SCALE: Semantic- and Confidence-Aware Conditional Variational Autoencoder for Zero-shot Skeleton-based Action Recognition

Soroush Oraki [1] Feng Ding [1] Jie Liang [2]

[1] School of Engineering Science, Simon Fraser University, Canada
[2] School of Information Science and Technology, Eastern Institute of Technology, Ningbo, China

OUTLINE

- Task
- Background & Motivation
- Method
- Experimental Setup
- Results
- Ablation
- Conclusion

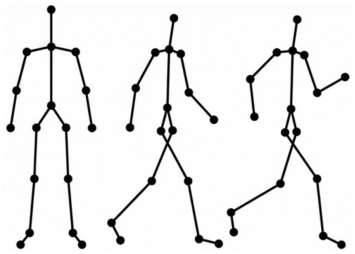
OUTLINE

- Task
- Background & Motivation
- Method
- Experimental Setup
- Results
- Ablation
- Conclusion

TASK

Zero-Shot Skeleton-based Action Recognition (ZSSAR)

- **Skeletons as input**
  - Compact 3D joint trajectories over time
  - Robust to background, lighting and viewpoint
  - Privacy-preserving



- **The open-world problem**
  - New actions appear constantly
  - Labeling skeletons for every class is impractical

- **Zero-shot setting**
  - Recognize unseen classes using only text descriptions of their labels
  - No training skeletons from the unseen classes

SEEN classes — {skeletons + text}

walking, waving, drinking, …

Model trains only on these

UNSEEN classes — {text only}

throwing, falling, …

At test time, recognize only from language description
- *First seen in test phase*

OUTLINE

- Task
- Background & Motivation
- Method
- Experimental Setup
- Results
- Ablation
- Conclusion

MOTIVATION

Where prior approaches struggle

Alignment-based

- Learn a shared skeleton-text embedding space
  - e.g. ReViSE, structured prompts (PURLS, STAR)
- **Limitation**
  - Brittle when label text underspecifies fine-grained motion; confusable unseen classes

Generative-based

Model class-conditional feature distributions

VAE-based

- Shared latent via VAEs (CADA-VAE, SynSE, SA-DVAE)
- No explicit uncertainty-aware discrimination of unseen classes

Diffusion-based

- Text-conditioned denoising (TDSM)
- Heavy training; stochastic, multi-trial inference

Our take: recast recognition as a deterministic, uncertainty-aware **energy ranking** — no explicit skeleton-text alignment, no test-time sampling.

