Bridging the Skeleton-Text Modality Gap

Diffusion-Powered Modality Alignment for Zero-shot Skeleton-based Action Recognition

Jeonghyeok Do

Munchurl Kim†

KAIST EE
KAIST ELECTRICAL ENGINEERING

VICLab Video and Image Computing Lab

Korea Advanced Institute of Science and Technology

†Corresponding Author

ICCV HONOLULU HAWAII
OCT 19-23, 2025

Motivation

- Zero-shot Skeleton-based Action Recognition (ZSAR)
  - The fully supervised skeleton-based action recognition methods perform well
  - **Annotating every possible action** is impractical

Action labels

Throw baseball

Throw basketball

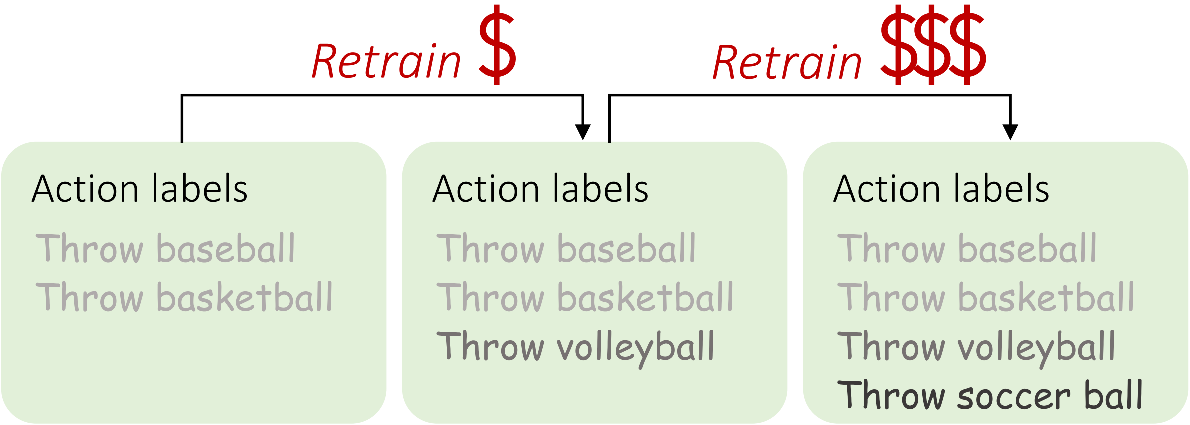
Throw volleyball

Throw soccer ball

Throw …

Motivation

- Zero-shot **S**keleton-based **A**ction **R**ecognition (**ZSAR**)
  - The fully supervised skeleton-based action recognition methods perform well
  - **Annotating every possible action** is **impractical**
  - **Retraining models** for new classes incurs a **significant cost**

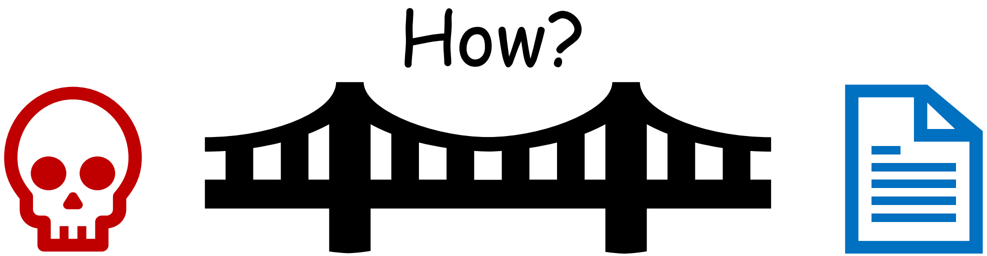


Motivation

- Zero-shot Skeleton-based Action Recognition (ZSAR)
  - Enabling **predictions for unseen actions** without requiring explicit training data
  - **Why ZSAR is possible?**
    - Human actions often *share common skeletal movement patterns* across related actions
    - ZSAR methods ***align*** *pre-learned skeleton features with text-based action descriptions*, allowing the models to **extrapolate from seen actions to unseen ones**

Motivation

- Zero-shot **S**keleton-based **A**ction **R**ecognition (**ZSAR**)
  - Significant challenges: *“the modality gap”*

