# SA-DVAE: Improving Zero-Shot Skeleton-Based Action Recognition by Disentangled Variational Autoencoders

**Sheng-Wei Li**, Zi-Xiang Wei, Wei-Jie Chen, Yi-Hsin Yu, Chih-Yuan Yang, Jane Yung-jen Hsu

Research Published in
European Conference on Computer Vision, 2024

# Introduction

Background and Motivation

# Motivation

Modern action recognition models (CNN[1], GCN[2]) require large amounts of data to learn effectively.

However, collecting and annotating large amounts of data can be impractical for several reasons:

- **Rarity of action classes**
- **High expense and time consumption**
- **Concerns over privacy**

This study focuses on the challenge of limited data availability in action recognition tasks, where some of the rare classes have no samples.

I.e., Zero-Shot Learning

Reference:
[1] Revisiting Skeleton-based Action Recognition.
[2] Skeleton-Based Action Recognition with Shift Graph Convolutional Network.

# Challenges in GZSL[1]

- **Seen Class Bias:** Predictions are usually biased towards the seen classes.

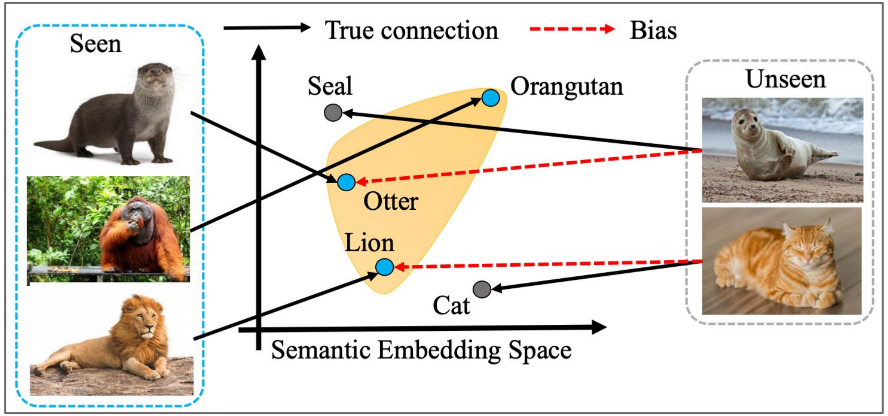



Illustration of misclassification of unseen classes into seen classes

- **Domain Shift:** The distributions of the seen and unseen classes may be different.

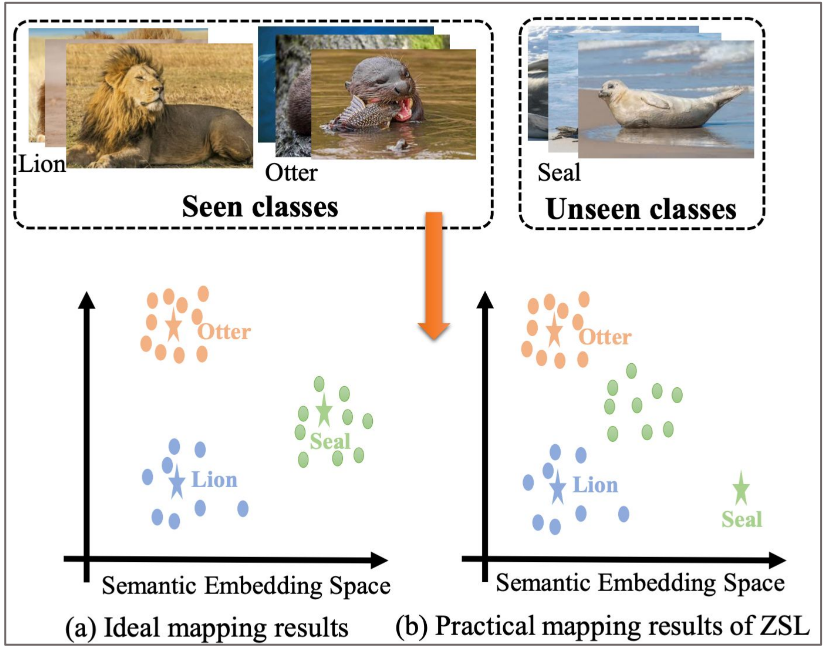


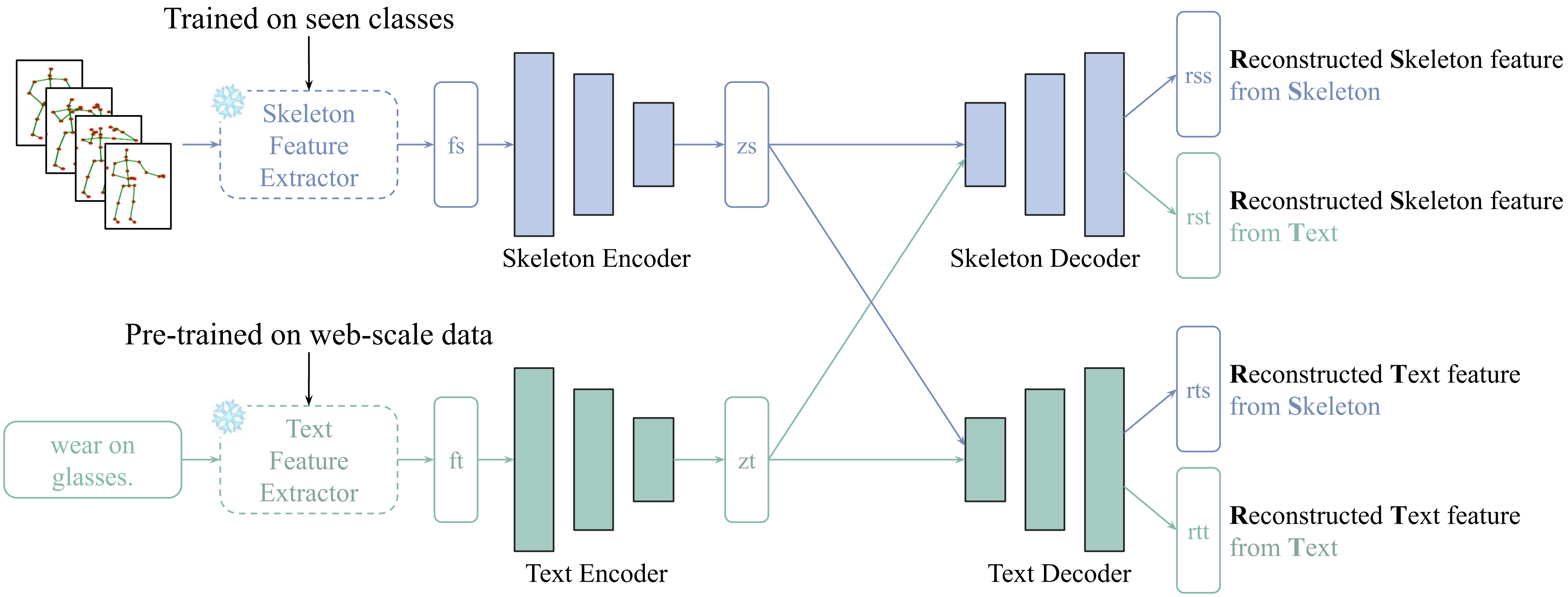

Illustration of seen/unseen domain shift.
★: The semantic embeddings, ●: The image samples


Reference:
[1] A Review of Generalized Zero-Shot Learning Methods.

# Related Work

Zero-Shot Learning on Action Recognition

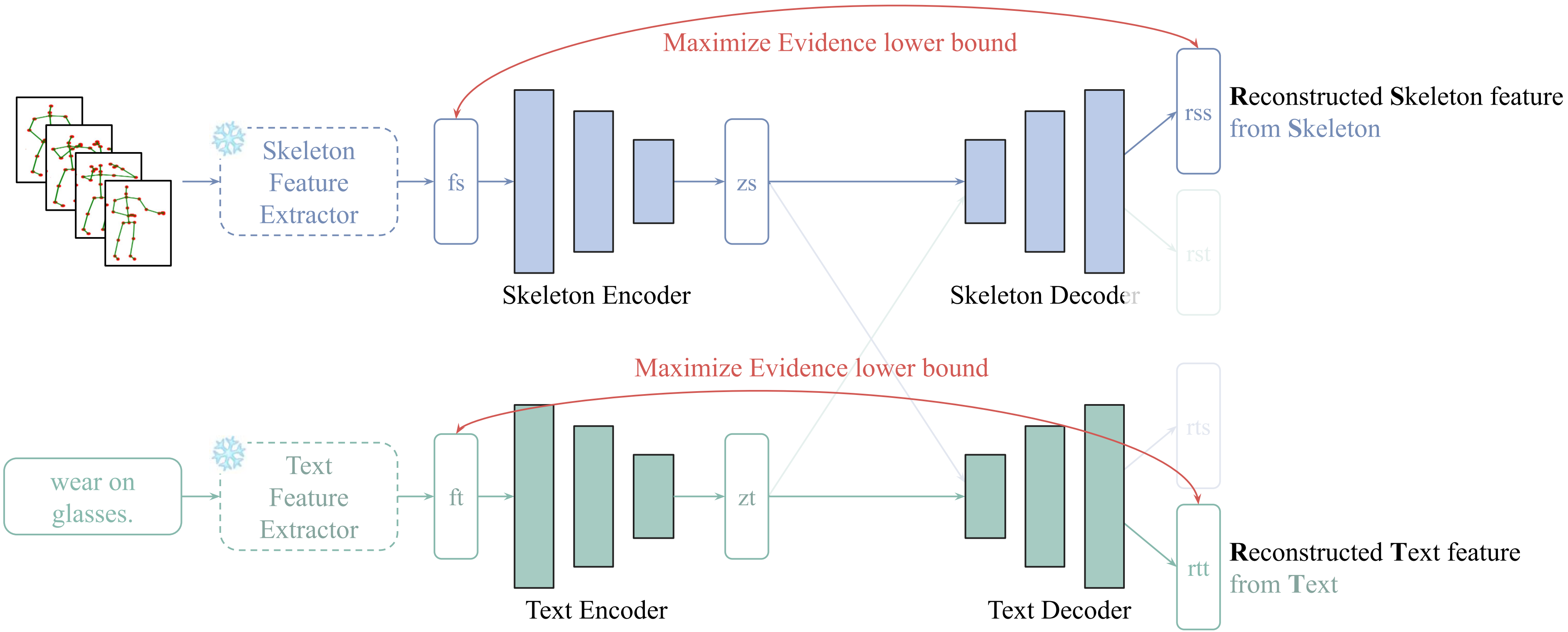
# SynSE[1]