BACKGROUND

Prior work

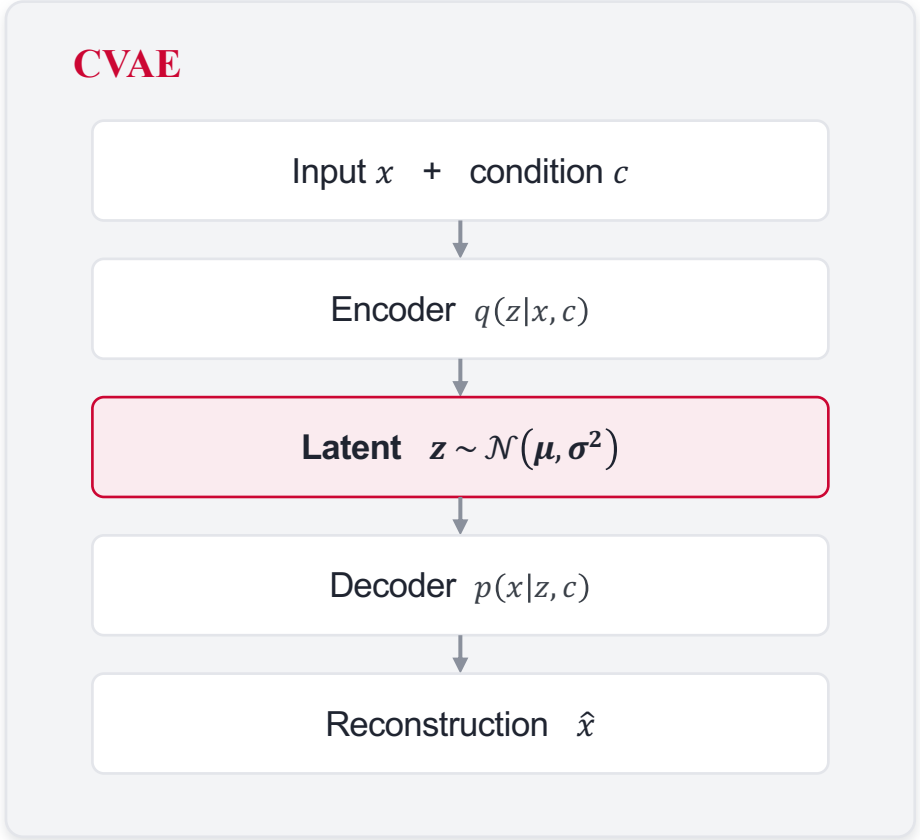
- **Cross-modal alignment**
  - Shared pose-language spaces: ReViSE, CADA-VAE
  - Structured language & richer semantics: SynSE (parts-of-speech), MSF
  - Disentangling / mutual information: SA-DVAE, SMIE
  - LLM part-aware descriptions & prompts: PURLS, STAR
- **Generative inference**
  - Text-conditioned diffusion: TDSM — strong, but heavy & stochastic
- **Energy-based & uncertainty-aware learning**
  - Energy as a unifying view of prediction & ranking [10]
  - EBMs for open-set recognition via energy scores [11]
- **The gap we address**
  - Most methods rely on explicit skeleton-text alignment
  - We instead rank class-conditional energies, with uncertainty built in

OUTLINE

- Task
- Background & Motivation
- Method
- Experimental Setup
- Results
- Ablation
- Conclusion

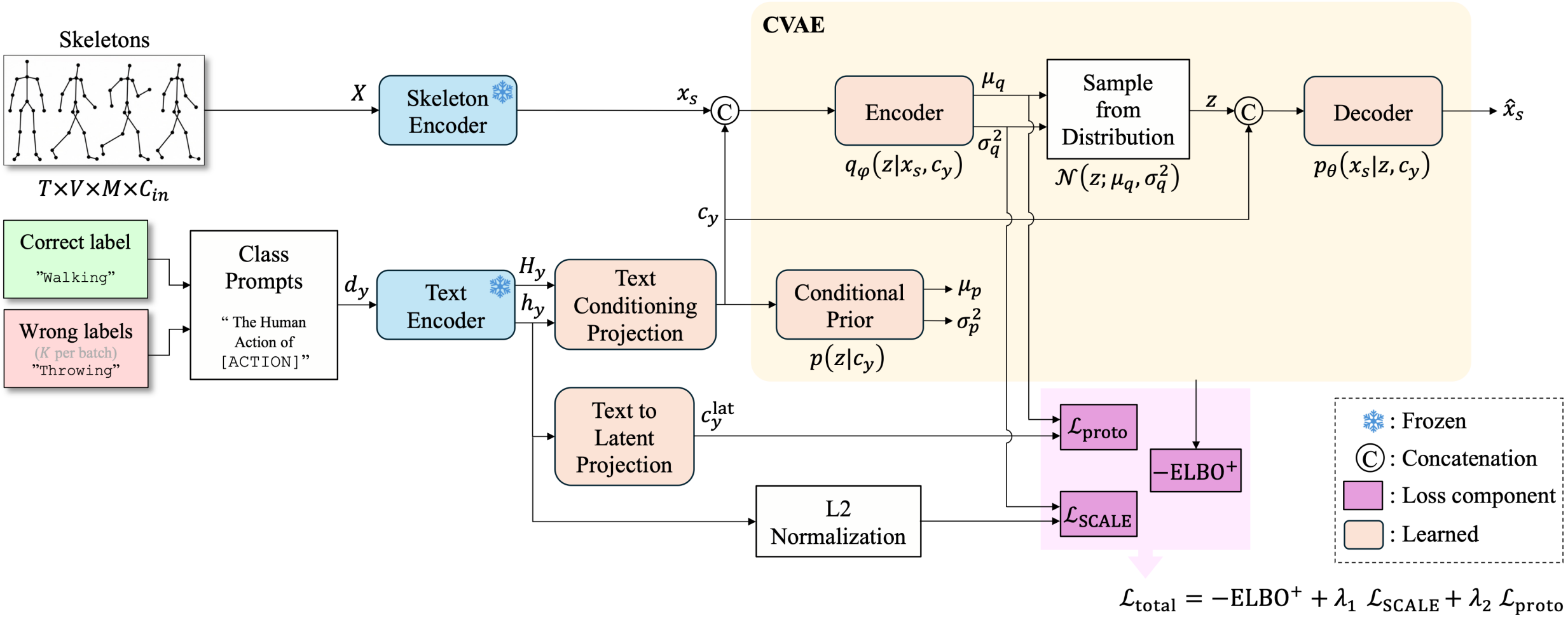
Conditional Variational Autoencoder (CVAE)