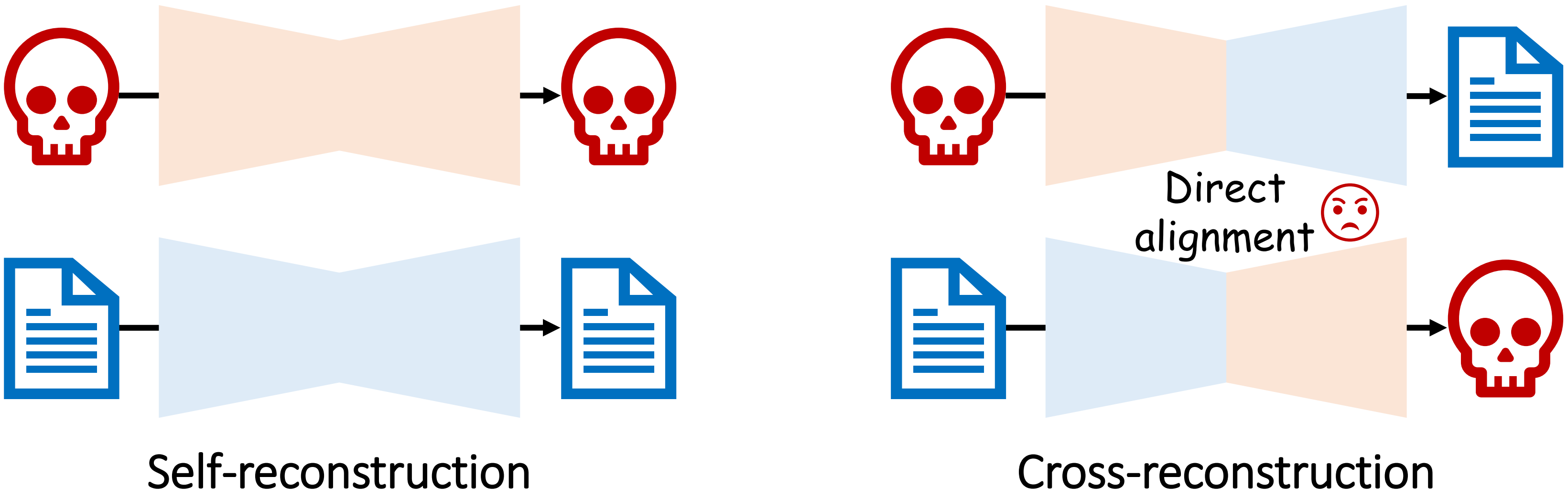


Temporal and spatial motion patterns

High-level semantic information

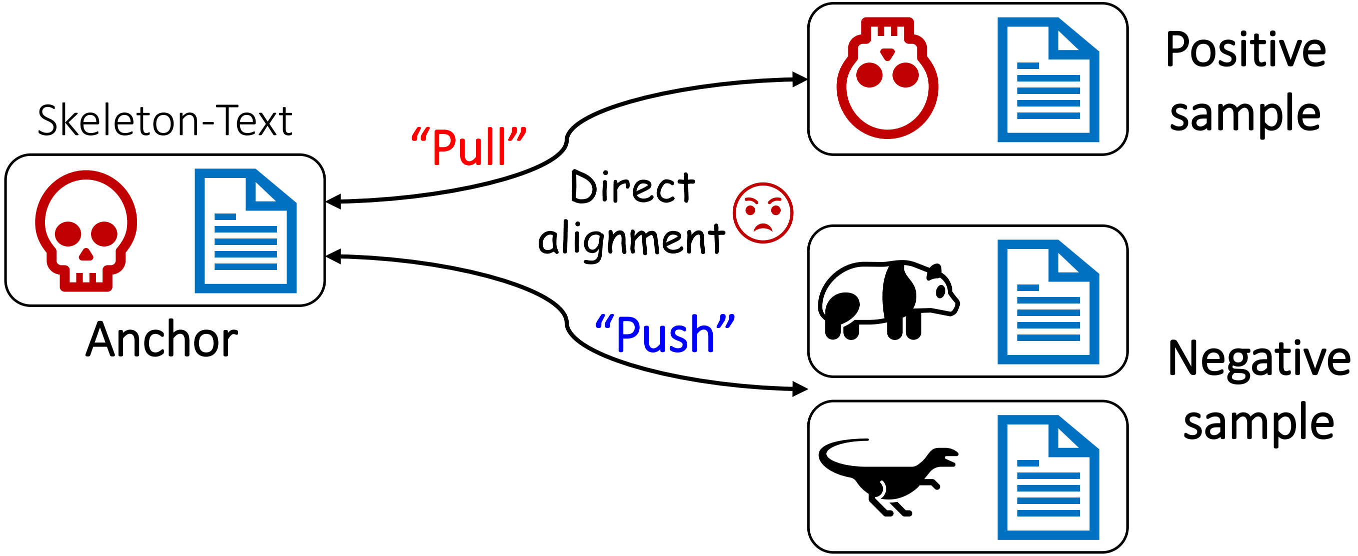
Motivation

- Previous ZSAR methods: VAE-based
  - Reconstructs skeleton-text feature pairs via **cross-reconstruction**
  - Recovers skeleton features from text and vice versa



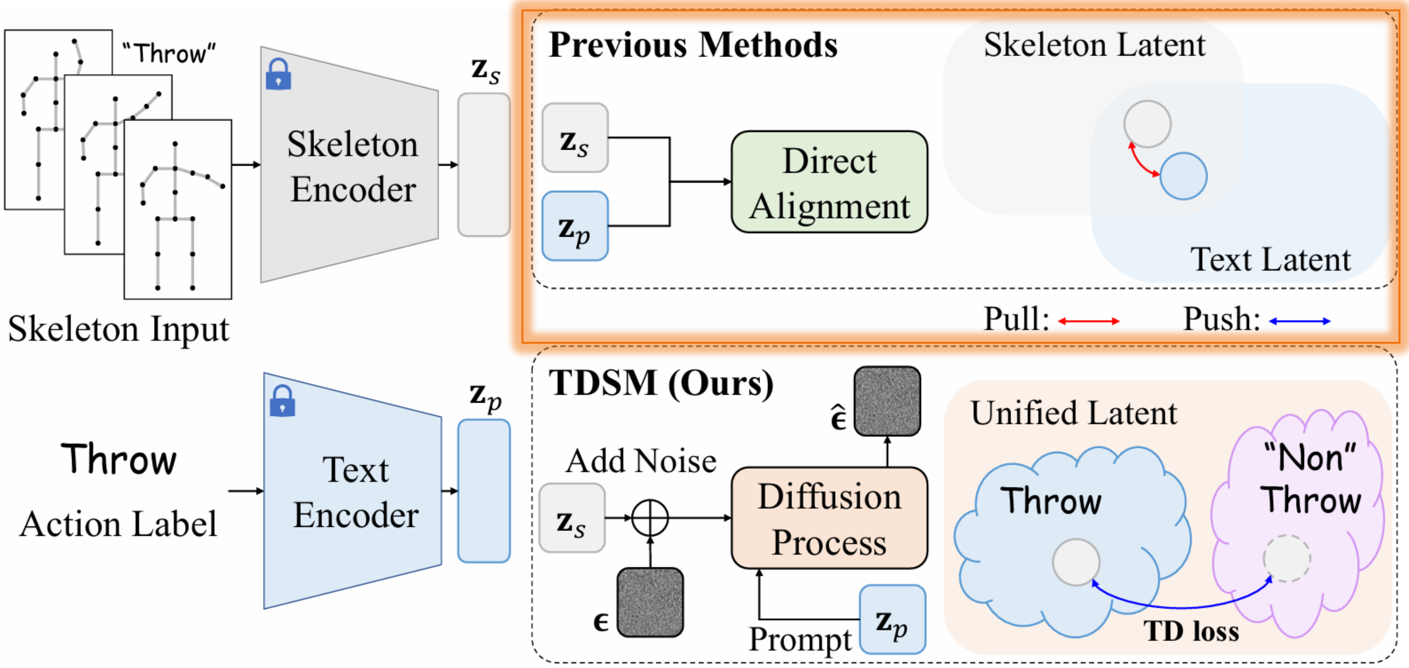
Motivation

- Previous ZSAR methods: Contrastive learning(CL)-based
  - Aligns skeleton and text features by **minimizing feature distance** through contrastive learning



Motivation

- Previous ZSAR methods: VAE-based, CL-based
  - Modality gap due to *direct alignment*



Motivation

- Proposed method: **T**riplet **D**iffusion for **S**keleton-Text **M**atching (**TDSM**)
  - Diffusion models effectively incorporate **conditioning signals** enabling strong cross-modal alignment

a space elevator, cinematic scifi art

A cheeseburger with juicy beef patties and melted cheese sits on top of a toilet that looks like a throne and stands in the middle of the royal chamber.

a hole in the floor of my bathroom with small gremlins living in it

a small office made out of car parts

This dreamlike digital art captures a vibrant, kaleidoscopic bird in a lush rainforest.

