, and diverse training subset. To narrow the gap between text and skeleton, we also propose refining the text by generating geometry-aware and more discriminative sentences to prevent confusion between similar action categories. To combine generated and real skeletons, we have designed a novel training objective that treats both types of samples as two populations rather than one, as a global shift remains. At inference time, we compensate for this shift in the generated latents, as the relative order between unseen categories is preserved. On NTU RGB+D 60 and NTU RGB+D 120, SIGMA sets a new state-of-the-art in zero-shot recognition on three of four splits, by up to 22 points, and reports the best generalized harmonic mean on the other two. Generated skeletons alone account for up to 7.4 points of harmonic mean over an identical system using only textual descriptions.

1 INTRODUCTION

Action recognition systems are trained on a fixed set of categories, yet deployment rarely respects those boundaries. For instance, a monitoring system in a care facility will encounter behaviours nobody thought to record; a robot sharing a workspace will see gestures absent from its training set. Annotating skeleton sequences for every category one might eventually care about is not a realistic answer. Zero-shot recognition takes the other route: it asks the model to recognize a motion it has never seen, from a textual description of that category alone. Skeletons suit this task well: they discard appearance, making them robust to lighting and background variation, and they are far less invasive than video, while remaining cheap enough to process at scale (Shahroudy et al., 2016; Liu et al., 2020; Cheng et al., 2020).

The difficulty is the distance between the two modalities. Methods built for this setting train a skeleton encoder and a text encoder to place their outputs in a common latent space, so that a sequence lands near the description of its category (Gupta et al., 2021; Li et al., 2024; Zhu et al., 2024; Wu et al., 2025). That alignment is learned from paired data, which exists only for the categories that were recorded. For every other category the model is handed a sentence and nothing else, and whatever it comes to know about how those motions look is inferred from the geometry of the text space and from whichever recorded categories happen to be described in similar terms. Two consequences follow. Nothing ties a description to the movement it denotes, so categories whose descriptions read alike are placed close together even when their motions are not. And those descriptions are usually written one category at a time, without reference to the categories an action will be confused with, which is precisely what a classifier needs.

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.

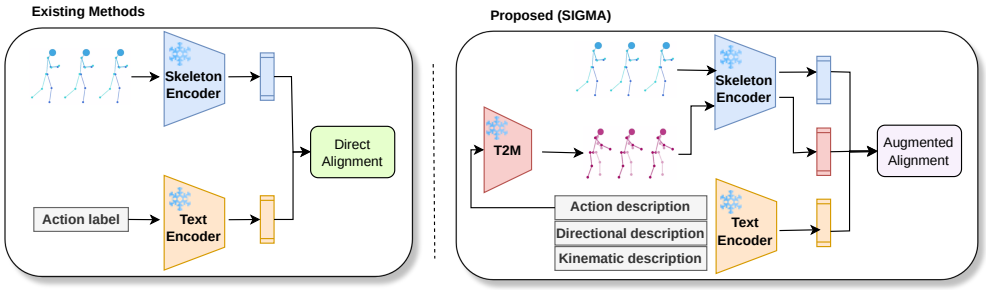

Figure 1: Where the evidence comes from. *(Left)* Existing methods encode the category name into a text feature and align it with the skeleton features of recorded sequences; for a category that was never recorded, that text feature is the only evidence the alignment ever receives. *(Right)* SIGMA writes three complementary descriptions per category and sends them both to the text encoder and to a text-to-motion generator. The generated sequences are brought into the recorded skeleton format and pass through the same frozen extractor as recorded ones, so the alignment is trained on skeleton evidence for categories that have no recording.

Text-to-motion generation changes what is available here. From a sentence, models such as MotionGPT3 (Zhu et al., 2025b) produce a sequence of body poses that is, in its coarse structure, the motion the sentence describes. For a category with no recording, this is a second source of evidence obtained from the only input the protocol allows, and it is of a different kind than text: it lives in the modality the recognizer consumes. Generative methods for this task already synthesize features for unseen categories directly in the latent space; we instead generate motion, which then passes through the same extractor as a recorded sequence and is therefore subject to the same constraints as real data rather than to those of the latent space.

That evidence cannot be used as it comes, and making it usable is most of the work. The joint layout does not match the one the extractor expects, limb proportions drift away from those of the recorded subjects, sequences unfold roughly twice as slowly as real ones, and the categories that involve two people cannot be produced at all, since the generator writes a single body per sentence. Quality also varies sharply between samples drawn from the same prompt, so feeding an unfiltered pool into training does more harm than good: it costs three points of harmonic mean relative to a curated subset of the same material.