- **Variational Autoencoder (VAE)**
  - Encoder $q(z|x)$: maps input to a latent distribution
  - Decoder $p(x|z)$: reconstructs the input from latent z
  - Trained by maximizing the “ELBO” = “reconstruction − KL”
- **Conditional VAE (CVAE)**
  - Add a condition c : $q(z|x,c)$, $p(x|z,c)$, prior $p(z|c)$
  - Reconstruction & likelihood become condition-aware
- **Why it fits ZSSAR**
  - Let c = text embedding (of labels) → evaluate likelihood for unseen classes
  - No unseen skeletons needed; ELBO acts as a compatibility score



METHOD

Training Phase



METHOD

Energy score

For a skeleton feature and a candidate class, the CVAE gives an evidence lower bound, which we read as a compatibility score and negate into an energy:

$$\mathrm{ELBO}(x_s, c_y) = \mathbb{E}_{q_\phi(z|x_s,c_y)}\left[\log p_\theta(x_s \mid z, c_y)\right] - \beta\, D_{\mathrm{KL}}\Big(q_\phi(z \mid x_s, c_y) \,\|\, p_\theta(z \mid c_y)\Big)$$

$$E(x_s, c_y) = -\mathrm{ELBO}(x_s, c_y)$$

Why condition BOTH the prior and the decoder on frozen text?

- **Decoder**
  - Reconstruction becomes class-specific: the same z decodes differently per class, so the likelihood term reflects class fit
- **Prior**
  - Each class gets its own latent region $p(z|c_y)$
  - The KL term is class-aware and computable for unseen classes

METHOD

Semantic- & Confidence-Aware Listwise Energy (SCALE)

ELBO alone does not separate competing classes very well.

- **(1) Semantic bias toward hard negatives**
  - Aggregate negatives with a soft-max weighted by text similarity s_j
    - confusable classes count more

$$\tilde{E}_{\text{neg}}(x_s) = \log \sum_{j=1}^{K} \exp\left(E(x_s, c_{y_j^-}) + \alpha s_j\right)$$

- **(2) Confidence from posterior uncertainty**
  - Variance u relaxes the margin & down-weights ambiguous samples; confident ones enforce stricter separation

$$u(x_s, c_{y^+}) = \frac{1}{D_z} \sum_{d=1}^{D_z} \sigma_{q,d}^2 \quad \text{(mean posterior variance)}$$

- **(3) Energy gap**
  - τ is base margin
  - β_u controls margin relaxation.

$$\Delta(x_s) = E(x_s, c_{y^+}) - \tilde{E}_{\text{neg}}(x_s) + \tau - \beta_u\, u(x_s, c_{y^+})$$

- **(4) SCALE loss**
  - Weighting based on uncertainty;

$$\mathcal{L}_{\text{SCALE}} = \exp\left(-\gamma\, u(x_s, c_{y^+})\right) \cdot \log\left(1 + \exp(\Delta(x_s))\right)$$