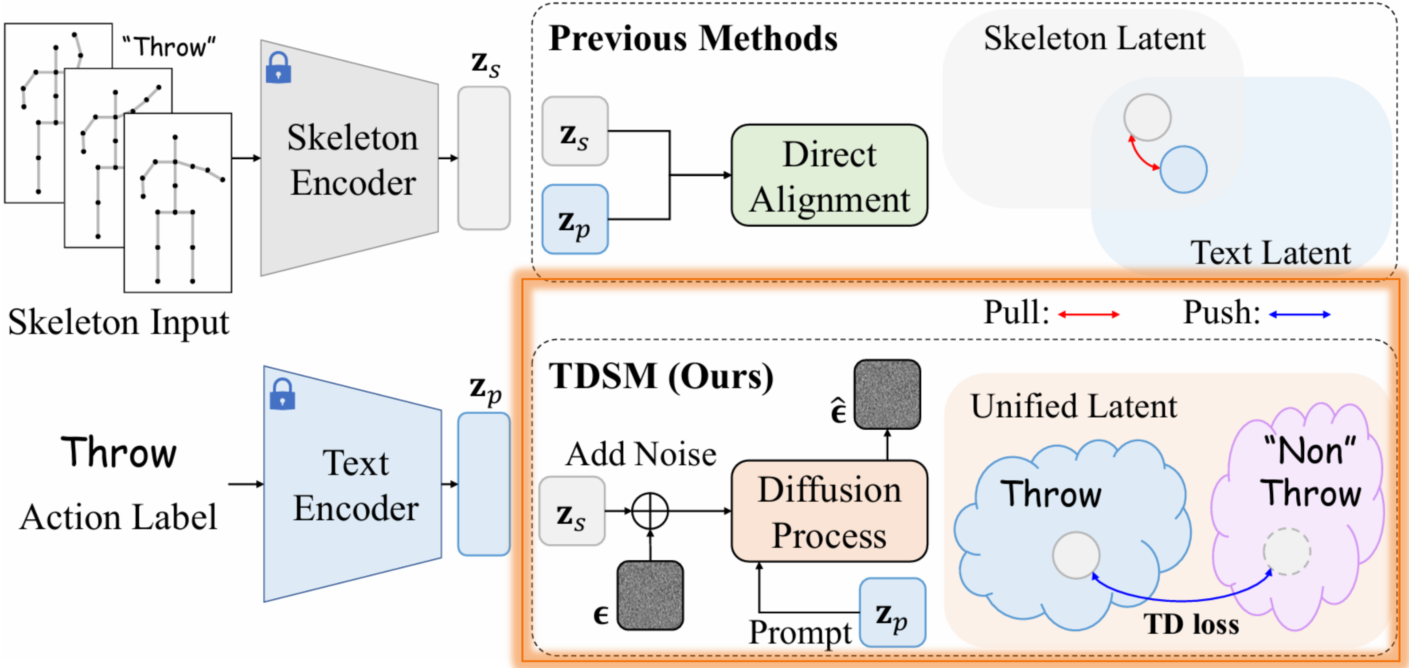
human life depicted entirely out of fractals

*e.g., **text-to-image** generation (Stable Diffusion v3.0)*

Rather than the generative ability, we are motivated by the **alignment property**

Motivation

- Proposed method: **T**riplet **D**iffusion for **S**keleton-Text **M**atching (**TDSM**)
  - Utilizes the cross-modality alignment power of diffusion models



Proposed Method

- We present a diffusion-based action recognition with zero-shot learning for skeleton inputs, **TDSM** which is the *first framework to apply diffusion models*

  - **Reverse diffusion process with text prompts**

    ↳ ***Implicitly align*** the *skeleton features with text prompts (action labels)*

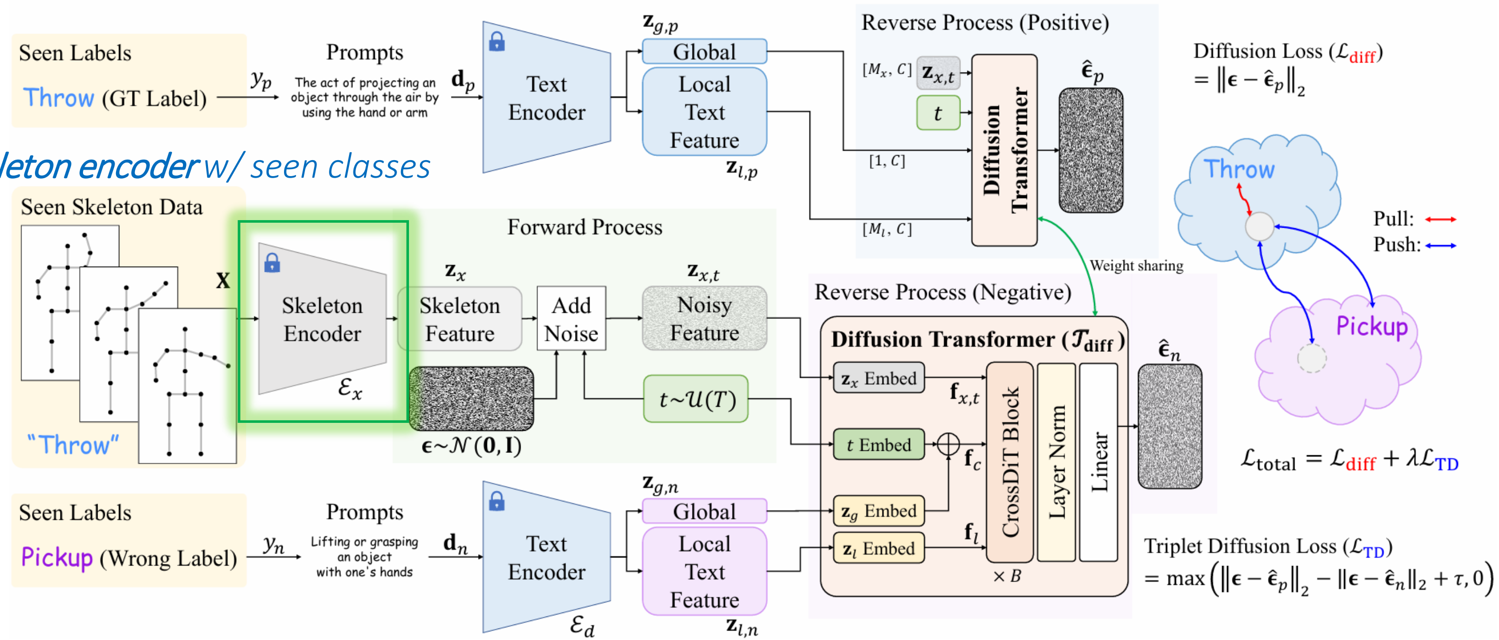
  - **Triplet diffusion (TD) loss**

    ↳ Enhance the model's ***discriminative power*** by *ensuring accurate denoising for correct skeleton-text pairs while suppressing it for incorrect pairs*