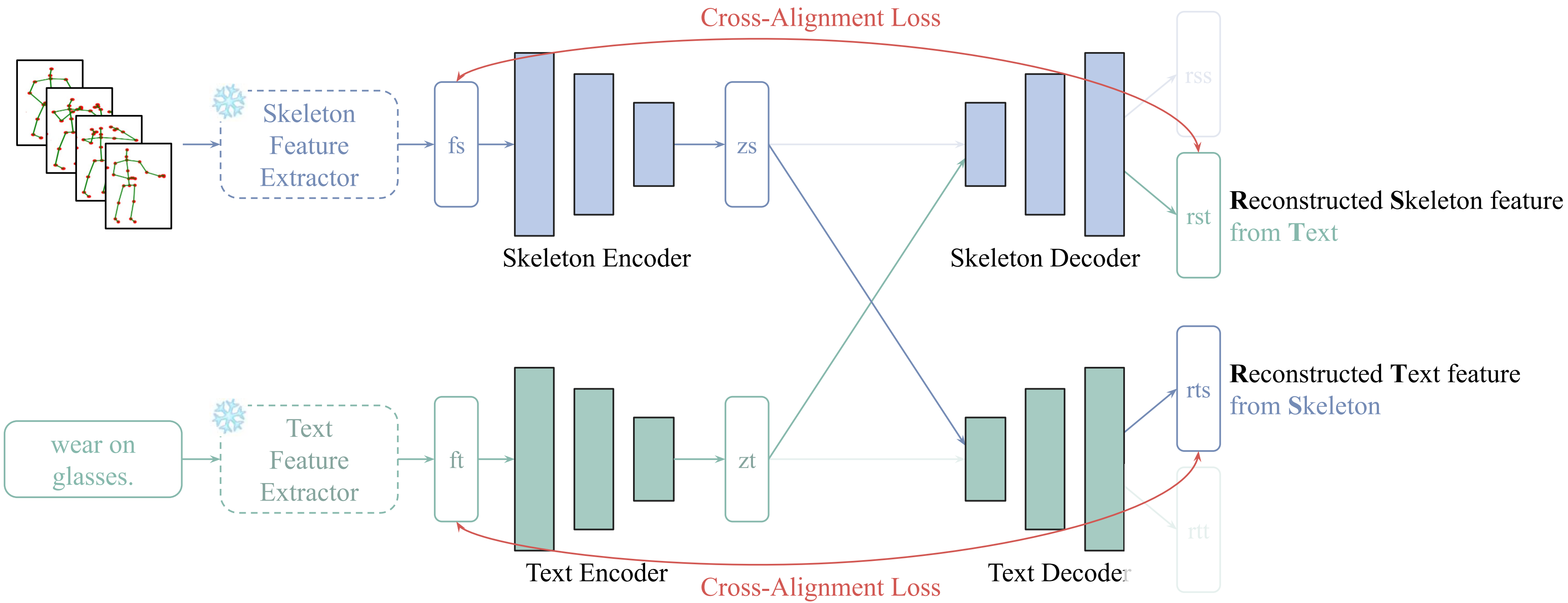
Reference:
[1] Syntactically Guided Generative Embeddings for Zero-Shot Skeleton Action Recognition.

# SynSE: Maximize ELBO

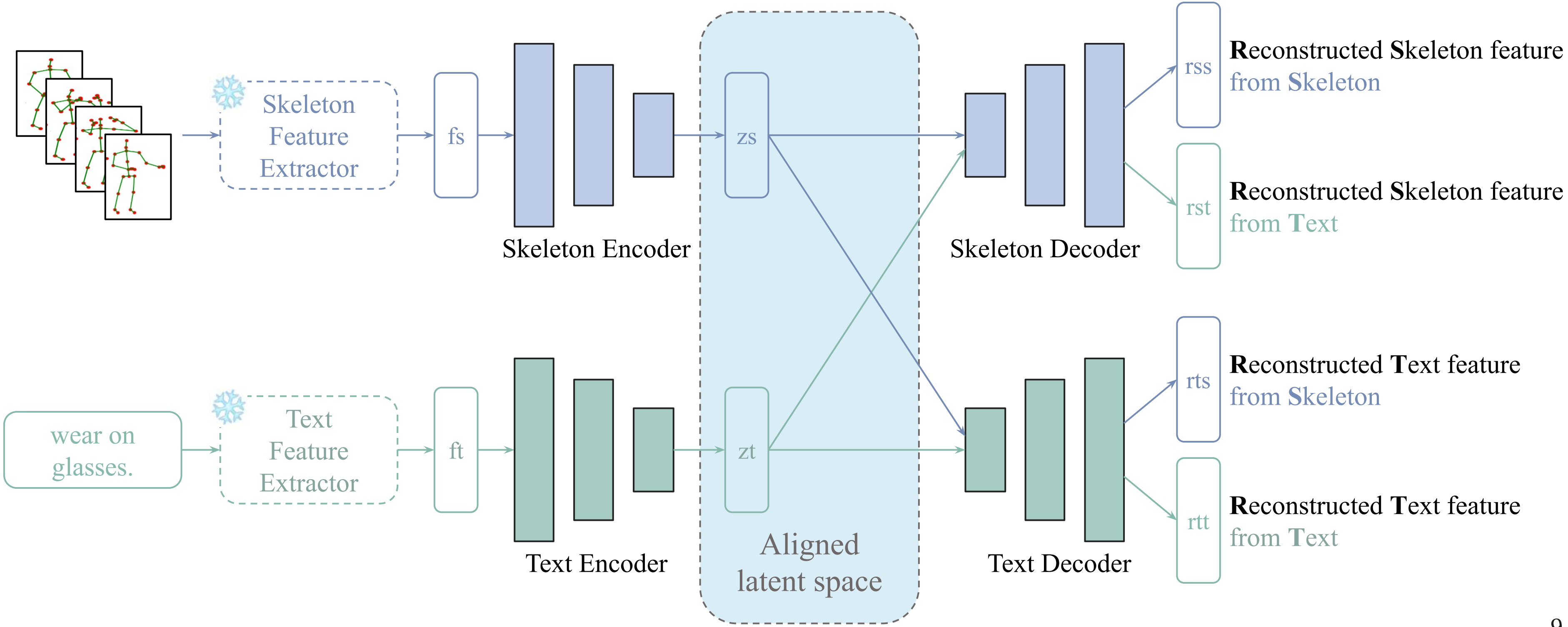



$$\mathrm{ELBO} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x) \| p_\theta(z))$$

# **SynSE:** Cross-Alignment



# SynSE: Generative Embedding Space with VAE

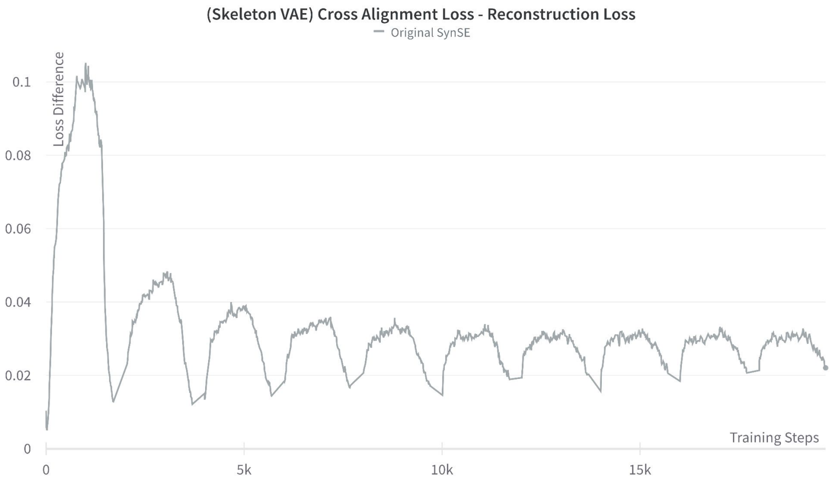


# Methodology

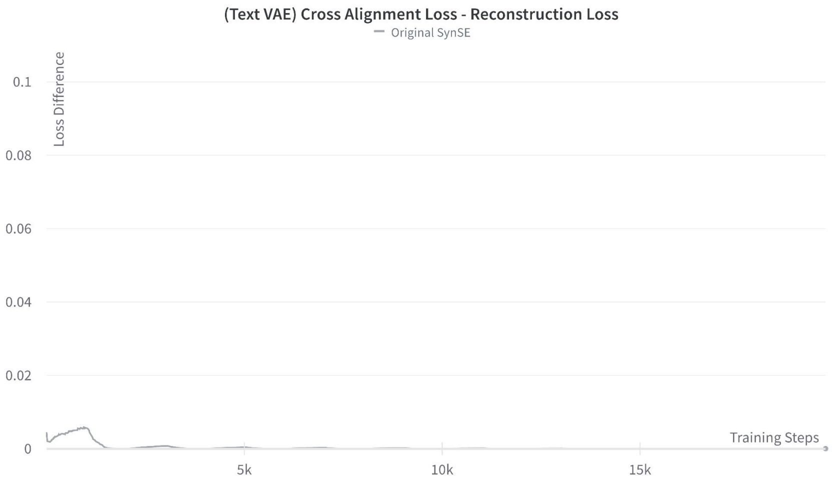
Observations and Proposed Method

# Observations about the Generative Alignment Module

While reproducing SynSE…




Loss Value of Reconstructing Skeletons from Text




Loss Value of Reconstructing Text from Skeletons

Reconstructing **skeletons from text** is much more difficult than reconstructing **text from skeletons**.