Once that evidence is in place, the text side becomes worth investing in. When the only bridge between a category name and a movement is the text space itself, a longer description mostly adds words; when skeleton evidence is present for the same category, a description that states what separates it from the categories it is confused with has something to attach to. We therefore have a language model write each category against its confusable set, along three complementary axes, and encode the result with a general-purpose text embedding model (Zhang et al., 2025) rather than with the image-text encoder this literature has settled on.

Our contributions are the following.

- **A novel ZS framework** that incorporates generated samples to bridge the modality gap by enhancing cross-modal alignment.
- **A context-aware physical class description module** that models actions at a finer level of detail, increases inter-class diversity, and provides more explainable semantic representations for the skeleton sequence.
- **Comprehensive experiments on two datasets under different data-split protocols**, demonstrating that SIGMA achieves competitive performance in GZSL and outperforms state-of-the-art methods in ZSL.

2 Related Work

2.1 Zero-Shot Skeleton-Based Action Recognition

Embedding-based. These methods project skeleton and text into a shared space and recognise by similarity. Early work matched global skeleton features to label embeddings, and later methods refined the granularity of that match: aligning body parts with the clauses that describe them (Zhu et al., 2024), or learning prompts on the skeleton side rather than only on the text side (**?**).**Generative-based.** A second line synthesises unseen-class representations from semantics and trains a classifier on them. Most of these build on paired variational autoencoders, one per modality, tied by reconstruction and alignment terms (Schonfeld et al., 2019; Gupta et al., 2021); subsequent work separates the part of the skeleton representation that text can explain from the part it cannot (Li et al., 2024), or splits the skeleton signal by frequency band so that small, fast gestures are not averaged away by slow ones (Wu et al., 2025). A parallel line replaces similarity matching altogether with a learned transport between the two latent distributions (Do & Kim, 2025; Chen et al., 2026). **Summary.** Whatever the architecture, all of these methods face the same constraint: for the categories being predicted, text is the only evidence available at training time. SIGMA keeps the shared-latent backbone this literature converged on and changes what is fed into it.

2.2 Semantic Descriptions for Unseen Categories

Since text is the only side the model can consult for an unseen category, how that text is written matters as much as how it is encoded. Label names alone are short and ambiguous, so recent methods replace them with richer material: descriptions written per body part (Zhu et al., 2024), prompts that a language model completes with the objects and settings an action involves (**?**), or class semantics adjusted using the categories that lie nearest in the text space (Chen et al., 2026). Our descriptions belong to this family, with one difference in how they are produced: each is written while the language model is shown the categories the action is actually confused with, so that the text states the distinguishing motion rather than describing the action in isolation.

2.3 Motion Generation as a Source of Training Data

Text-to-motion generation. Motion synthesis has moved from short, constrained clips to models that take a free-form sentence and return a plausible sequence of poses, trained on motion-caption corpora (Guo et al., 2022); recent systems treat motion as a modality alongside language inside a single model, which improves how faithfully a generated sequence follows the wording it came from (Zhu et al., 2025b). These models are evaluated on how natural and how text-faithful their output looks to a human or to a learned critic, a different standard from the one that matters here. **Synthetic data for recognition.** Training a recognizer on generated samples is established in vision, with the recurring finding that raw generated material is rarely usable as is and that selection matters more than volume. Skeletons add constraints images do not have: a fixed joint layout, plausible body proportions, the right speed, and, for interactions, two bodies in a coherent spatial relation, none of which a model trained to produce natural-looking motion guarantees. **Summary.** We use such a generator as a component and measure it against the standard that matters for recognition, namely whether its output is close enough to recorded data to be learned from; Section 3 addresses each of the constraints above in turn.

3 Method

SIGMA keeps the shared-latent backbone this literature converged on and changes what the model is given for the categories it must predict. Three components carry that change: how the skeleton side is built for categories that were never recorded (Section 3.1), how the semantic side is written and encoded (Section 3.2), and how both are aligned during training (Section 3.3). Section 3.4 describes how the trained model commits to a label.