Proposed Method

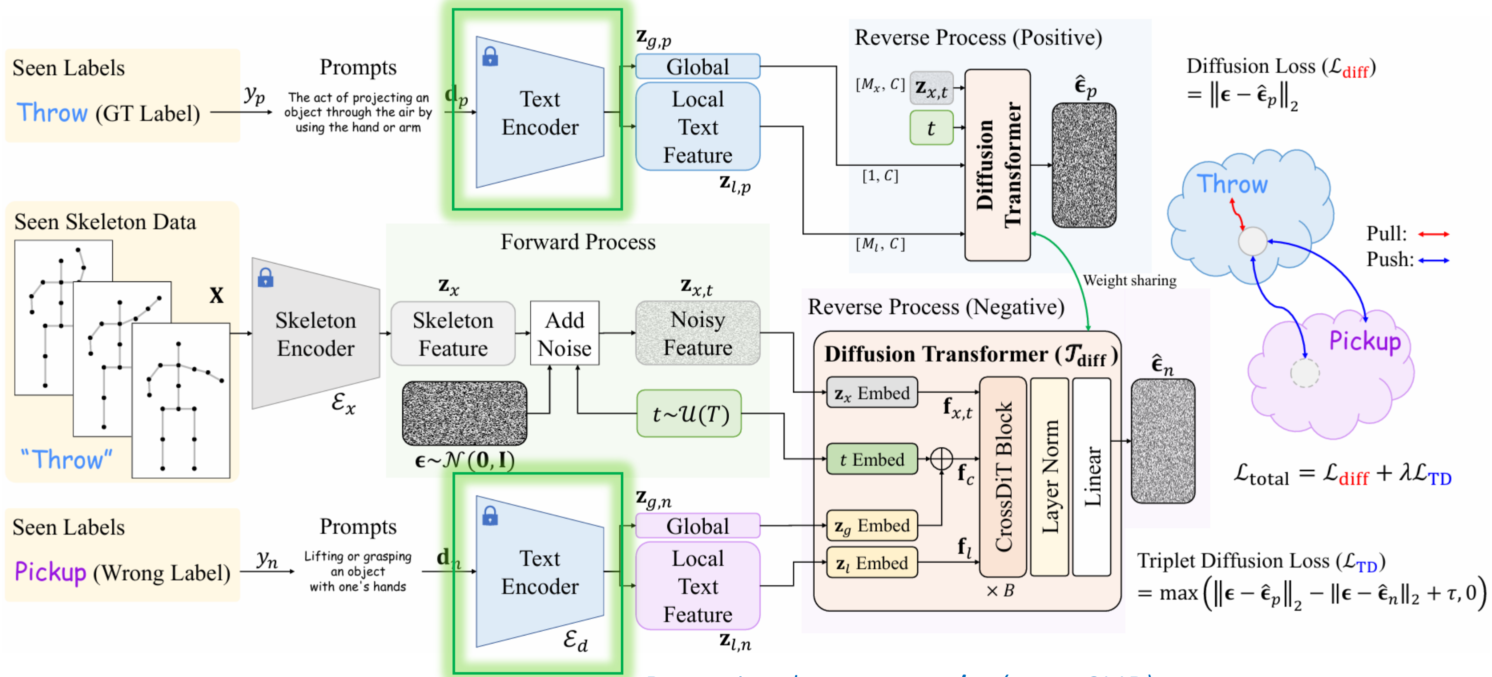
- Training framework of our **TDSM**: embedding inputs
  - Performs the diffusion process in a compact latent space

*Pretrains **skeleton encoder** w/ seen classes*



Proposed Method

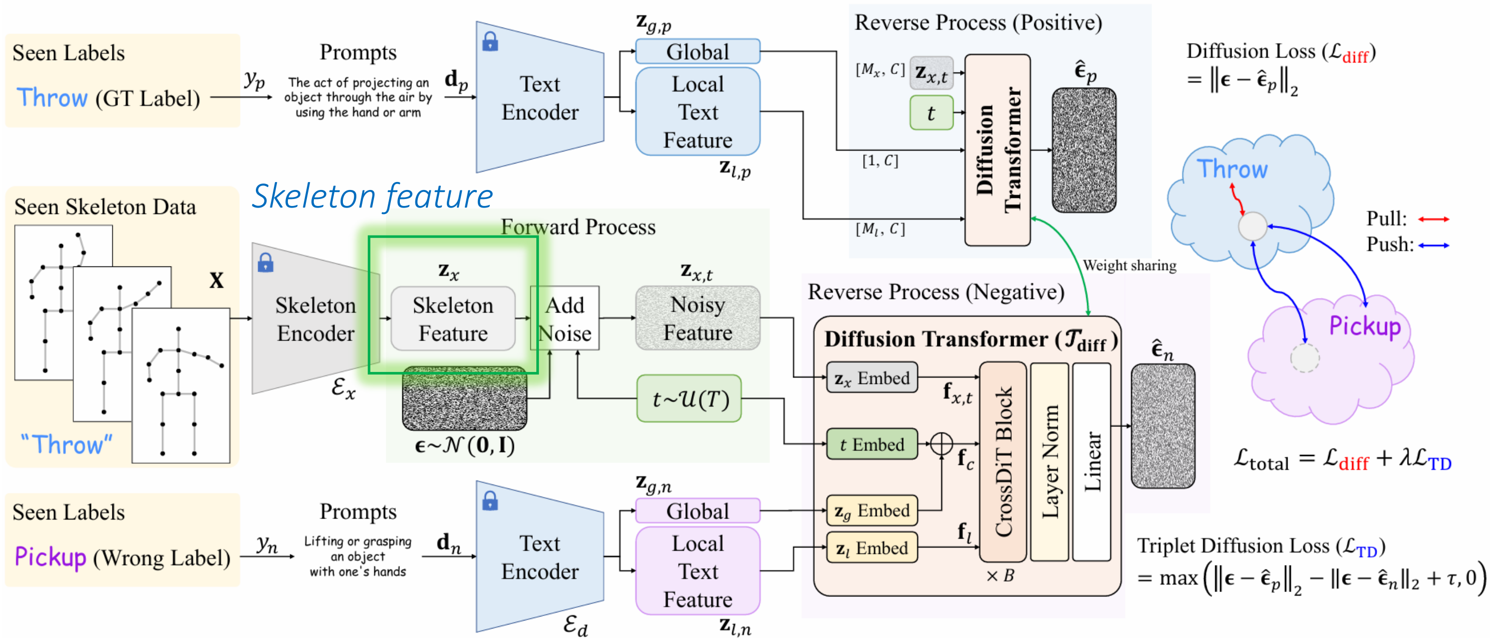
- Training framework of our **TDSM**: embedding inputs
  - Performs the diffusion process in a compact latent space