# Observations

Reconstructing **skeletons from text** is much more difficult than reconstructing **text from skeletons**.

For action recognition datasets:

- Skeletons contain both **semantic info** and **instance-specific style** e.g., person, viewpoint, etc.
- Class labels contain **only semantic info**

→ The both modalities are asymmetrical.

# Observations

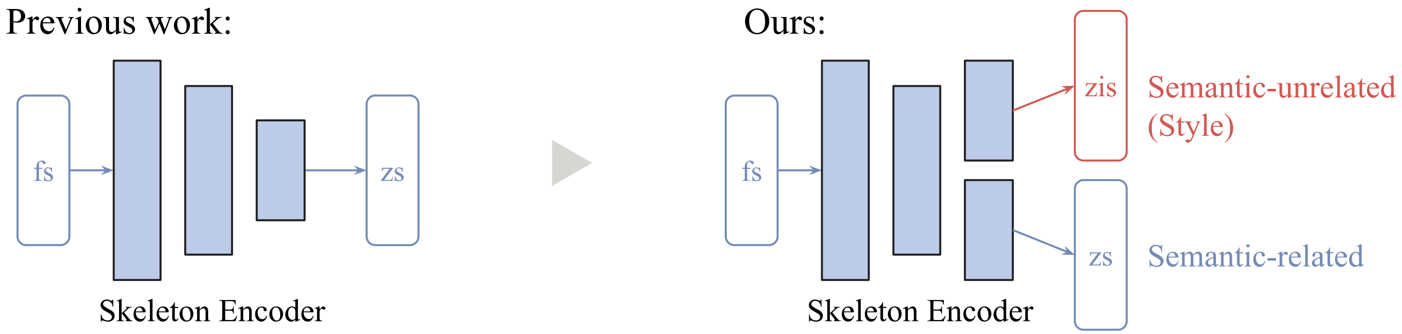
Since the both modalities are asymmetrical,

→ Design asymmetrical VAEs by applying feature disentanglement.

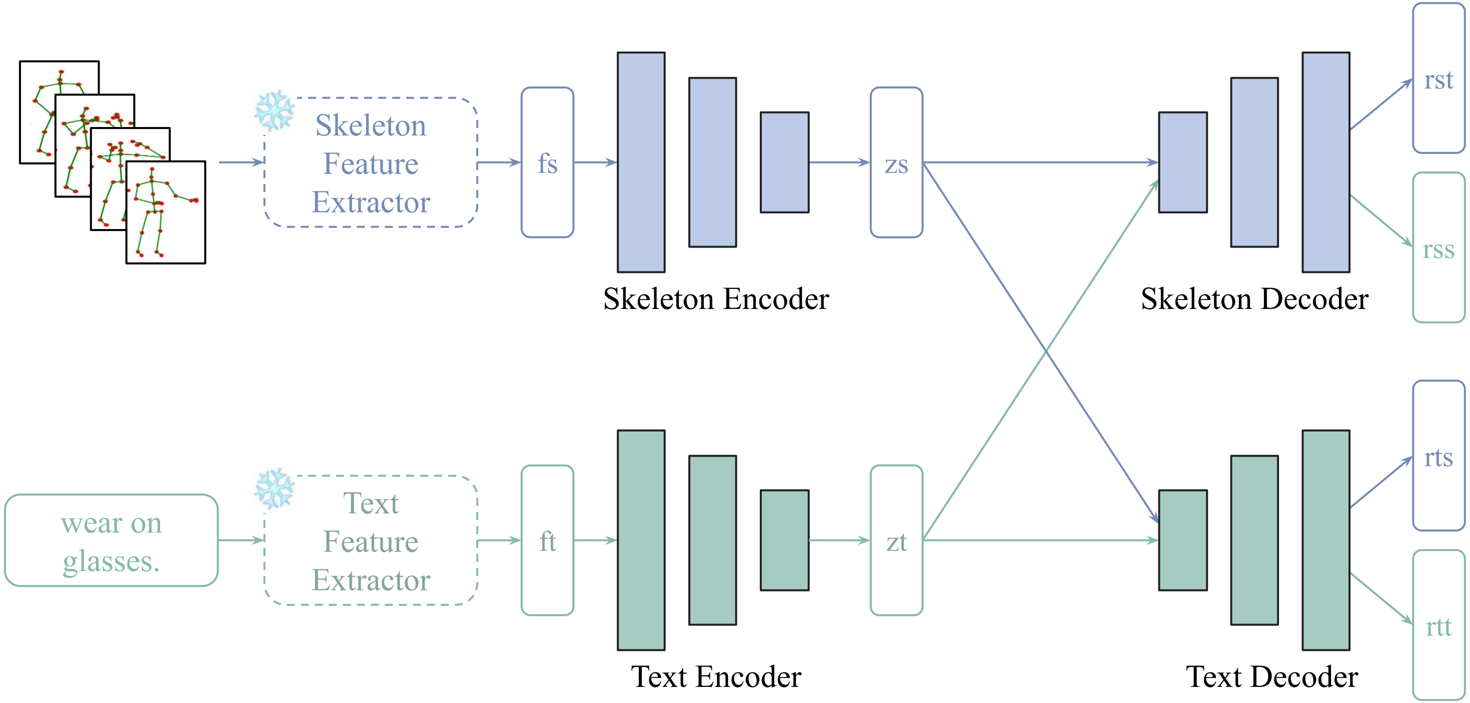
Skeleton Encoder encodes the feature into:

a. Semantic-related: Skeleton latent (*zs*)

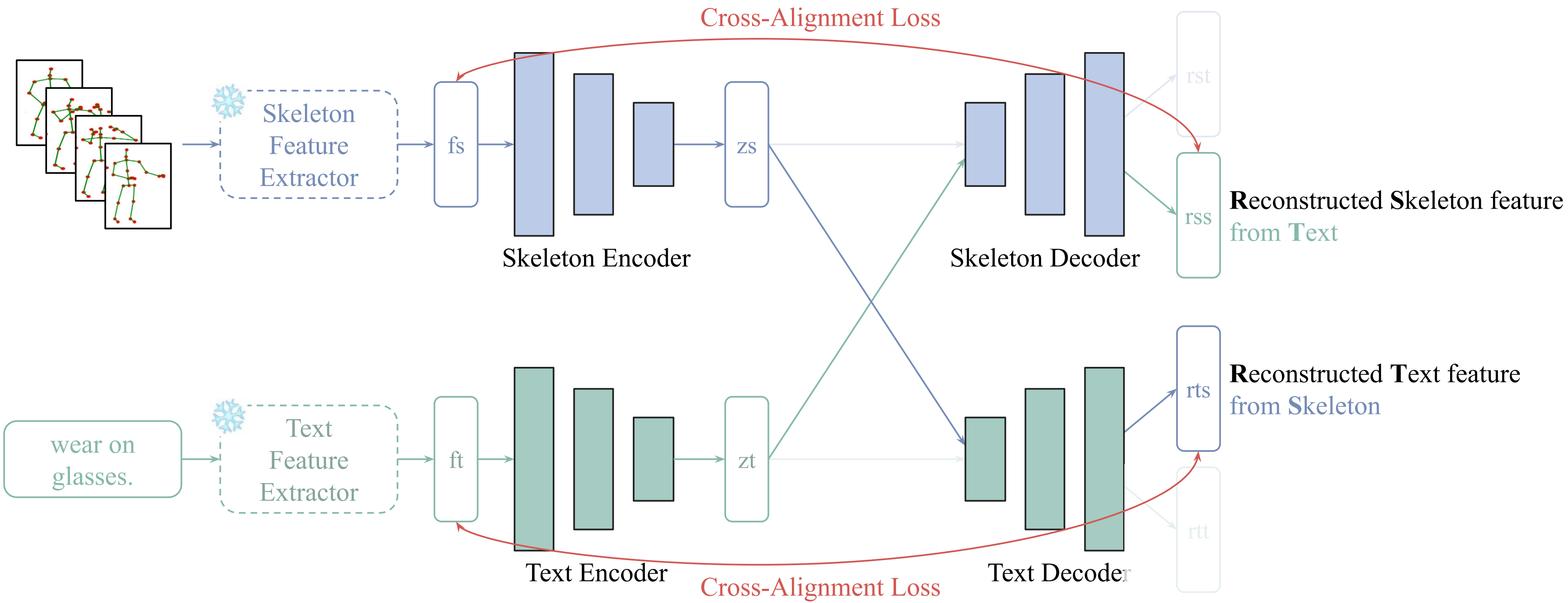
b. Semantic-unrelated: Instance style latent (*zis*)