Let $\mathcal{S}$ and $\mathcal{U}$ denote the seen and unseen category sets, with $\mathcal{S} \cap \mathcal{U} = \emptyset$. A skeleton sequence is summarised by a frozen skeleton feature extractor into $s \in \mathbb{R}^{d_s}$, written s_o when the sequence is an original recording and s_a when it is artificial, that is, produced by the generator of Section 3.1. A category c is summarised by a frozen text feature extractor into $t_f \in \mathbb{R}^{d_t}$, written t_f^c where the category must be made explicit. Two variational encoders map these into a shared latent space of

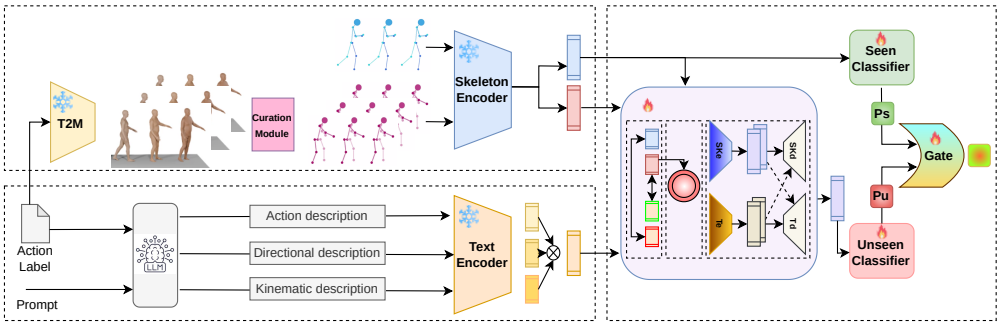

Figure 2: Overview of the SIGMA framework.

dimension d: $E_s(s)$ returns (μ_s, σ_s) and $E_t(t_f)$ returns (μ_t, σ_t), with decoders D_s, D_t and latents z_s, z_t.

3.1 Skeleton Feature Extraction for Unseen Categories

For a seen category, s comes from a recording. For an unseen one, no recording exists, and this section builds a substitute.

Generation. Each of the three descriptions of Section 3.2 is submitted to MotionGPT3 (Zhu et al., 2025b) under several random seeds, which yields a pool of candidate sequences per category rather than a single one. The generator returns poses in its own representation: a different joint set, its own body scale, and its own frame rate.

Retargeting. Three corrections bring that output into the recorded format. The sequence is resampled to the frame rate of the recorded data; its joints are mapped onto the recorded layout, the surplus joints of the target skeleton being filled by replication from their parent since the generator does not model them; and each bone is rescaled to the median length measured on the training sequences of seen categories, applying the correction along the kinematic chain from the pelvis outwards, which preserves joint angles while changing segment lengths.

Composing two-body interactions. A sentence describes one body and the generator returns one body, so categories such as two people walking towards each other cannot be produced directly. We assemble them from two independently generated sequences placed on a common timeline. What the generator does not provide is the spatial relation between the actors, and both parts of it are derived from the action verb rather than from any recording: for an approach the bodies are rotated to face one another, for a separation back to back, and for a static interaction they keep their generated orientation. The initial gap is set from the typical interpersonal distance of two-body categories among the seen classes, and extended by the distance each body covers when the verb implies locomotion, so that the pair ends at that distance. The composed pair is rotated as a whole by one of three yaw angles, reproducing the recorded viewpoints. Because the rule comes from the verb, it can be validated on splits where these categories are seen and applied unchanged where they are not (Section 4).

Encoding. Recorded and generated sequences are then passed through the same frozen Shift-GCN encoder (Cheng et al., 2020), pre-trained on the seen categories of the split, which yields the d_s-dimensional feature s. Generated material is therefore subject to the same representation as real data, not to a separate branch.

Curation. Generated samples of a category vary widely in quality, and injecting the full pool degrades training. Selecting the best ones is straightforward when real sequences of that category are available, and impossible to justify here, since those sequences are exactly what the model must not have seen. Our criterion uses only the generated material. Let $s_a^1, \dots, s_a^n$ be the encoded candidates of an unseen category c and $\bar{s}_a$ their centroid. We discard the candidates furthest from $\bar{s}_a$, keeping the closest $\lceil \rho n \rceil$, which removes degenerate generations without appealing to any external reference. Among those, a fixed budget of K samples is chosen by farthest-point selection,

$$s_a^{(1)} = \arg\min_i \|s_a^i - \bar{s}_a\|, \qquad s_a^{(k+1)} = \arg\max_i \min_{j \le k} \|s_a^i - s_a^{(j)}\|, \tag{1}$$

so the retained set covers the spread of the category instead of collapsing onto its mode. K is the same for every category, which prevents a large pool from dominating training.