METHOD

Overall Training Loss

Latent prototype contrast

- Treat the posterior mean μ_q as the sample's latent code
- Pull it toward its text-derived prototype $c_y{}^{\text{lat}} = g(h_y)$
- Cosine-similarity InfoNCE

$$\mathcal{L}_{\text{proto}}(x_s, y^+) = -\log \frac{\exp\left(\text{sim}(\mu_q(x_s, c_{y^+}), c_{y^+}^{\text{lat}})/\lambda\right)}{\sum_j \exp\left(\text{sim}(\mu_q(x_s, c_{y^+}), c_j^{\text{lat}})/\lambda\right)}$$

***Goal**: Organize the latent space by semantics, without direct skeleton-text feature matching.*

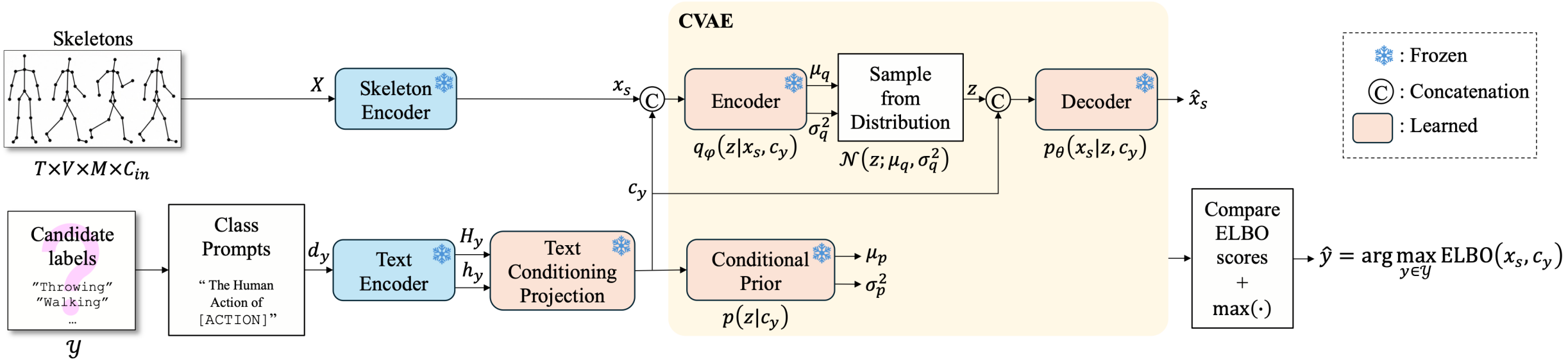
Overall Training Loss

$$\mathcal{L}_{\text{total}} = -\text{ELBO}^+ + \lambda_1 \cdot \mathcal{L}_{\text{SCALE}} + \lambda_2 \cdot \mathcal{L}_{\text{proto}}$$

- $-ELBO^+$ · generative fit for the true class
- $\mathcal{L}_{SCALE}$ · uncertainty-aware listwise discrimination
- $\mathcal{L}_{proto}$ · semantic structure in latent space

METHOD

Inference Phase



1 **Encode once**

Extract skeleton feature x_s with the frozen encoder; a single forward pass.

2 **Score every class**

For each candidate unseen class, condition on its text and evaluate $ELBO(x_s, c_y)$.

3 **Pick the best**

Predict $\hat{y} = \arg\max ELBO$.

OUTLINE

- Task
- Background & Motivation
- Method
- Experimental Setup
- Results
- Ablation
- Conclusion

EXPERIMENTS

Datasets & Setup

- **Datasets**
  - NTU RGB+D 60 — 56,880 sequences, 60 classes
  - NTU RGB+D 120 — 114,480 sequences, 120 classes
  - 8 zero-shot splits: SynSE (milder) & PURLS (harder)
- **Protocol**
  - Unseen skeletons excluded from training (zero-shot)
  - Text descriptions = only supervision for unseen classes
- **Model**
  - Skeleton encoder: frozen Shift-GCN
  - Text encoder: frozen CLIP ViT-B/32, simple label prompts
  - Only CVAE + projection/prototype heads trained
- **Training**
  - Optimizer: AdamW
  - β cyclical annealing: to avoid posterior collapse
  - Single NVIDIA RTX 4090