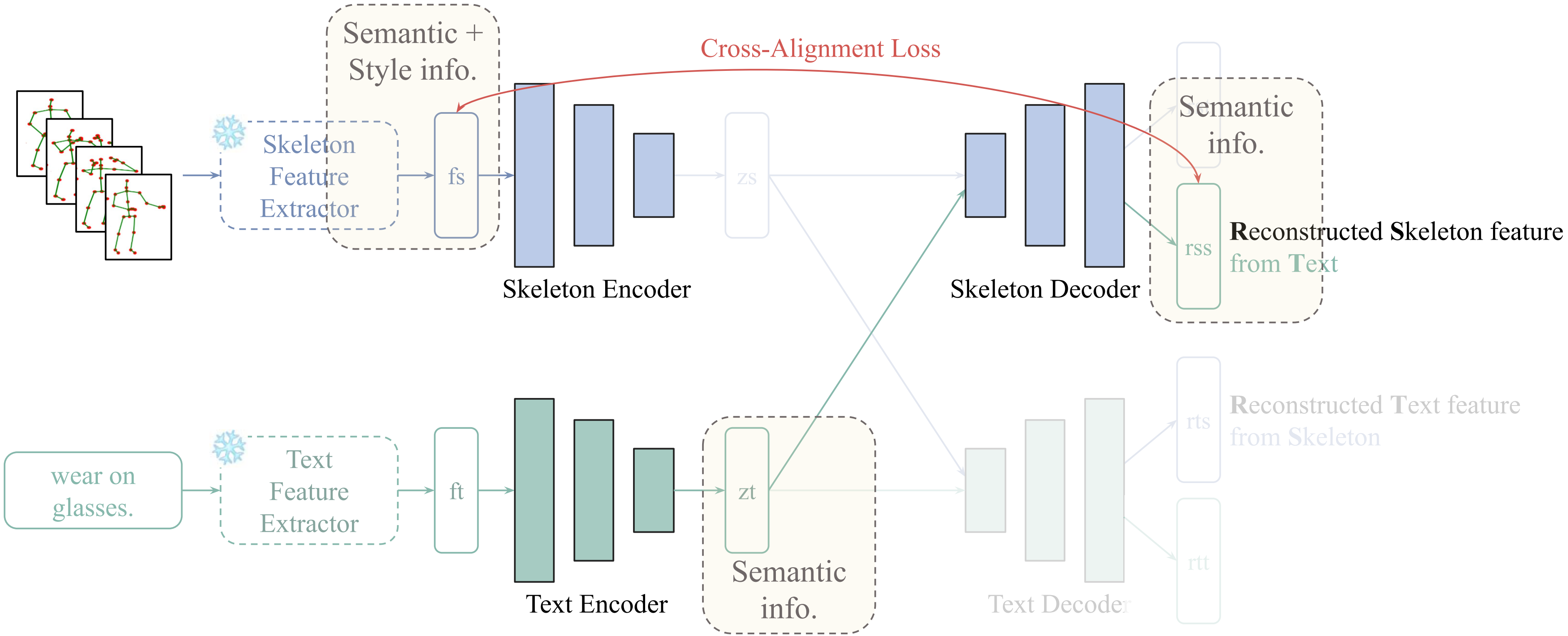
# Review SynSE



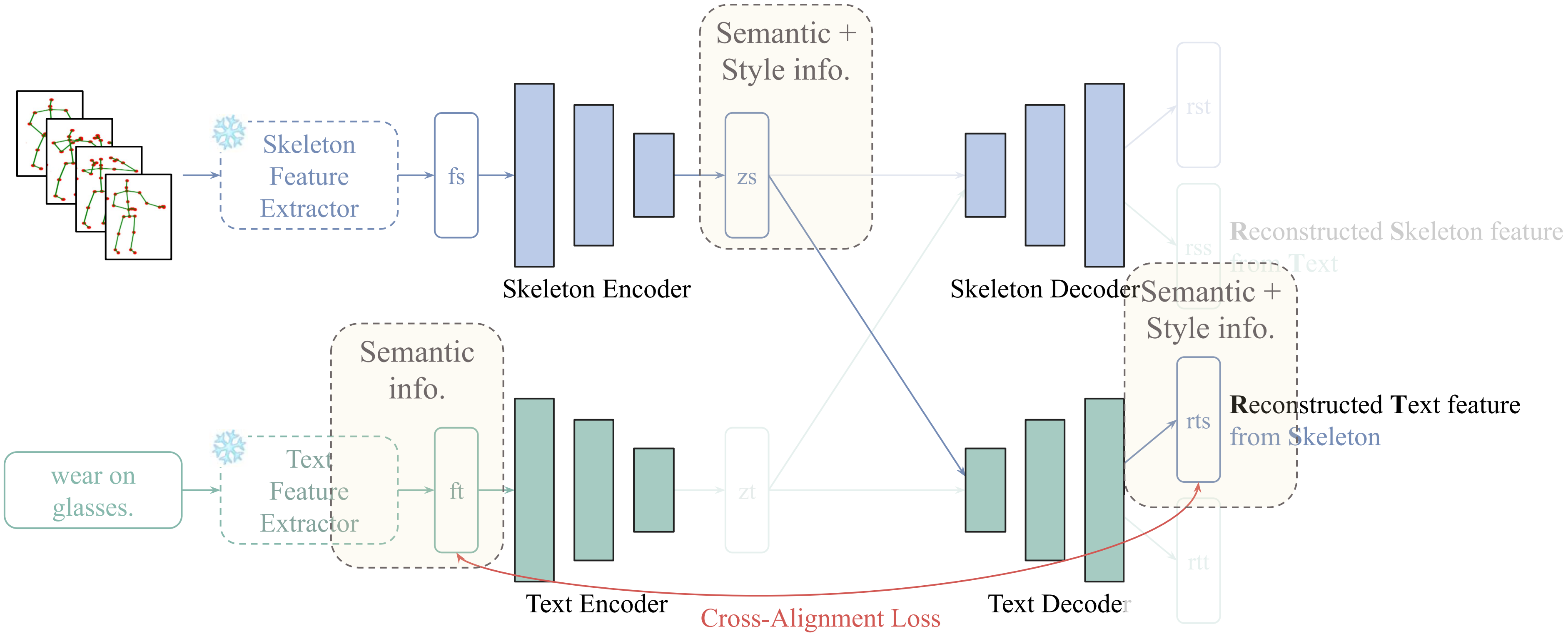
# **Review SynSE:** Cross-Alignment



# **Review SynSE:** Reconstruct Skeletons From Text

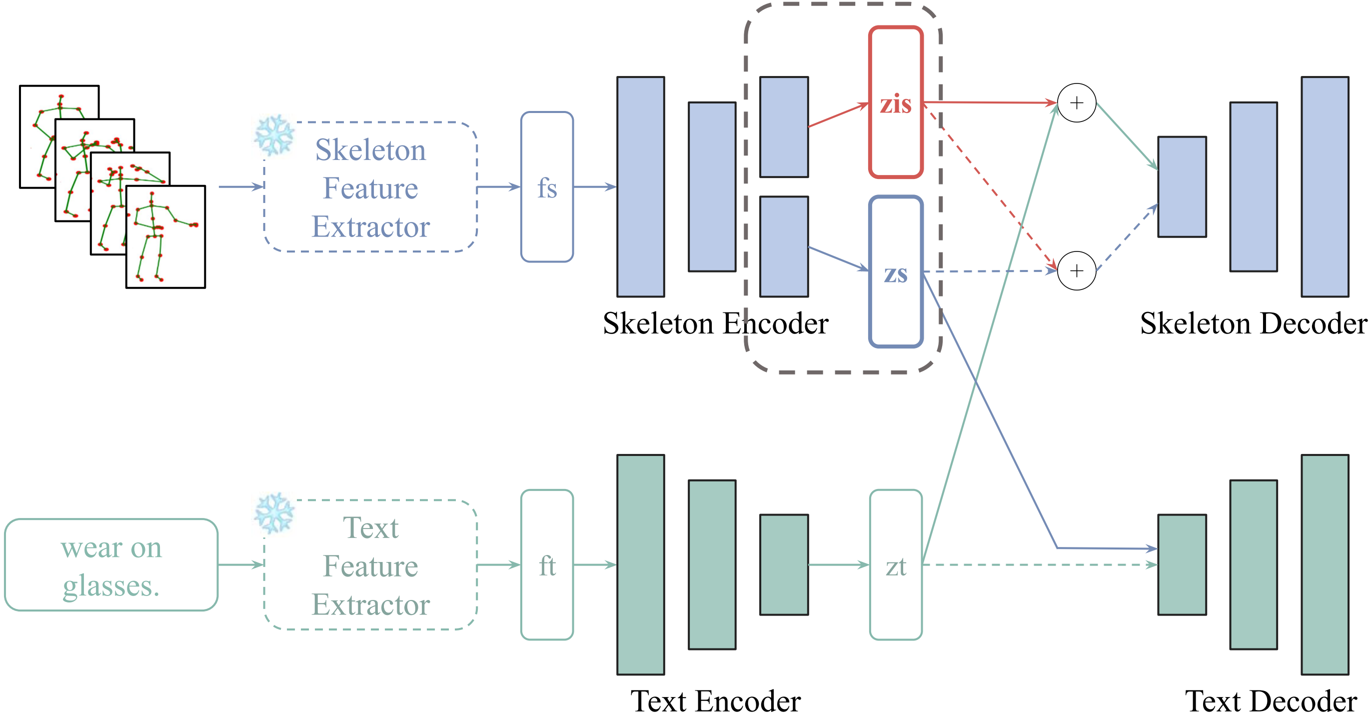


# **Review SynSE:** Reconstruct Text From Skeletons

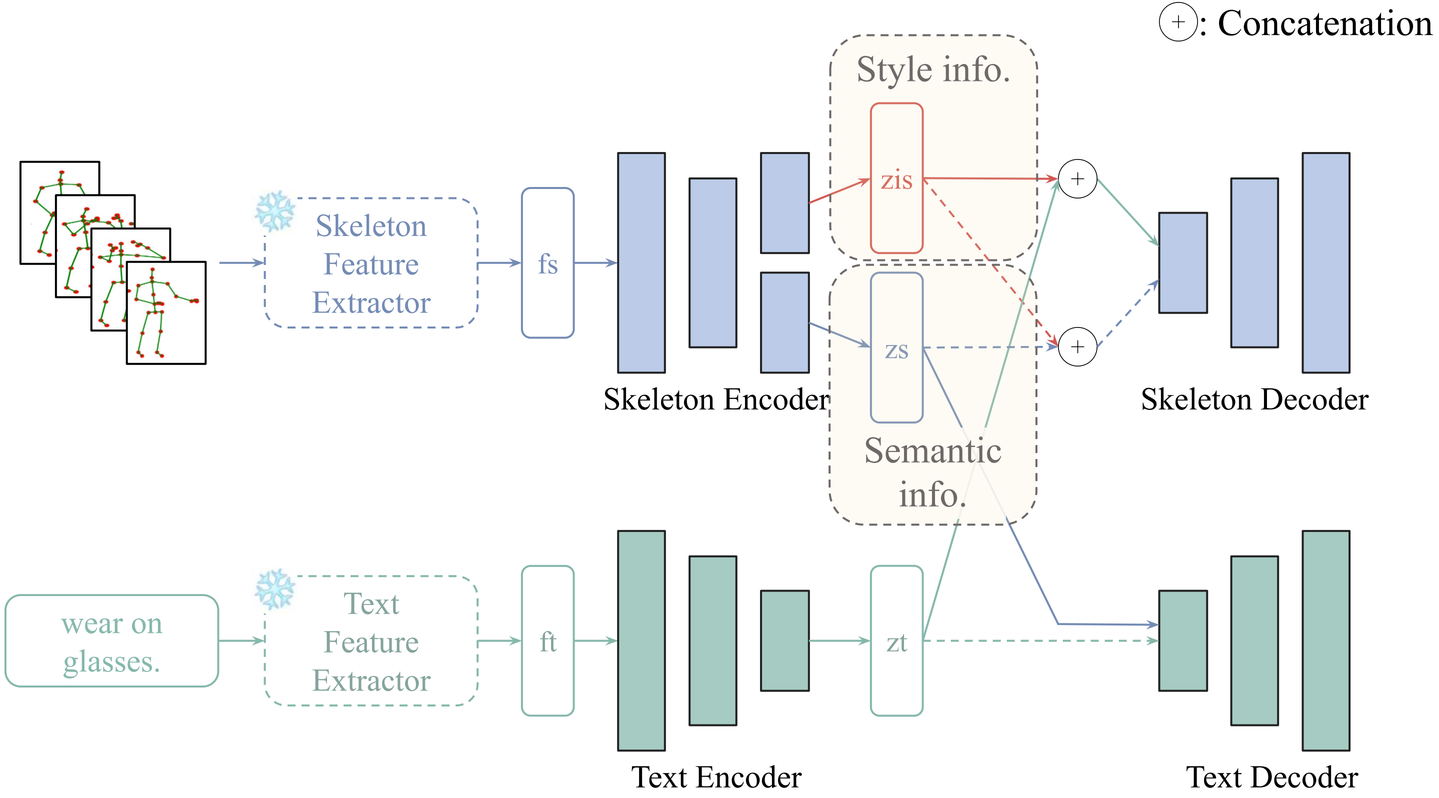