3.2 Semantic Feature Extraction

A description is the only evidence the model receives about an unseen category, and the failure mode it must avoid is not vagueness but non-discrimination: two categories whose descriptions are individually accurate can still be phrased so similarly that their embeddings nearly coincide.

Writing. For every category we prompt Claude Sonnet three times, along complementary axes: the action label (LB), a directional description (DIR) stating where the limbs travel, and a kinematic description (KD) stating which body parts move, in what order, and against what reference. In the DIR and KD prompts the model is given the categories the action is most often confused with and is asked to state what separates them, so the text carries the distinguishing motion rather than a description of the action in isolation.

Encoding. The three texts are embedded independently into t_{lb}, t_{dir} and t_{kd} (Zhang et al., 2025), each ℓ_2-normalised, and concatenated into $t_f = [\, t_{\text{lb}} \| t_{\text{dir}} \| t_{\text{kd}} \,]$, so that t_f carries three views of the same category rather than one.

3.3 Zero-Shot Training Objective

Both modalities are now vectors, and training aligns them. The skeleton encoder E_s applies the frequency-enhanced reweighting of Wu et al. (2025) to s before projecting it into the latent space; that module is used unchanged and we claim no contribution there.

Reconstruction and regularisation. Each modality is reconstructed from its own latent and from the other's, and both posteriors are regularised towards a standard normal,

$$\begin{aligned}\mathcal{L}_{\text{VAE}} = \sum_{m\in\{s,t\}} \Big(\|v_m - D_m(z_m)\|^2 + \beta\, \mathrm{KL}\big(q(z_m \mid v_m) \,\|\, \mathcal{N}(0,I)\big)\Big) \\ + \gamma\Big(\|s - D_s(z_t)\|^2 + \|t_f - D_t(z_s)\|^2 \Big),\end{aligned} \tag{2}$$

where $v_s = s$ and $v_t = t_f$.

Cross-modal alignment. Alignment uses the calibrated loss of Wu et al. (2025), which pushes a matched pair together only in proportion to how much closer it already is than a mismatched one, and is therefore robust to the ambiguous skeleton-text pairs this task produces,

$$\begin{aligned}\mathcal{L}_{\text{align}} = \frac{\lambda}{|B|} \sum_{i\in B} \frac{1}{1+\exp\big((\|t_f^i - D_t(z_s^{i^-})\|^2 - \|t_f^i - D_t(z_s^i)\|^2)/\lambda\big)} \\ + \frac{\lambda}{|B|} \sum_{i\in B} \frac{1}{1+\exp\big((\|s^i - D_s(z_t^{i^-})\|^2 - \|s^i - D_s(z_t^i)\|^2)/\lambda\big)},\end{aligned} \tag{3}$$

where i^- is a negative drawn by permutation inside the batch B and λ is a temperature.

Handling two populations in one batch. The curated generated samples enter training as additional elements of the batch, paired with the description of their category, so they are subject to Eq. 2 and Eq. 3 exactly as recorded samples are: a generated skeleton is reconstructed from its own description and aligned with it. How many of them a batch receives is set by a target exposure rather than by the size of the pool, so a category with many generated candidates does not dominate training. Two further adjustments follow from their being a different population. First, the text branch of Eq. 3 draws its negatives by permutation inside the batch; with recorded and generated samples mixed, a recorded sample could draw a generated one as its negative and be pushed away from it for reasons unrelated to its category. We therefore permute the two populations separately in that branch, so a recorded sample only ever draws a recorded negative. Second, a contrastive term acts on the generated elements B_a alone, pulling each generated latent towards the semantic embedding of its own category and away from those of all other categories, seen and unseen,

$$\mathcal{L}_{\text{ctr}} = -\frac{1}{|B_a|} \sum_{i\in B_a} \log \frac{\exp\big(\langle \tilde{z}_s^i,\ \tilde{t}_f^{c_i}\rangle/\tau\big)}{\sum_{c\in\mathcal{S}\cup\mathcal{U}} \exp\big(\langle \tilde{z}_s^i,\ \tilde{t}_f^{c}\rangle/\tau\big)}, \tag{4}$$