*Pretrained **text encoder** (e.g., CLIP)*

Proposed Method

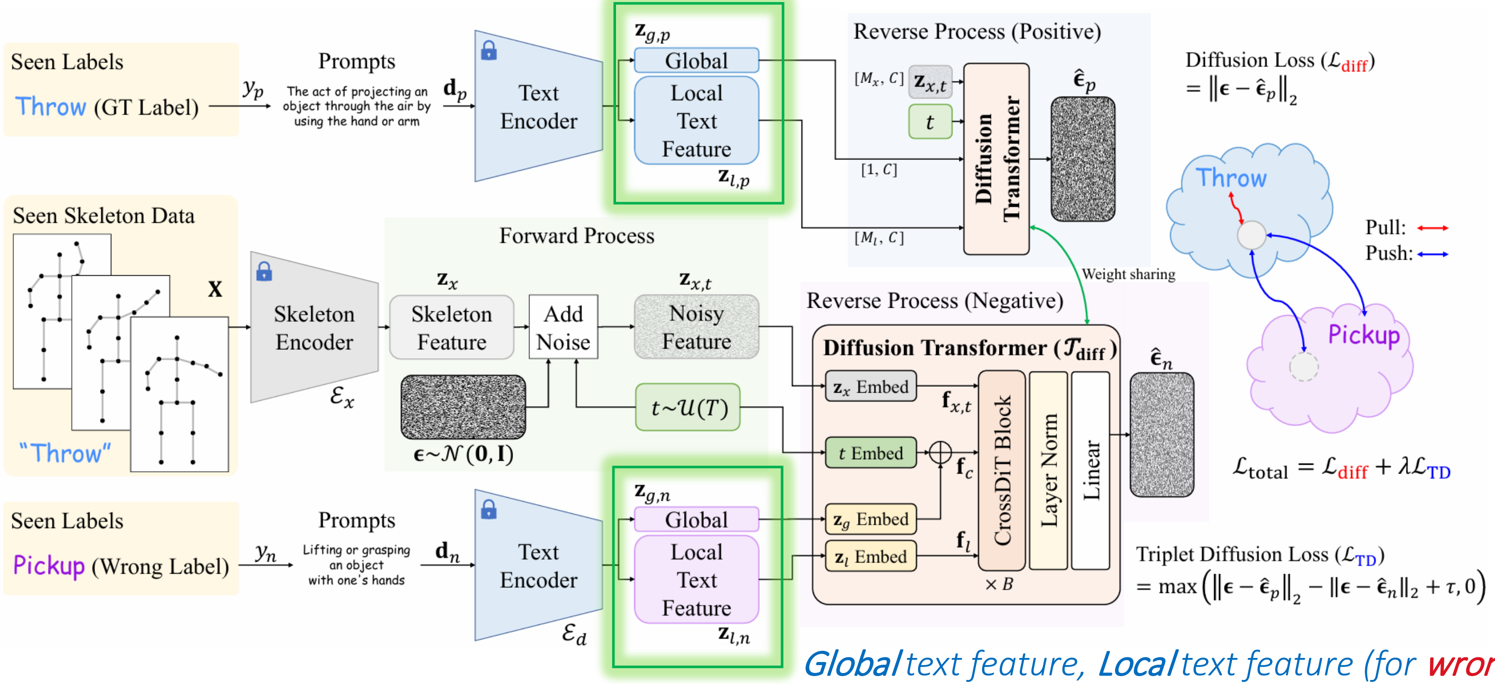
- Training framework of our **TDSM**: embedding inputs
  - **Embeds** skeleton and prompt input



Proposed Method

- Training framework of our **TDSM**: embedding inputs
  - **Embeds** skeleton and prompt input

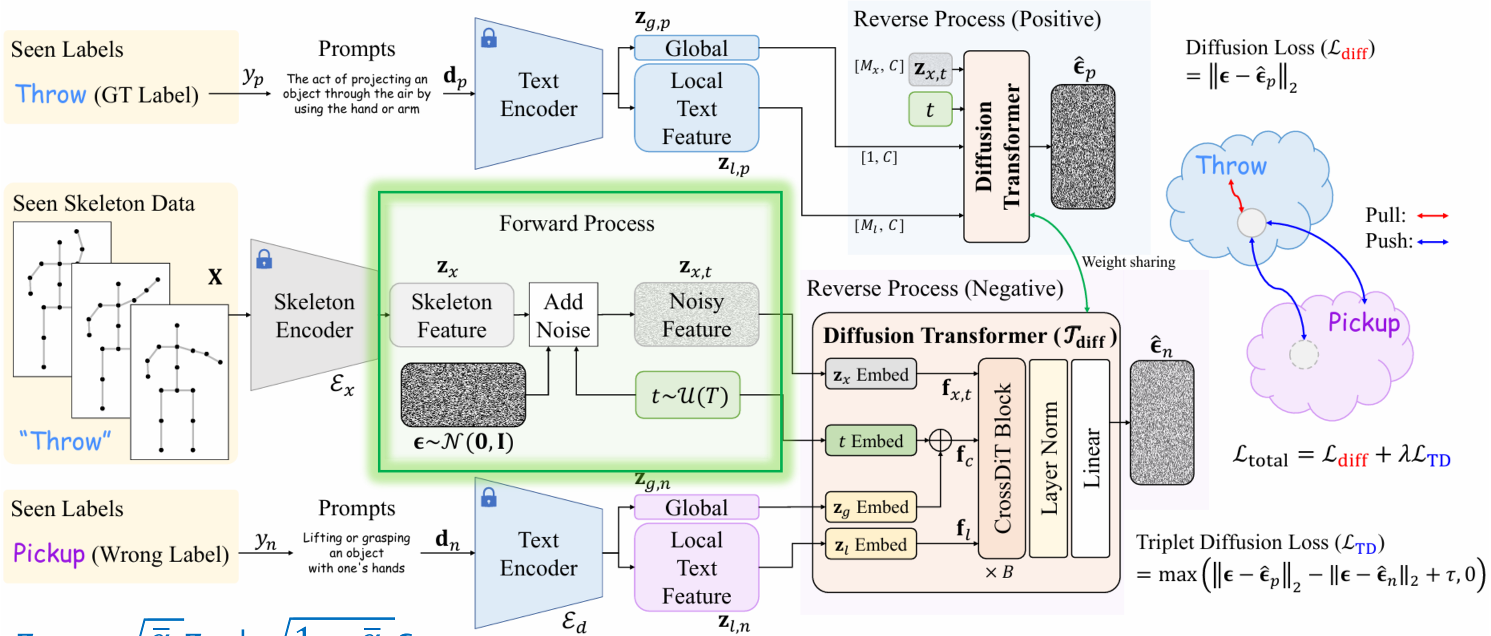
Global text feature, *Local* text feature (for *GT* label)



Global text feature, *Local* text feature (for *wrong* label)