# SA-DVAE

Introduce feature disentanglement

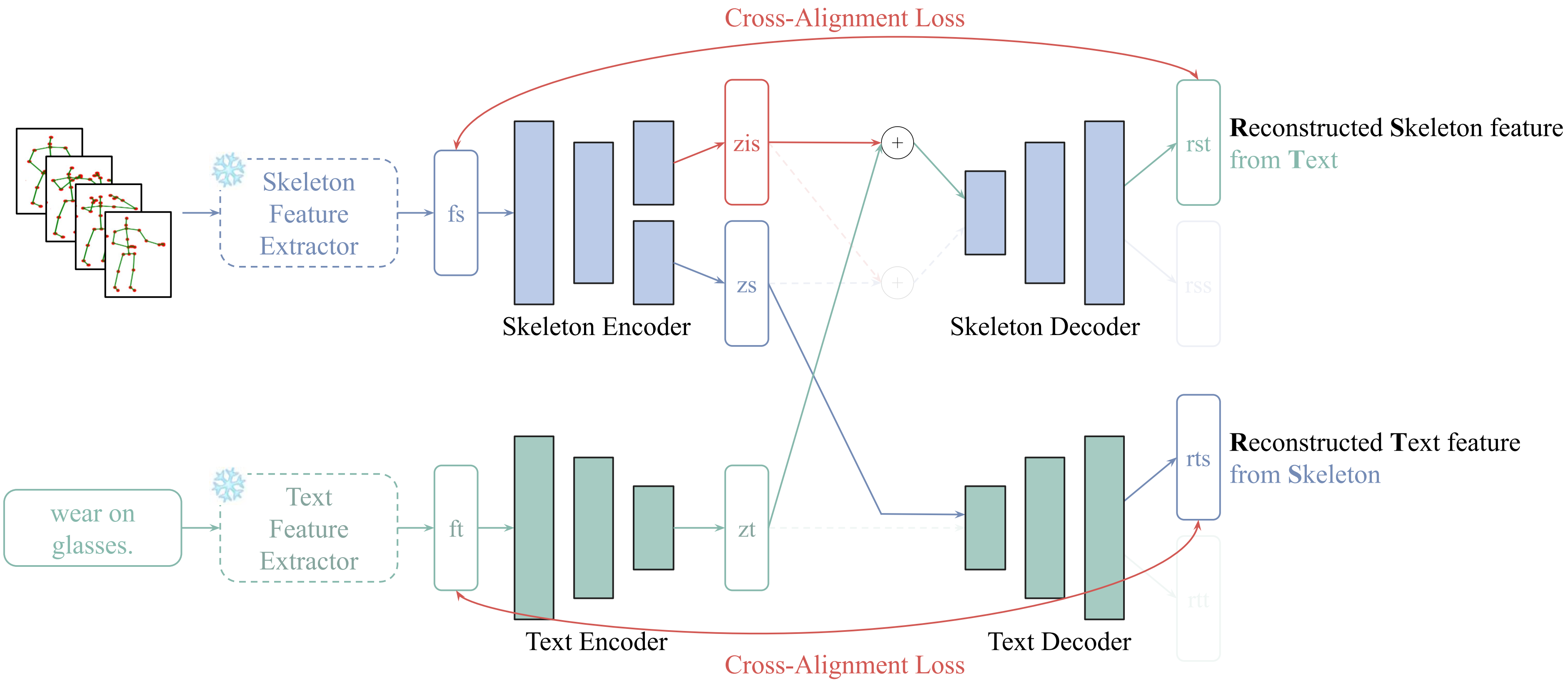


# SA-DVAE

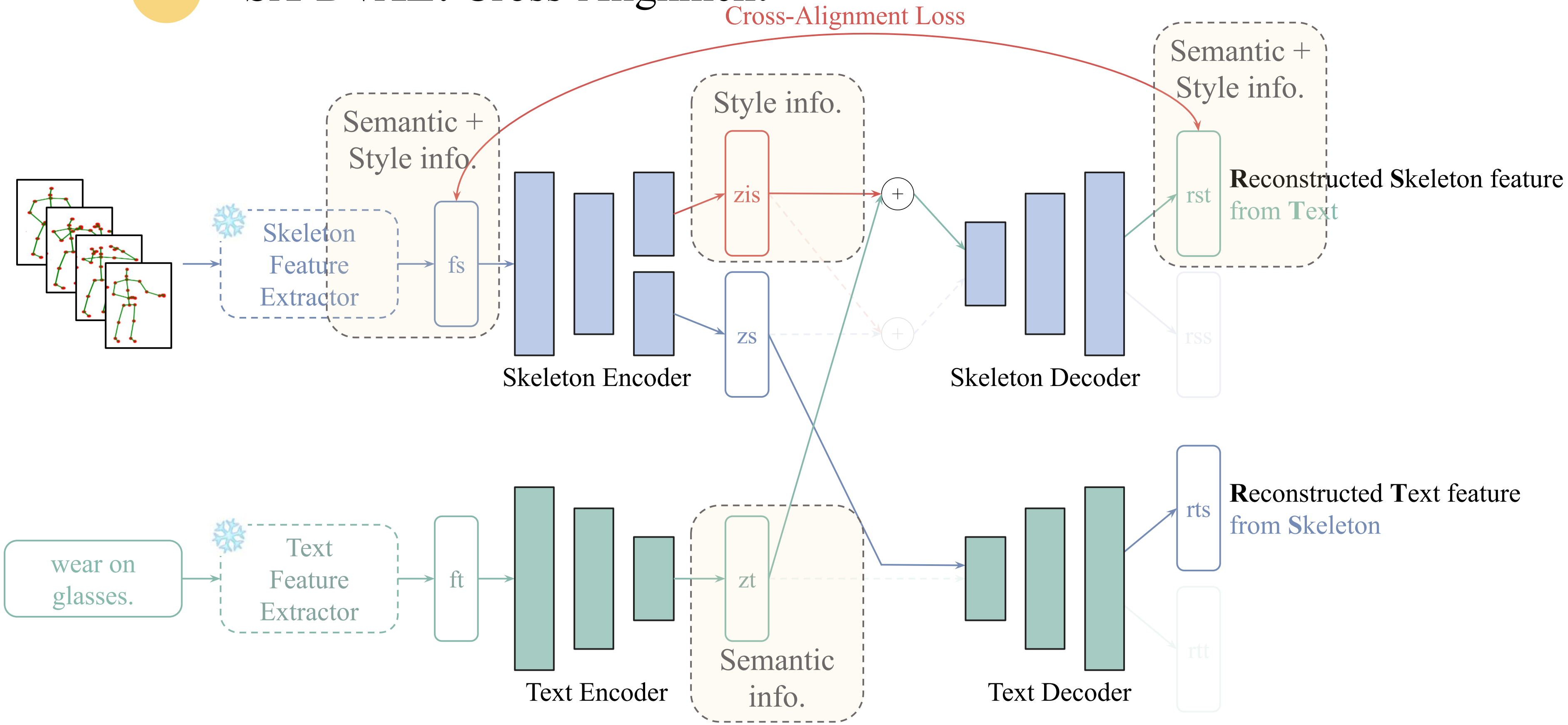


+: Concatenation
Style info.
zis
Skeleton Feature Extractor
fs
zs
Skeleton Encoder
Skeleton Decoder
Semantic info.
wear on glasses.
Text Feature Extractor
ft
zt
Text Encoder
Text Decoder