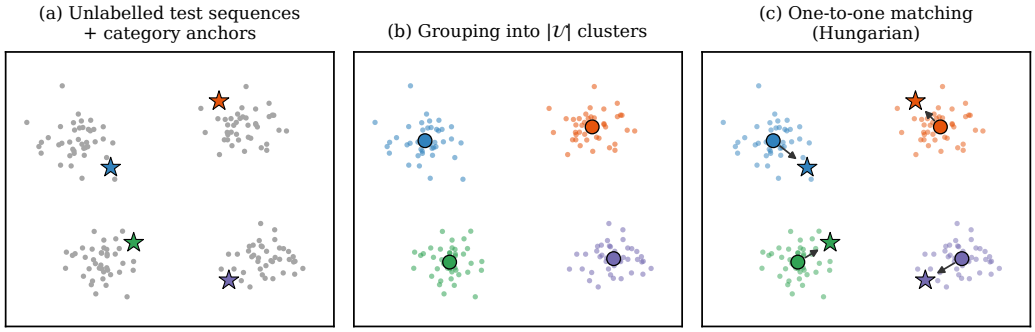

Figure 3: Set-level assignment. *(a)* The unlabelled evaluation sequences are mapped to latents, alongside one anchor per unseen category. *(b)* They are grouped into $|\mathcal{U}|$ clusters, whose centres $\bar{z}_k$ summarise where the evaluation set actually concentrates. *(c)* Clusters and anchors are matched one-to-one by minimising the total distance of Eq. 7. A cluster may therefore receive an anchor that is not its nearest, which is what prevents two clusters from claiming the same category and leaving another unpredicted.

where $\tilde{\cdot}$ denotes ℓ_2 normalisation and the semantic embeddings are held fixed. Its negatives are category-level rather than sample-level, which is what a generated sample can be held responsible for: it should land in its category's region, not resemble any particular recording. The term is switched on only once the alignment has stabilised, since applying it from the first cycle sharpens boundaries around latents that are not yet placed correctly. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \alpha\,\mathcal{L}_{\text{align}} + \delta\,\mathcal{L}_{\text{ctr}}. \tag{5}$$

3.4 Inference

Per-sequence decision. A sequence is assigned to the unseen category whose semantic posterior explains its latent best,

$$\hat{y}(s) \;=\; \arg\min_{c\in\mathcal{U}} \sum_{j=1}^{d} \frac{(z_{s,j}-\mu_{t,c,j})^2}{\sigma^2_{t,c,j}}, \qquad z_s = \mu_s. \tag{6}$$

Set-level assignment. Injection does more than supply training material: the generated latents place and separate the regions the unseen categories occupy, so those categories are arranged in the latent space rather than scattered. Deciding each sequence in isolation discards that arrangement. We instead group the unlabelled evaluation sequences into as many clusters as there are unseen categories and match clusters to categories by minimising the total distance between cluster centres $\bar{z}_k$ and category anchors a_c,

$$\pi^\star \;=\; \arg\min_{\pi\in\Pi} \sum_{k} \|\bar{z}_k - a_{\pi(k)}\|, \tag{7}$$

where $\bar{z}_k$ is the centre of cluster k, $a_c = \mu_{t,c}$ is the anchor of category c, and Π is the set of one-to-one matchings, solved exactly by the Hungarian algorithm.

Among several candidate clusterings we retain the one whose matching cost is lowest, a criterion that requires no labels. The anchor a_c is the semantic posterior mean of category c; validation retains it over its combination with the centroid of the retained generated samples on all four splits. The one-to-one constraint is what the per-sequence decision lacks: a cluster lying between two anchors cannot be absorbed by both, so a category can no longer go unpredicted while a neighbour takes its sequences.

Generalized setting. When seen categories are also present at test time, sequences are first routed to one side by the gating mechanism used throughout this literature, and the per-sequence decision of Eq. 6 is applied on the unseen side. Set-level assignment is not used here: the sequences routed to the unseen side include misrouted seen ones, so the assumption the matching rests on, that the set partitions exactly into the unseen categories, no longer holds. All generalized results we report are therefore obtained with per-sequence decisions, and the gain from injection they show is independent of the set-level component.

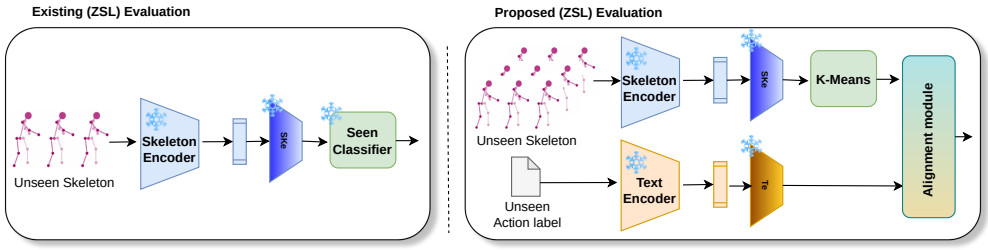

Figure 4: ZSL evaluation.