OUTLINE

- Task
- Background & Motivation
- Method
- Experimental Setup
- Results
- Ablation
- Conclusion

RESULTS

State-of-the-art comparison (Top-1 %)

Method	Venue	NTU-60 (Acc %)				NTU-120 (Acc %)			
		55/5	48/12	40/20	30/30	110/10	96/24	80/40	60/60
ReViSE [1]	ICCV'17	53.9	17.5	24.3	14.8	55.0	32.4	19.5	8.3
JPoSE [2]	ICCV'19	64.8	28.8	20.1	12.4	51.9	32.4	13.7	7.7
CADA-VAE [3]	CVPR'19	76.8	29.0	16.2	11.5	59.5	35.8	10.6	5.7
SynSE [4]	ICIP'21	75.8	33.3	19.9	12.0	62.7	38.7	13.6	7.7
SMIE [5]	ACMM'23	78.0	40.2	–	–	65.7	45.3	–	–
SA-DVAE [6]	ECCV'24	82.4	41.4	–	–	68.8	46.1	–	–
STAR [7]	ACMM'24	81.4	45.1	–	–	63.3	44.3	–	–
PURLS [8]	CVPR'24	79.2	41.0	31.0	23.5	72.0	52.0	28.4	19.6
TDSM [9]	ICCV'25	**86.5**	**56.0**	**36.1**	25.9	**74.2**	**65.1**	**37.0**	**27.2**
SCALE (Ours)	ICPR'26	**84.5**	**50.0**	**35.5**	**27.7**	**73.6**	**54.9**	**32.1**	**23.5**

RESULTS

SCALE vs. TDSM

TDSM is the only method that outperforms SCALE in terms of accuracy, but at a very different cost (per sample):

107×

fewer parameters

2.43 M vs 261 M

227×

faster inference

0.454 ms vs 103 ms

~16,000×

fewer FLOPs

0.002 vs 32.0 GFLOPs

TDSM runs 10 stochastic trials per sample at inference; Our SCALE method uses a single deterministic pass.

OUTLINE

- Task
- Background & Motivation
- Method
- Experimental Setup
- Results
- Ablation
- Conclusion

ABLATION 1

Effect of loss components

Training Loss	NTU-60 (Acc, %)		NTU-120 (Acc, %)	
	[55/5]	[48/12]	[110/10]	[96/24]
$-\text{ELBO}^{+}$	78.7	43.3	65.1	47.1
$-\text{ELBO}^{+} + \mathcal{L}_{\text{proto}}$	$81.1^{\uparrow 2.4}$	$45.7^{\uparrow 2.4}$	$69.0^{\uparrow 3.9}$	$50.6^{\uparrow 3.5}$
$-\text{ELBO}^{+} + \mathcal{L}_{\text{SCALE}}$	$82.7^{\uparrow 4.0}$	$46.6^{\uparrow 3.3}$	$70.6^{\uparrow 5.5}$	$52.2^{\uparrow 5.1}$
$-\text{ELBO}^{+} + \mathcal{L}_{\text{SCALE}} + \mathcal{L}_{\text{proto}}$	$\mathbf{84.5}^{\uparrow 5.8}$	$\mathbf{50.0}^{\uparrow 6.7}$	$\mathbf{73.6}^{\uparrow 8.5}$	$\mathbf{54.9}^{\uparrow 7.8}$

Takeaways

- **Each term helps on its own**
- **Best when combined**
  - Latent structure + uncertainty-aware discrimination are both needed

ABLATION 2

Effect of semantic similarity bias α

α	NTU-60 (Acc, %)		NTU-120 (Acc, %)	
	[55/5]	[48/12]	[110/10]	[96/24]
$\alpha = 0$	83.9	49.6	72.9	54.3
$\boldsymbol{\alpha = 1.0}$	**84.5**$^{\uparrow 0.6}$	**50.0**$^{\uparrow 0.4}$	**73.6**$^{\uparrow 0.7}$	**54.9**$^{\uparrow 0.6}$
$\alpha = 2.0$	84.3$^{\uparrow 0.4}$	50.1$^{\uparrow 0.5}$	73.5$^{\uparrow 0.6}$	54.4$^{\uparrow 0.1}$
$\alpha = 3.0$	84.4$^{\uparrow 0.5}$	49.8$^{\uparrow 0.2}$	73.3$^{\uparrow 0.4}$	54.5$^{\uparrow 0.2}$