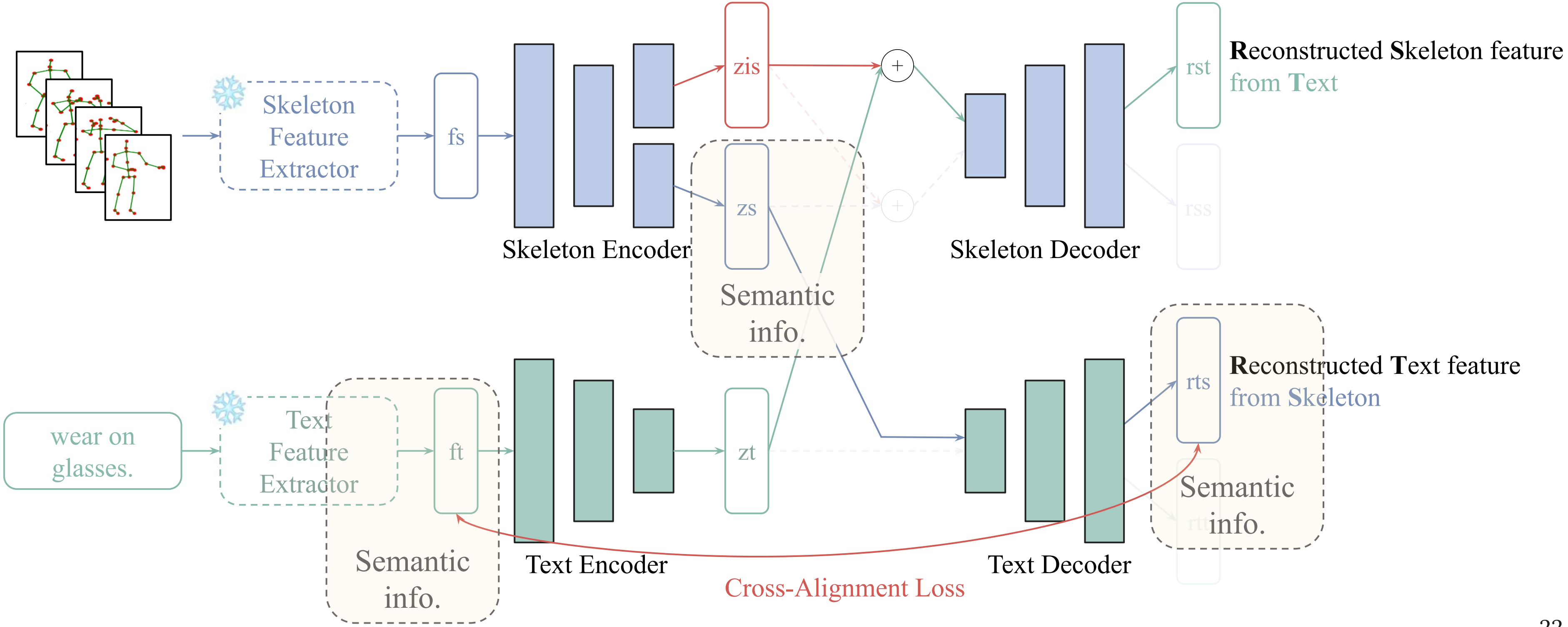
# **SA-DVAE:** Cross-Alignment



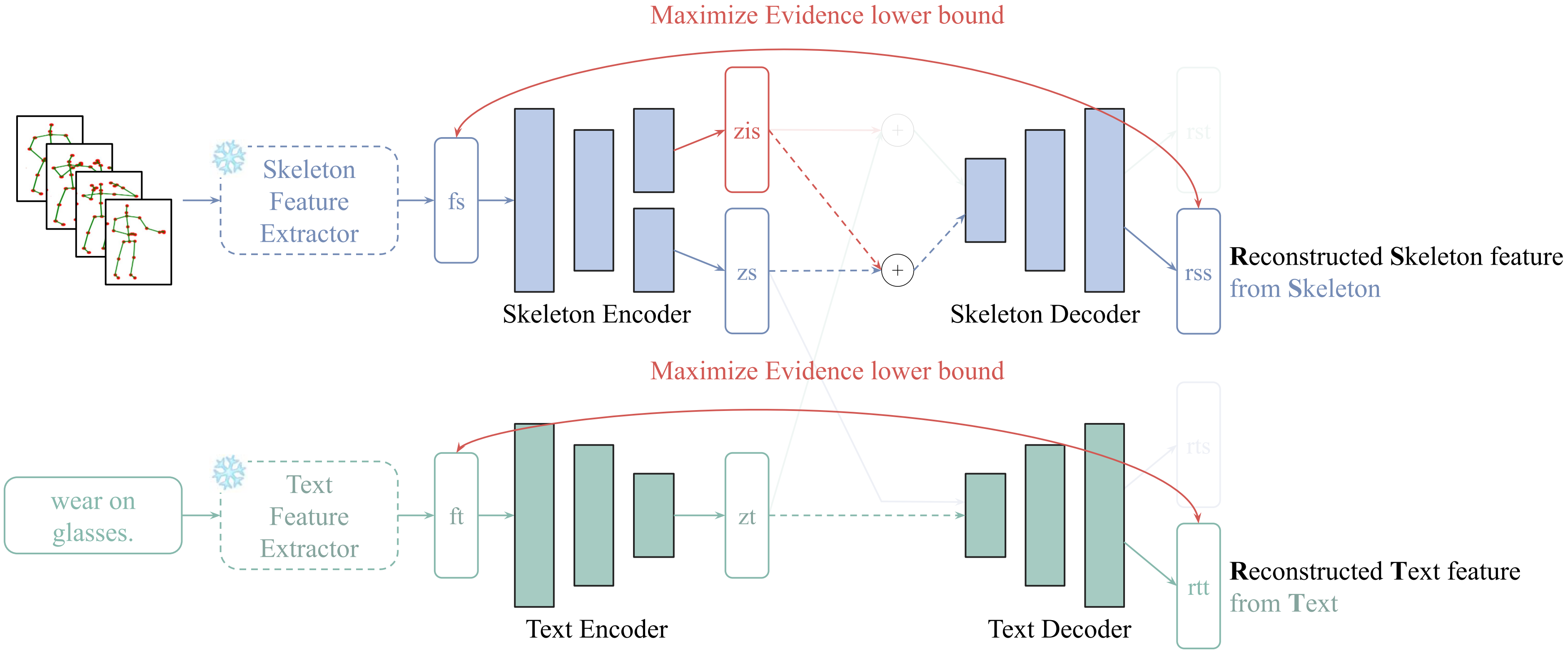
# SA-DVAE: Cross-Alignment



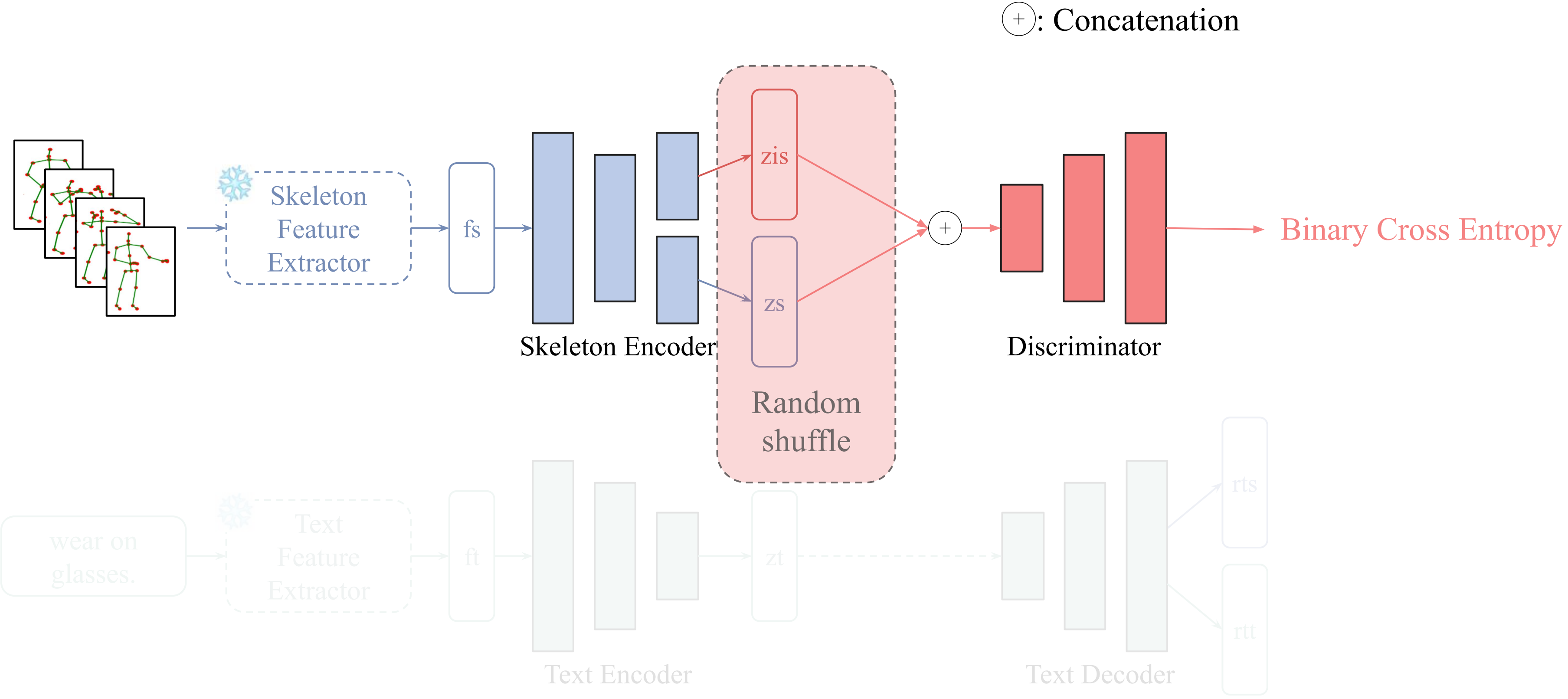
# **SA-DVAE:** Cross-Alignment



# SA-DVAE: Maximize ELBO



# SA-DVAE: Adversarial Total Correlation Penalty

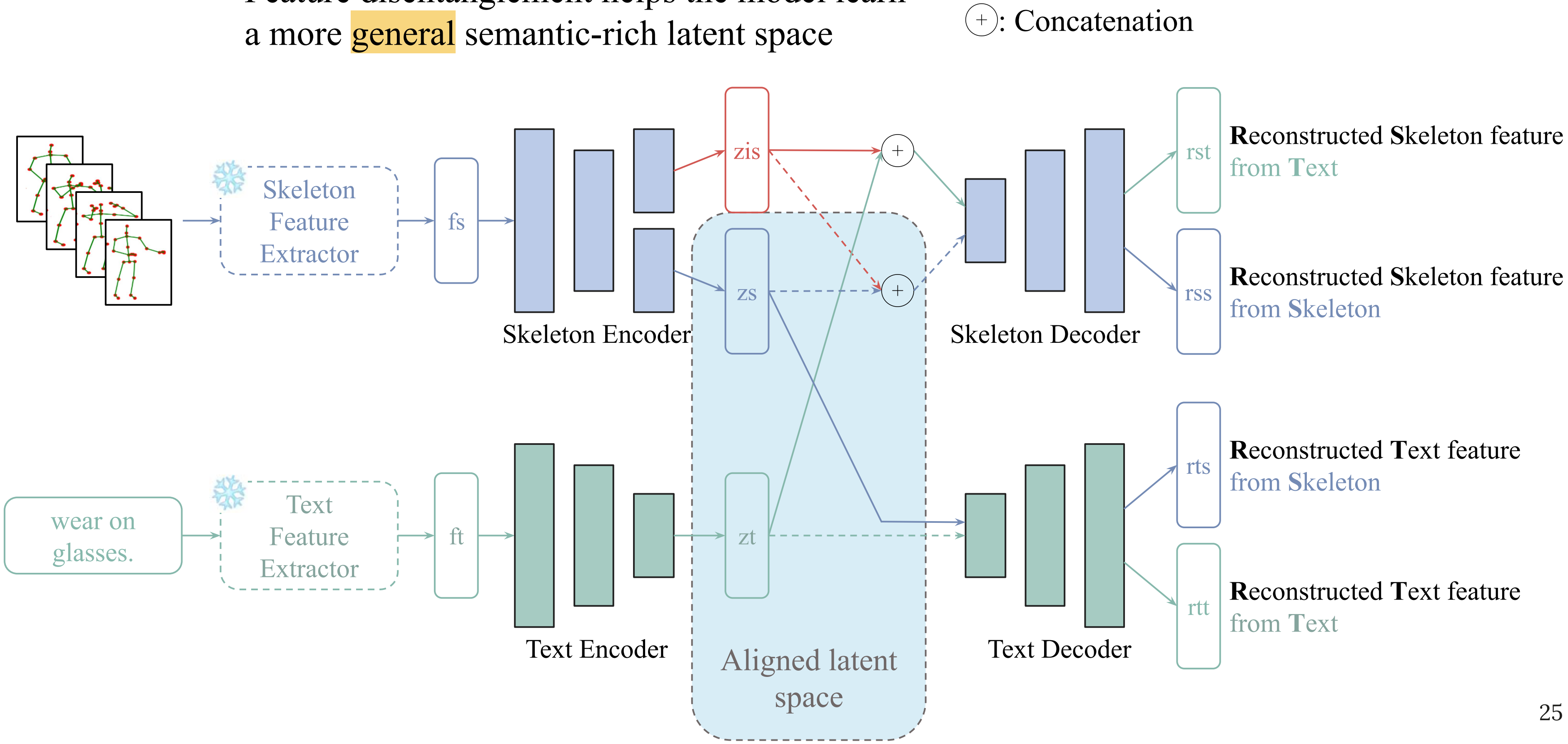