4 Experiments

4.1 Datasets and Splits

NTU RGB+D 60 (Shahroudy et al., 2016) contains 56,880 skeleton sequences over 60 action categories, performed by 40 subjects and captured by Kinect sensors with 25 joints per body. **NTU RGB+D 120** (Liu et al., 2020) extends it to 120 categories and 114,480 sequences from 106 subjects. Both are used under the cross-subject protocol. We follow the seen/unseen splits of SynSE (Gupta et al., 2021), in continuity with Wu et al. (2025): 55/5 and 48/12 on NTU-60, 110/10 and 96/24 on NTU-120, where the second number is the count of categories held out from training. Each split additionally provides a validation partition, built by holding out categories from the seen set, on which every design choice is made.

Table 1: Zero-shot learning results. Best in red, second in blue. † marks the transductive regime, in which the label is decided over the unlabelled evaluation set as a whole (Eq. 7) rather than independently per sequence.

Methods	**NTU-60 (ACC,%)**		**NTU-120 (ACC,%)**	
	55/5 split	48/12 split	110/10 split	96/24 split
CADA-VAE Schonfeld et al. (2019) (CVPR 19)	76.8	29.0	59.5	35.8
SynSE Gupta et al. (2021) (ICIP 21)	75.8	33.3	62.7	38.7
SMIE Zhou et al. (2023) (ACMMM 23)	78.0	40.2	65.7	45.3
STAR Chen et al. (2024) (ACMMM 24)	81.4	45.1	63.3	44.3
GZSSAR Li et al. (2023) (ICIG 23)	83.3	49.8	72.0	60.7
PURLS Zhu et al. (2024) (CVPR 24)	79.2	41.0	72.0	52.0
SA-DVAE Li et al. (2024) (ECCV 24)	82.4	41.4	68.8	46.1
FS-VAE Wu et al. (2025) (ICCV 25)	**86.9**	57.2	74.4	62.5
SCoPLe Zhu et al. (2025a) (CVPR 25)	84.1	53.0	74.5	52.2
TDSM Do & Kim (2025) (ICCV 25)	86.5	56.0	74.2	65.1
PGFA Zhou et al. (2025) (TIP 25)	80.3	56.0	80.0	59.4
SIGMA (Ours)	84.2	67.3	80.0	64.6
SIGMA† (Ours)	84.0	**78.9**	**91.7**	**81.3**

4.2 Comparison with the State of the Art

Zero-shot recognition (ZSL). Table 1 shows that SIGMA's gains depend on the nature of the unseen categories. The largest improvement occurs on 48/12, where unseen classes include two-body interactions and temporally mirrored actions that text alone describes ambiguously; generated motion evidence helps resolve these distinctions. In contrast, gains are smaller on 55/5 and 96/24, where several unseen classes depend on hand-object interactions or fine-grained gestures that a 25-joint skeleton cannot represent. These splits are therefore closer to saturation, limiting the benefit of motion generation.

Generalized zero-shot recognition (GZSL). Table 2 shows the effect of motion injection. SIGMA achieves the highest unseen accuracy on 55/5 and 110/10 and the second highest on 48/12, indicating that motion evidence better populates unseen regions of the latent space and reduces their absorption into seen classes. This comes with a small seen-accuracy trade-off on 55/5, but overall yields the best harmonic mean on 48/12 and 110/10 and ties on 96/24, with only a 0.1-point loss on 55/5. On 96/24,

the larger unseen set leads to a stronger seen-side performance and limits the benefit of favouring unseen classes.

Table 2: Generalized zero-shot results. H is the harmonic mean. Best per column in red, second in blue; ties share the first rank. GZSSAR results are those reproduced by Wu et al. (2025) from the released code.