Proposed Method

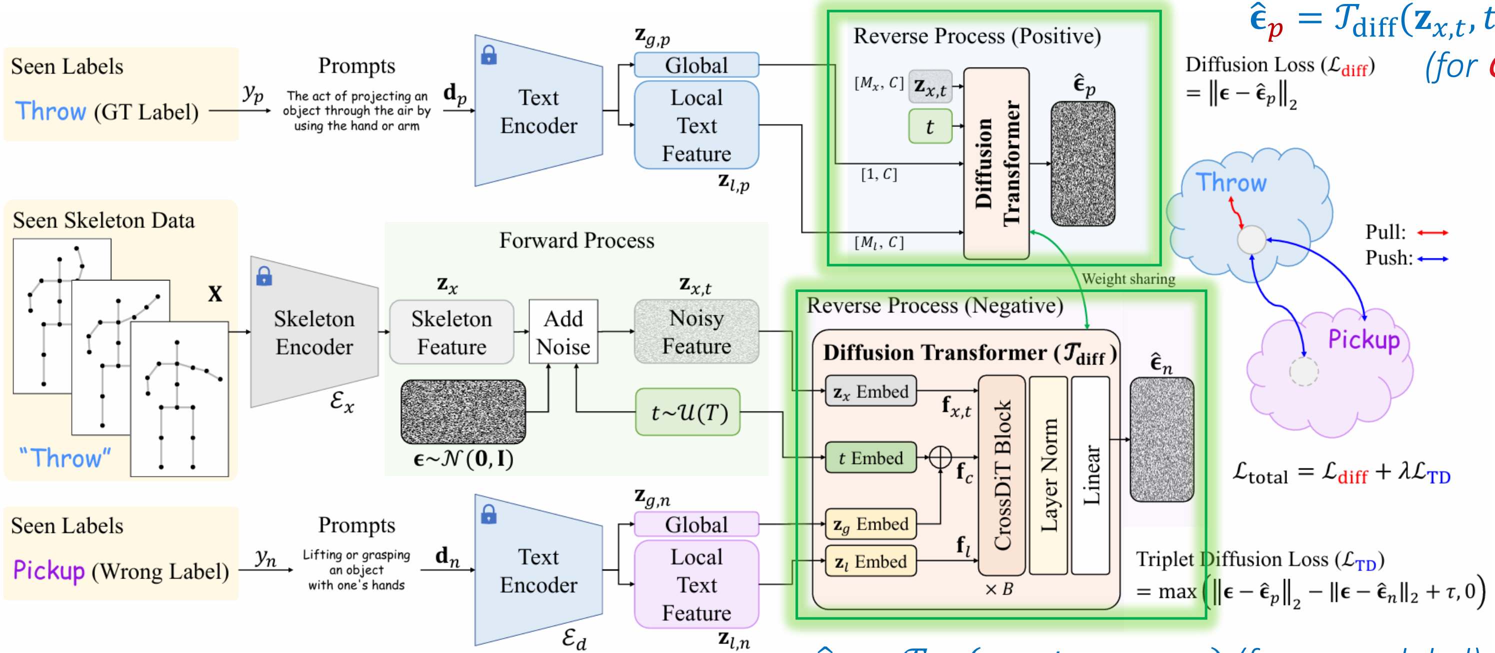
- Training framework of our **TDSM**: diffusion process (forward process)
  - Random Gaussian noise is added to the skeleton feature at a random timestep



$$\mathbf{z}_{x,t} = \sqrt{\bar{\alpha}_t}\mathbf{z}_x + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$$

Proposed Method

- Training framework of our **TDSM**: diffusion process (**reverse process**)
  - Network **predicts noise** from *noisy skeleton feature* conditioned on *text features*

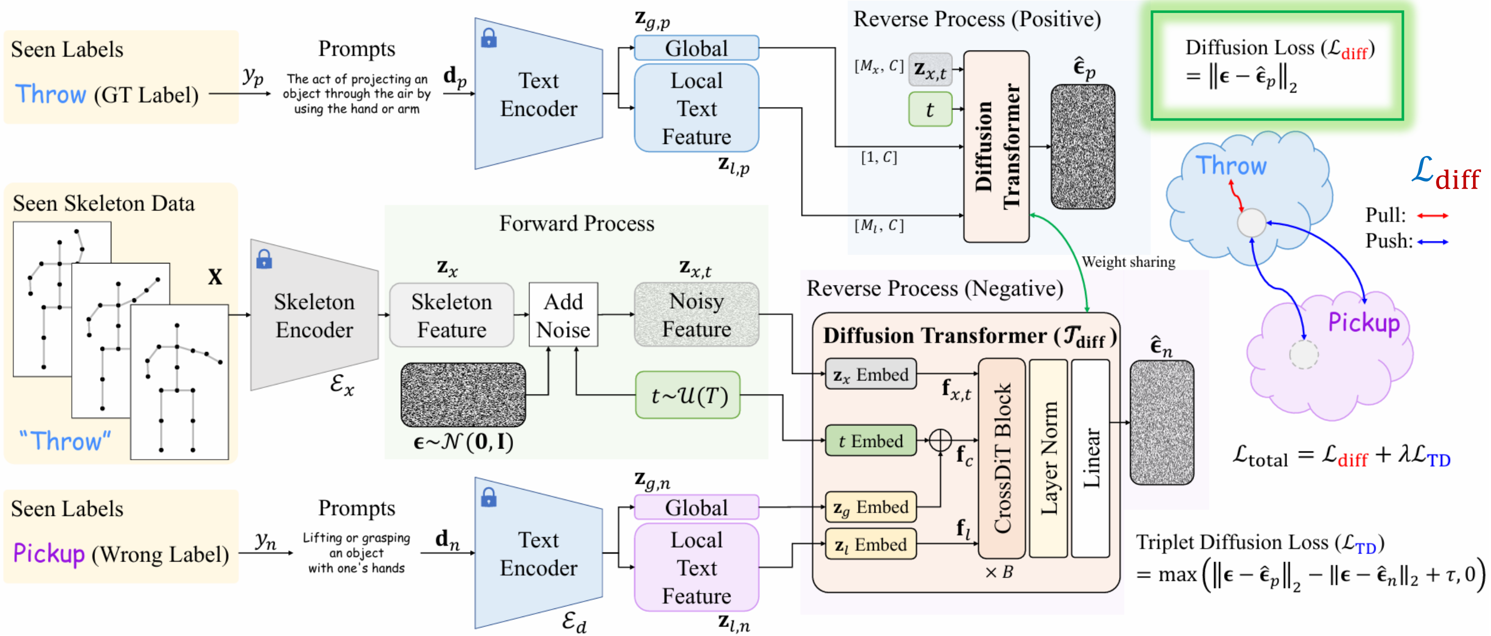


$\hat{\epsilon}_p = \mathcal{T}_{\text{diff}}(\mathbf{z}_{x,t}, t; \mathbf{z}_{g,p}, \mathbf{z}_{l,p})$ *(for GT label)*

$\hat{\epsilon}_n = \mathcal{T}_{\text{diff}}(\mathbf{z}_{x,t}, t; \mathbf{z}_{g,n}, \mathbf{z}_{l,n})$ *(for wrong label)*