Reference:


[2] Chen et al. Semantics Disentangling for Generalized Zero-Shot Learning. ICCV 2021

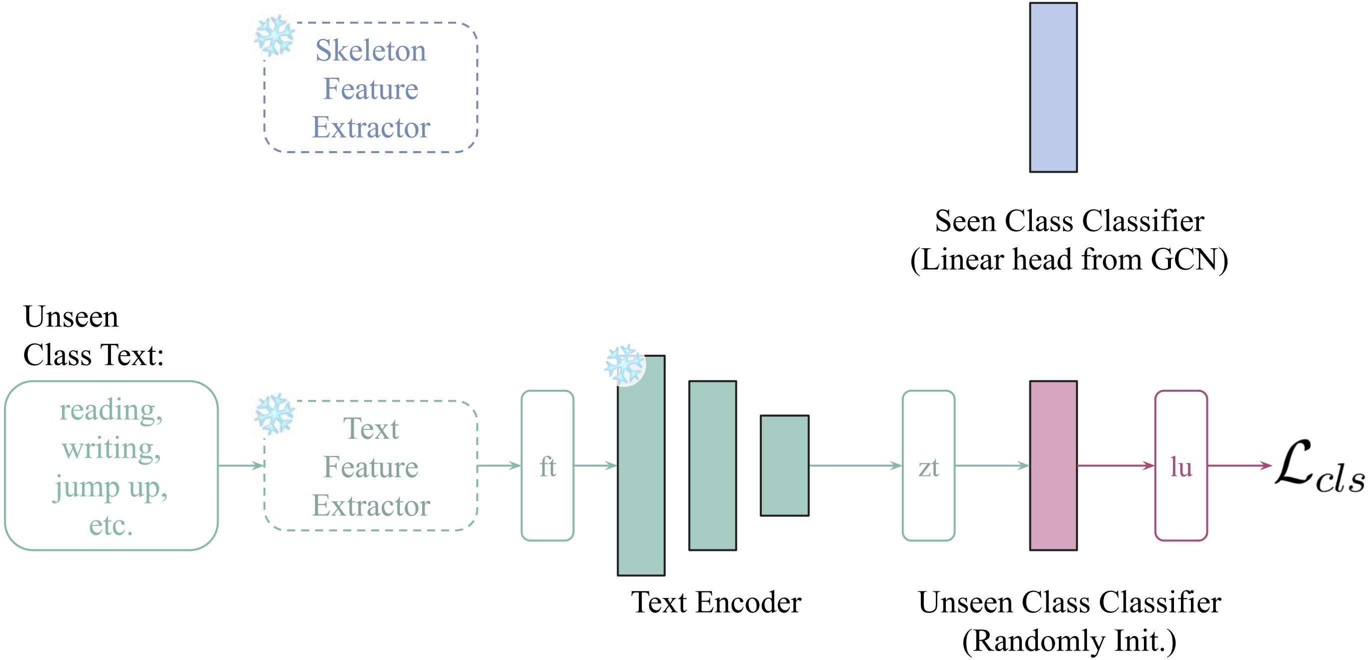
## **Stage 1:** Aligned VAEs

Feature disentanglement helps the model learn a more general semantic-rich latent space