Takeaways

- **Hard-negative emphasis helps**
  - α = 1.0 beats α = 0 by +{0.4-0.7}
- **But more is not better**
  - α = 2.0 or 3.0 give marginal or slightly worse results
  - Overweighting a few negatives hurts generalization

ABLATION 3

Effect of uncertainty modeling

β_u: *Margin relaxation* γ: *Loss reweighting*

Uncertainty Modeling	β_u	γ	NTU-60 (Acc, %)		NTU-120 (Acc, %)	
			[55/5]	[48/12]	[110/10]	[96/24]
Fixed margin, No reweighting	0	0	83.7	48.2	73.1	53.2
Uncertainty-adaptive margin only	0.5	0	84.3$^{\uparrow 0.6}$	49.4$^{\uparrow 1.2}$	73.5$^{\uparrow 0.4}$	54.5$^{\uparrow 1.3}$
Uncertainty-based reweighting only	0	2.0	84.0$^{\uparrow 0.3}$	48.6$^{\uparrow 0.4}$	73.3$^{\uparrow 0.2}$	53.9$^{\uparrow 0.7}$
Full $\mathcal{L}_{\text{SCALE}}$ loss	**0.5**	**2.0**	**84.5**$^{\uparrow 0.8}$	**50.0**$^{\uparrow 1.8}$	**73.6**$^{\uparrow 0.5}$	**54.9**$^{\uparrow 1.7}$

- Both mechanisms help and are complementary
  - the adaptive margin matters most on harder splits

OUTLINE

- Task
- Background & Motivation
- Method
- Experimental Setup
- Results
- Ablation
- Conclusion

CONCLUSION

Contributions

✓ Alignment-free ZSSAR

Recognition as a deterministic, listwise comparison of class-conditional energies from a lightweight CVAE.

✓ Uncertainty-aware discrimination

Posterior variance adapts margins and reweighting; semantic bias targets hard negatives.

✓ 2nd-best accuracy

- Beats prior VAE- & alignment-based methods on NTU-60/120; best on the hardest split.
- Not better than the SOTA diffusion-based model, but much faster and more efficient with relatively comparable accuracy.

✓ Tiny & deterministic

Lightweight, single-pass, and fast inference

REFERENCES

[1] Hubert Tsai, Y.H., Huang, L.K., Salakhutdinov, R.: Learning robust visual-semantic embeddings. In: Proceedings of the IEEE International conference on Computer Vision. pp. 3571–3580 (2017)

[2] Wray, M., Larlus, D., Csurka, G., Damen, D.: Fine-grained action retrieval through multiple parts-of-speech embeddings. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 450–459 (2019)

[3] Schonfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z.: 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)

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

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

[6] Li, S.W., Wei, Z.X., Chen, W.J., Yu, Y.H., Yang, C.Y., Hsu, J.Y.j.: Sa-dvae: Improving zero-shot skeleton-based action recognition by disentangled variational autoencoders. In: European Conference on Computer Vision. pp. 447–462. Springer (2024)

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

[8] Zhu, A., Ke, Q., Gong, M., Bailey, J.: Part-aware unified representation of language and skeleton for zero-shot action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18761–18770 (2024)

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

[10] LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F., et al.: A tutorial on energy-based learning. Predicting structured data 1(0) (2006)

[11] Al Rahhal, M.M., Bazi, Y., Al-Dayil, R., Alwadei, B.M., Ammour, N., Alajlan, N.: Energy-based learning for open-set classification in remote sensing imagery. International Journal of Remote Sensing 43(15-16), 6027–6037 (2022)

SFU SIMON FRASER UNIVERSITY

ICPR 2026 International Conference on Pattern Recognition LYON

Thank you!

Questions?