Proposed Method

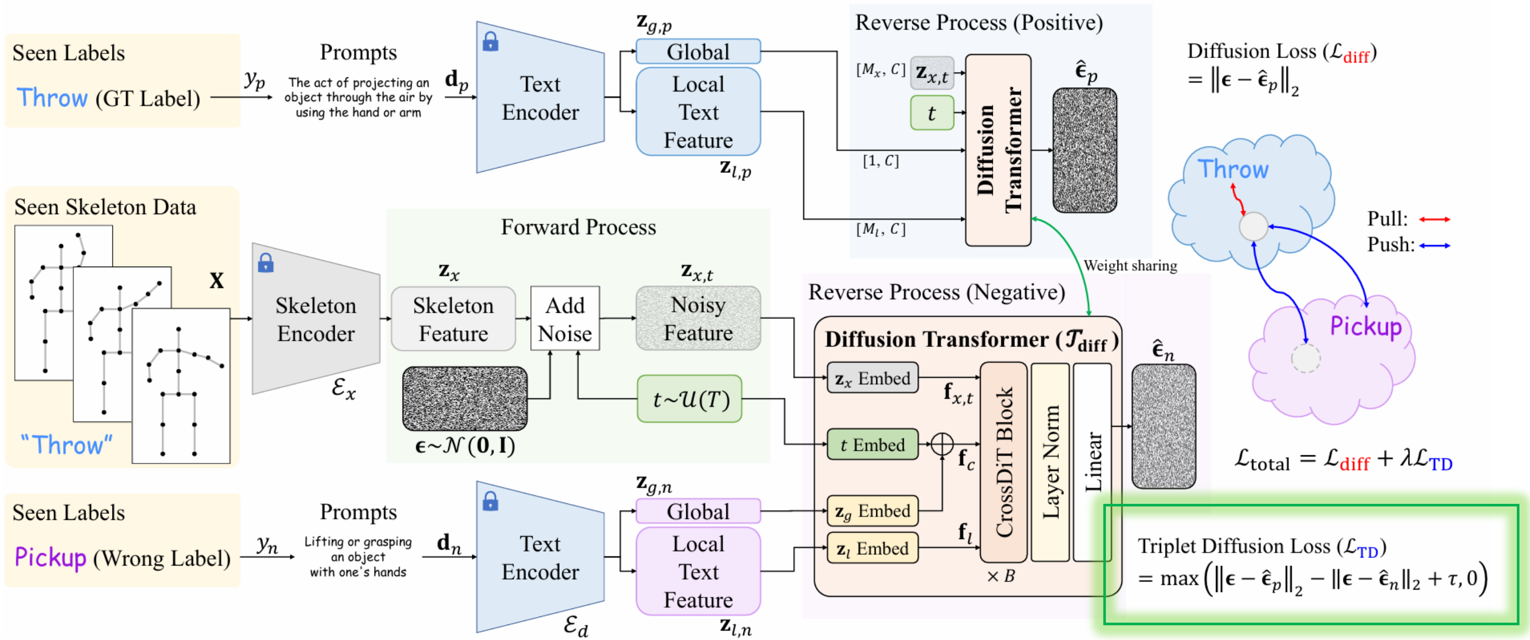
- Training framework of our **TDSM**: loss function (**diffusion loss**)
  - Diffusion loss ensures *accurate denoising* for positive skeleton-text(GT) pair



$$\mathcal{L}_{\text{diff}} = \|\epsilon - \hat{\epsilon}_p\|_2$$

Proposed Method

- Training framework of our **TDSM**: loss function (**triplet diffusion (TD) loss**)
  - TD loss enhances the ability to *differentiate between* GT/wrong label predictions



$$\mathcal{L}_{\text{TD}} = \max\left(\|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_p\|_2 - \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_n\|_2 + \tau, 0\right)$$

Proposed Method

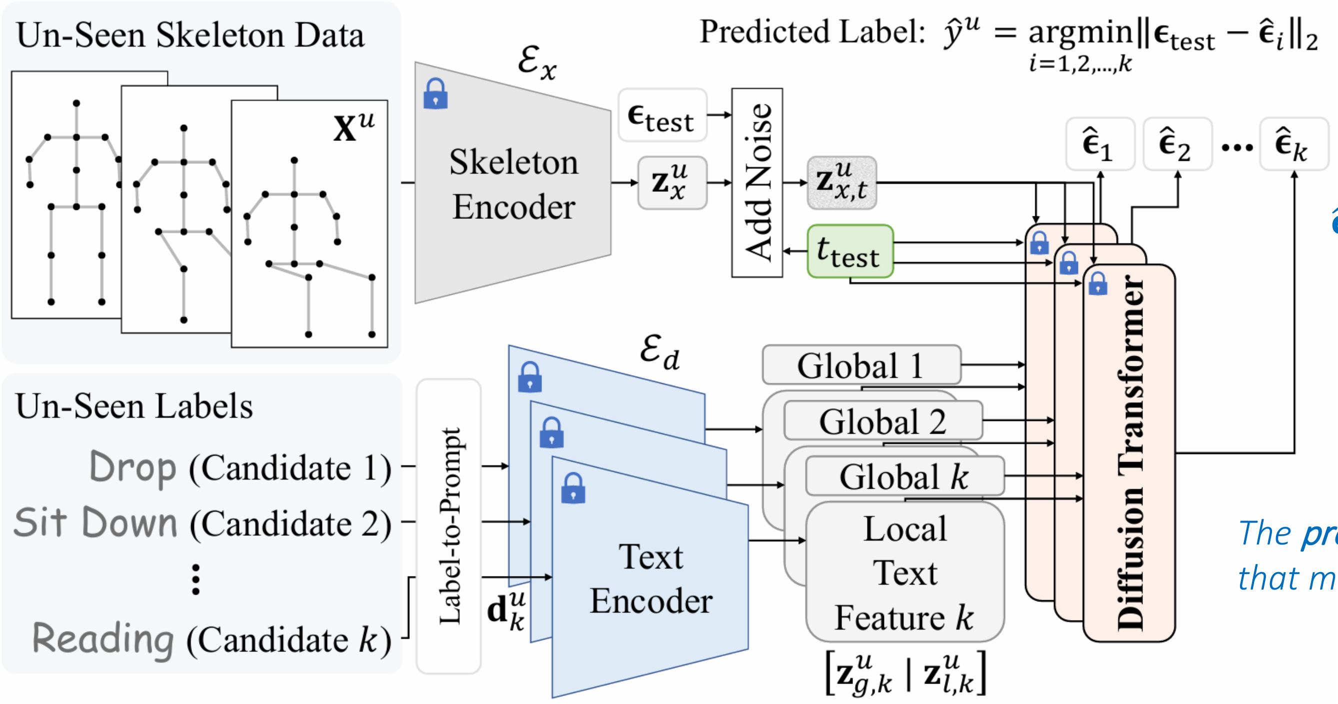
- Inference phase of our **TDSM**

  - **Enhance discriminative fusion through the TD loss**
    - *Denoise GT skeleton-text pairs* effectively while *preventing the fusion of incorrect pairs* within the *seen dataset*

  - **Selective denoising process promotes a robust fusion of skeleton and text features**
    - Allow the model to develop a *discriminative feature space* that can *generalize to unseen action labels*

Proposed Method

- Inference phase of our **TDSM**
  - *One-step inference* at a *fixed timestep (t_{test})* and *fixed noise ($\boldsymbol{\epsilon}_{\text{test}}$)*



$\hat{\boldsymbol{\epsilon}}_k = \mathcal{T}_{\text{diff}}(\mathbf{z}_{x,t}^u, t_{\text{test}}; \mathbf{z}_{g,k}^u, \mathbf{z}_{l,k}^u)$
(for un-seen label k)