Methods	NTU-60 (55/5 split)			NTU-60 (48/12 split)			NTU-120 (110/10 split)			NTU-120 (96/24 split)		
	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H
CADA-VAE Schonfeld et al. (2019) (CVPR 19)	69.4	61.8	65.4	51.3	27.0	35.4	47.2	49.8	48.4	41.1	34.1	37.3
SynSE Gupta et al. (2021) (ICIP 21)	61.3	56.9	59.0	52.2	27.9	36.3	52.5	57.6	54.9	56.4	32.2	41.0
STAR Chen et al. (2024) (ACMMM 24)	69.0	69.9	69.4	**62.7**	37.0	46.6	59.9	52.7	56.1	51.2	36.9	42.9
GZSSAR Li et al. (2023) (ICIG 23)	66.8	70.7	68.7	54.8	41.4	47.1	58.1	57.8	58.0	59.2	45.9	51.7
SA-DVAE Li et al. (2024) (ECCV 24)	62.3	70.8	66.3	50.2	36.9	42.6	61.1	59.8	60.4	58.8	35.8	44.5
FS-VAE Wu et al. (2025) (ICCV 25)	**77.0**	74.5	**75.7**	56.2	48.6	52.1	59.2	67.9	63.3	57.8	**51.9**	**54.7**
SCoPLe Zhu et al. (2025a) (CVPR 25)	69.6	71.9	70.8	54.5	**61.8**	57.9	**63.5**	61.1	62.3	53.3	51.8	52.2
SIGMA (Ours)	71.6	**80.0**	75.6	57.6	60.3	**58.9**	62.3	**68.6**	**65.3**	**60.6**	49.9	**54.7**

5 Ablation Study

Every experiment below is run on all four splits with the rest of the pipeline held fixed, and reports zero-shot accuracy per sequence.

5.1 Contribution of Each Component

Table 3 turns the three components on in turn: the context-aware descriptions of Section 3.2 (SD), the injection of curated generated skeletons of Section 3.1 (INJ), and the permutation and contrastive adjustments that the mixed population requires, from Section 3.3 (CL). CL is only defined when generated samples are present and is never enabled on its own.

Read as a whole, the table says something more specific than a list of additive gains. Injection driven by the bare category label, in the third row, is worth 0.5 points on average and is *below* the baseline on both NTU-60 splits: generated skeletons dropped into a latent space organised by category names alone land outside the region that names them about as often as inside it. The descriptions are what make them usable. Once the descriptions are in place, adding the same generated skeletons is worth 2.5 points on average, and the two together reach 72.8 against 65.0 for neither, more than the sum of what each returns alone. The contribution is joint rather than separable, which is what one should expect from a mechanism whose purpose is to tie a description to movement.

The training adjustments add a further 1.3 points on average and are the least uniform of the three: they are worth 1.5, 2.8 and 3.5 points on 55/5, 110/10 and 96/24, and cost 2.7 on 48/12, where the configuration without them is the better one. We report and release the configuration that is best on average and apply it unchanged to all four splits, rather than selecting the training objective per split.

5.2 Which Description Axes Carry the Signal

Table 4 varies the text handed to the text encoder and leaves everything else fixed, injected skeletons included. No axis is sufficient on its own: the three single-axis variants sit within 2.5 points of one another on average and all three are at least 8.5 points below the full set. Nor is any pair a substitute for the third, which costs 2.4 points on average even in its best form. The two enriched axes are also not interchangeable, and which one matters depends on the split: the directional axis alone collapses on 110/10, eleven points below the bare label, while on 48/12 and 96/24 the pair that includes the kinematic axis is respectively six and four points ahead of the pair that includes the directional one.

The last two rows isolate what the cross-category context contributes, and the answer is that it contributes nearly all of it. The *independent* variant uses the same three axes, written by the same language model under the same instructions, except that each category is described on its own rather than against the categories it is confused with. It averages 65.9, against 65.5 for the bare category label: three sentences of kinematic and directional detail per category are worth four tenths of a point over the label alone when they are written in isolation, and 8.6 points when they are written against the confusable set. The loss relative to the full descriptions is 8.2 points on average and reaches 12.0 on 110/10. What the language model contributes is therefore not description in the ordinary sense, but the contrast a classifier needs and a category name does not carry.

Table 3: Influence of each component. Context-aware semantic descriptions (SD), injection of curated generated skeletons (INJ), and the permutation and contrastive adjustments the mixed population requires (CL). Zero-shot accuracy per sequence (%). CL requires INJ and is therefore never enabled on its own. The last row is the configuration we adopt and apply unchanged to all four splits.

Components			**NTU-60 (ACC,%)**		**NTU-120 (ACC,%)**		**Mean**
SD	INJ	CL	55/5 split	48/12 split	110/10 split	96/24 split	
✗	✗	✗	82.4	54.5	73.7	49.4	65.0
✓	✗	✗	81.3	59.3	78.9	61.7	70.3
✗	✓	✓	81.6	53.7	75.1	51.6	65.5
✓	✓	✗	82.7	70.0	77.3	61.1	72.8
✓	✓	✓	**84.2**	**67.3**	**80.1**	**64.6**	**74.1**

Table 4: Which description axes carry the signal. Zero-shot accuracy per sequence (%), same backbone, same injected skeletons, same hyperparameters; only the text given to the text encoder changes. *Independent* uses the same three axes written by the same model under the same instructions, each category described on its own rather than against the categories it is confused with.

Description axes	55/5	48/12	110/10	96/24	Mean
Label only	81.6	53.7	75.1	51.6	65.5
Directional only	82.4	55.0	64.0	53.8	63.8
Kinematic only	81.0	51.9	68.1	51.0	63.0
Label + directional	82.7	60.7	76.7	55.8	69.0
Label + kinematic	81.2	66.9	78.0	60.3	71.6
Directional + kinematic	82.6	63.5	71.7	59.8	69.4
All three, independent	80.1	60.5	68.1	55.1	65.9
All three, context-aware (SIGMA)	**84.2**	**67.3**	**80.1**	**64.6**	**74.1**

Table 5: Which generated samples to keep. Harmonic mean (%) in the generalized setting on the 55/5 split of NTU-60, at a fixed budget of fifteen samples per category; only the rule that picks them changes.

Selection rule	**NTU-60 55/5**
Full pool, no selection	72.6
Closest to the synthetic centroid	75.1
Random subset	74.6
Proximity then diversity (ours)	**75.6**

5.3 Which Generated Samples to Keep, and How Many

The selection rule of Eq. 1 matters more than the amount of generated material (Table 5). Unfiltered material is the worst of the four rules, three points below the rule we adopt and below even a random subset of the same size: what hurts is not the absence of generated samples but the presence of the ones the generator got wrong. Raising the budget does not help either — on 110/10, moving from fifteen to twenty-four samples per category lowers the harmonic mean from 65.3 to 63.1 — since the samples admitted beyond the fifteenth are by construction the ones the selection rule ranked last.

6 Conclusion

We investigated whether text-to-motion generation can provide supervision for action categories that were never recorded. Our approach combines generated motion with recorded skeletons through format alignment, compositional synthesis of unsupported interactions, and category-agnostic selection of diverse samples. This provides up to a 7% improvement in harmonic mean over an otherwise identical description-only system and substantially improves zero-shot recognition over prior methods across all splits and protocols.

7 AI USE STATEMENT

In this work, we used generative AI tools for language editing, proofreading, and improving the clarity and presentation of the manuscript. We also used generative AI tools to assist with brainstorming and refining the wording of some descriptions and explanations.

We did not use generative AI tools to generate scientific results, conduct experiments, make experimental or methodological decisions, generate or modify datasets, or produce the final figures and tables. These uses are therefore not applicable to this work.

REFERENCES

Yang Chen, Jingcai Guo, Tian He, Xiaocheng Lu, and Ling Wang. Fine-grained side information guided dual-prompts for zero-shot skeleton action recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 778–786, 2024.

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

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

Jeonghyeok Do and Munchurl Kim. Bridging the skeleton-text modality gap: Diffusion-powered modality alignment for zero-shot skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12757–12768, October 2025.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5152–5161, June 2022.

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

Ming-Zhe Li, Zhen Jia, Zhang Zhang, Zhanyu Ma, and Liang Wang. Multi-semantic fusion model for generalized zero-shot skeleton-based action recognition. In *International Conference on Image and Graphics*, pp. 68–80. Springer, 2023.

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 (ECCV)*, 2024.

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 TPAMI*, 2020.

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 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. *CVPR*, 2016.

Wenhan Wu, Zhishuai Guo, Chen Chen, Hongfei Xue, and Aidong Lu. Frequency-semantic enhanced variational autoencoder for zero-shot skeleton-based action recognition. In *ICCV*, 2025.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

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

Yujie Zhou, Wenwen Qiang, Anyi Rao, Ning Lin, Bing Su, and Jiaqi Wang. Zero-shot skeleton-based action recognition via mutual information estimation and maximization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5302–5310, 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 Computer Vision and Pattern Recognition Conference*, 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 (CVPR)*, pp. 13876–13885, June 2025a.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. Motiongpt3: Human motion as a second modality, 2025b. URL https://arxiv.org/abs/2506.24086.