*The **predicted label** is the one that minimizes $\|\boldsymbol{\epsilon}_{\text{test}} - \hat{\boldsymbol{\epsilon}}_k\|_2$*

Experiment Results

- **Quantitative Results** (Top-1 Acc ↑)
  - SynSE (standard) and PURLS (extreme) benchmarks

X/Y split
X: the # of seen classes
Y: the # of unseen classes

Methods	Publications	SysSE NTU-60 (Acc, %)		PURLS		SysSE NTU-120 (Acc, %)		PURLS	
		55/5 split	48/12 split	40/20 split	30/30 split	110/10 split	96/24 split	80/40 split	60/60 split
ReViSE [26]	ICCV 2017	53.91	17.49	24.26	14.81	55.04	32.38	19.47	8.27
JPoSE [67]	ICCV 2019	64.82	28.75	20.05	12.39	51.93	32.44	13.71	7.65
CADA-VAE [54]	CVPR 2019	76.84	28.96	16.21	11.51	59.53	35.77	10.55	5.67
SynSE [20]	ICIP 2021	75.81	33.30	19.85	12.00	62.69	38.70	13.64	7.73
SMIE [77]	ACM MM 2023	77.98	40.18	-	-	65.74	45.30	-	-
PURLS [79]	CVPR 2024	79.23	40.99	<u>31.05</u>	<u>23.52</u>	<u>71.95</u>	<u>52.01</u>	<u>28.38</u>	<u>19.63</u>
SA-DVAE [38]	ECCV 2024	<u>82.37</u>	41.38	-	-	68.77	46.12	-	-
STAR [8]	ACM MM 2024	81.40	<u>45.10</u>	-	-	63.30	44.30	-	-
TDSM (Ours)	**-**	**86.49**	**56.03**	**36.09**	**25.88**	**74.15**	**65.06**	**36.95**	**27.21**

Experiment Results

- **Quantitative Results** (Top-1 Acc ↑)
  - SMIE (generalization) benchmark: three distinct split

X/Y split
X: the # of seen classes
Y: the # of unseen classes

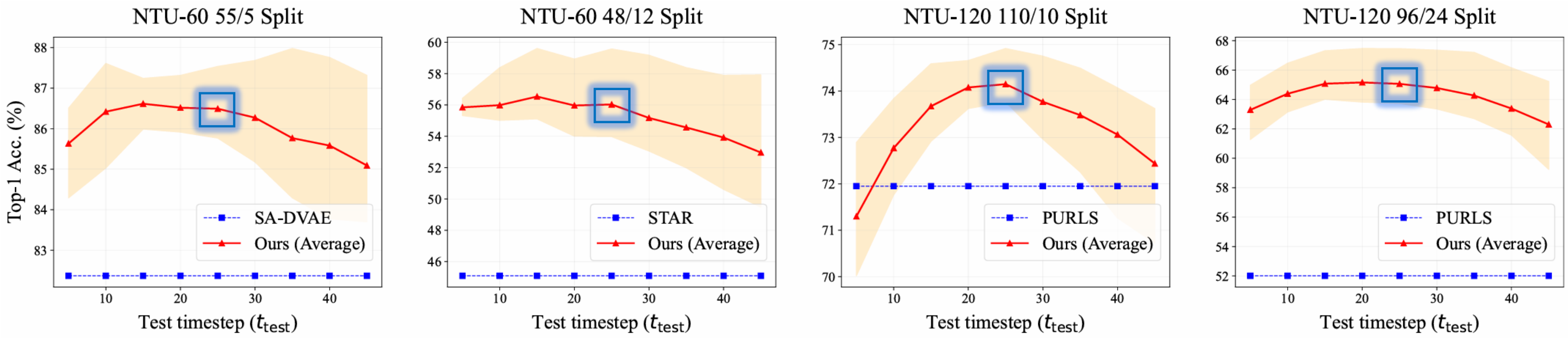
Methods	NTU-60 (Acc, %)	NTU-120 (Acc, %)	PKU-MMD (Acc, %)
	55/5 split	110/10 split	46/5 split
ReViSE [26]	60.94	44.90	59.34
JPoSE [67]	59.44	46.69	57.17
CADA-VAE [54]	61.84	45.15	60.74
SynSE [20]	64.19	47.28	53.85
SMIE [77]	65.08	46.40	60.83
SA-DVAE [38]	84.20	50.67	66.54
STAR [8]	77.50	-	70.60
TDSM (Ours)	**88.88**	**69.47**	**70.76**

Average of the three splits

Experiment Results

Ablation Study: Effect of varying inference timesteps t_{test}

Orange area
: the variation across 10 different random noise



Inference timestep: $t_{\text{test}} = T/2$
(empirically determined)

Ablation Study: Loss function & Text feature types

$\mathcal{L}_{\text{diff}}$	$\mathcal{L}_{\text{TD}}$	NTU-60 (Acc, %) 55/5 split	NTU-60 (Acc, %) 48/12 split	NTU-120 (Acc, %) 110/10 split	NTU-120 (Acc, %) 96/24 split
✓		79.87	53.03	72.44	57.65
	✓	80.90	54.36	70.73	60.95
✓	✓	**86.49**	**56.03**	**74.15**	**65.06**

Global $\mathbf{z}_g$	Local $\mathbf{z}_l$	NTU-60 (Acc, %) 55/5 split	NTU-60 (Acc, %) 48/12 split	NTU-120 (Acc, %) 110/10 split	NTU-120 (Acc, %) 96/24 split
✓		83.41	51.50	70.14	61.90
	✓	83.33	52.63	69.95	62.10
✓	✓	**86.49**	**56.03**	**74.15**	**65.06**

Experiment Results

- Ablation Study: Impact of total timesteps T

Inference timestep: $t_{\text{test}} = T/2$ (empirically determined)

Total T	NTU-60 (Acc, %)		NTU-120 (Acc, %)	
	55/5 split	48/12 split	110/10 split	96/24 split
1	85.03	44.10	69.91	60.35
10	84.51	50.89	69.97	62.04
50	**86.49**	**56.03**	**74.15**	**65.06**
100	83.48	56.27	71.05	64.57
500	81.34	53.43	71.93	60.81

- Ablation Study: Effect of noise ϵ during training

Gaussian noise ϵ	NTU-60 (Acc, %)		NTU-120 (Acc, %)	
	55/5 split	48/12 split	110/10 split	96/24 split
Fixed	76.40	44.25	64.01	52.21
Random	**86.49**	**56.03**	**74.15**	**65.06**

Regularization mechanism prevents overfitting

Bridging the Skeleton-Text Modality Gap

Diffusion-Powered Modality Alignment for Zero-shot Skeleton-based Action Recognition

For more details, please visit here

Project Page

Thank You!

KAIST EE
KAIST ELECTRICAL ENGINEERING

VICLab Video and Image Computing Lab

ICCV HONOLULU HAWAII
OCT 19-23, 2025