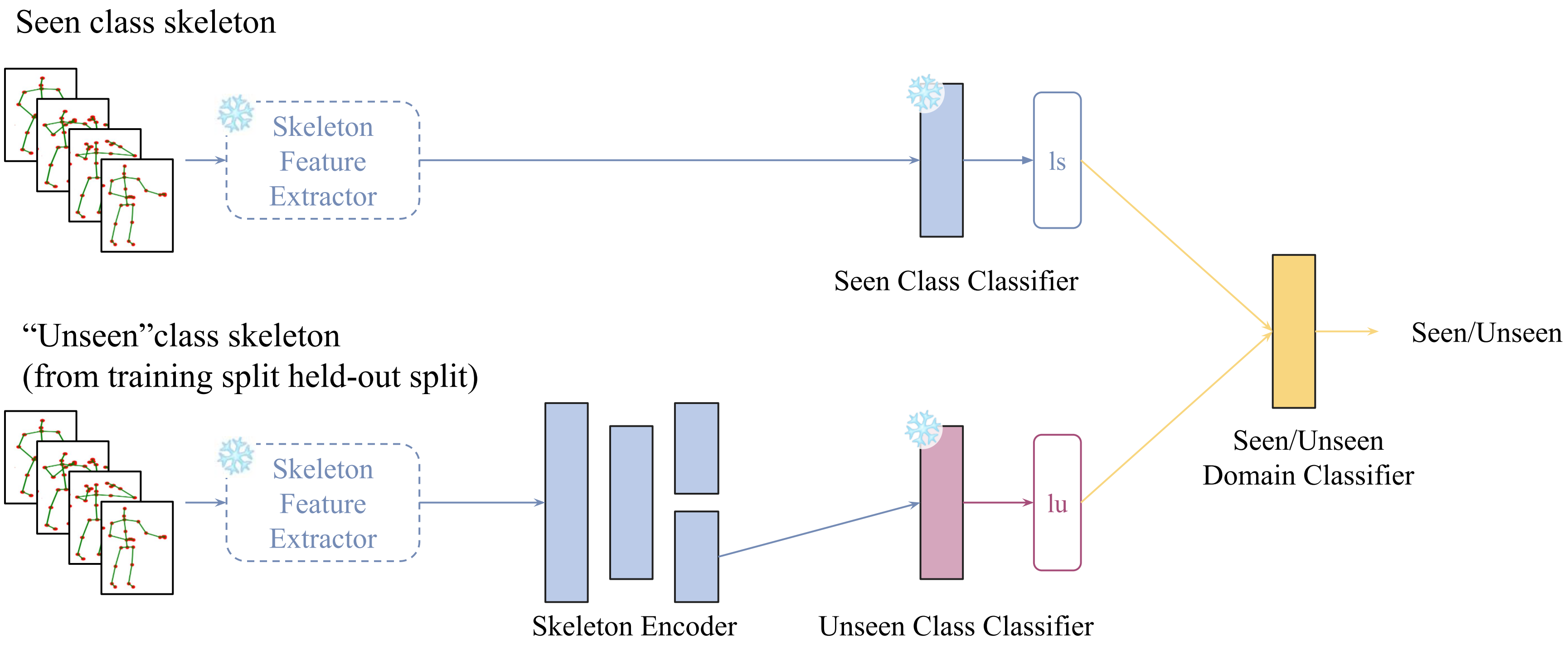
## **Stage 2:** Seen and Unseen Classifier

The unseen classifier handles unseen class predictions



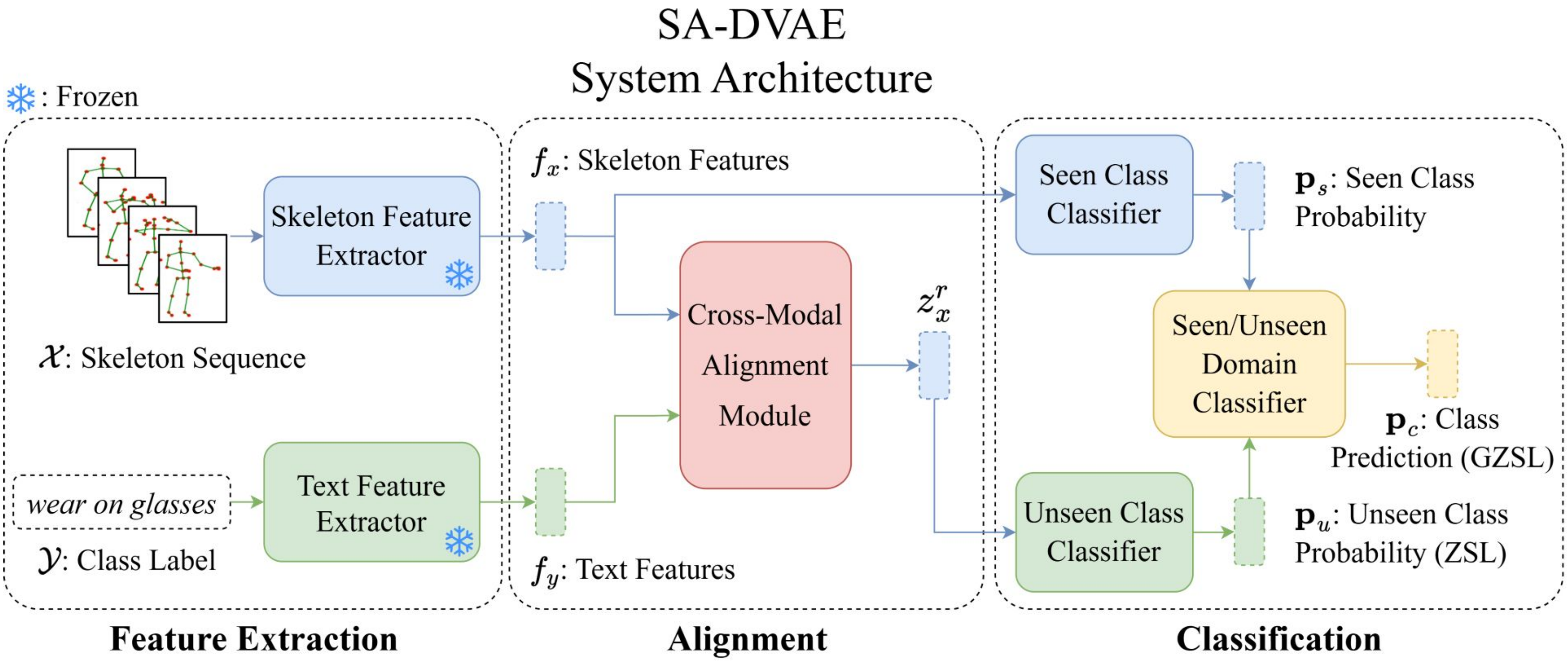
## **Stage 3:** Seen/Unseen Domain Classifier[1]

The seen/unseen domain classifier hinders the model from being biased toward seen classes.



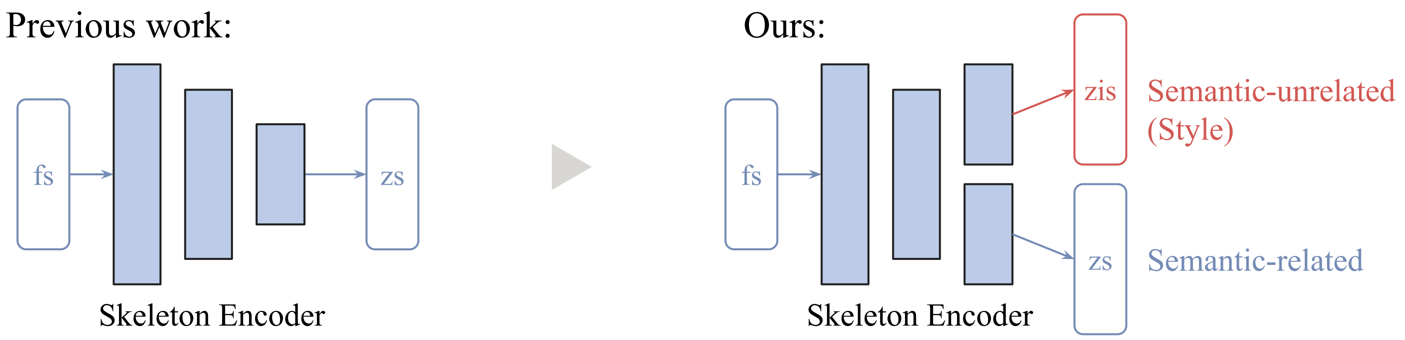

Reference:
[1] Adaptive Confidence Smoothing for Generalized Zero-Shot Learning.

# System Diagram

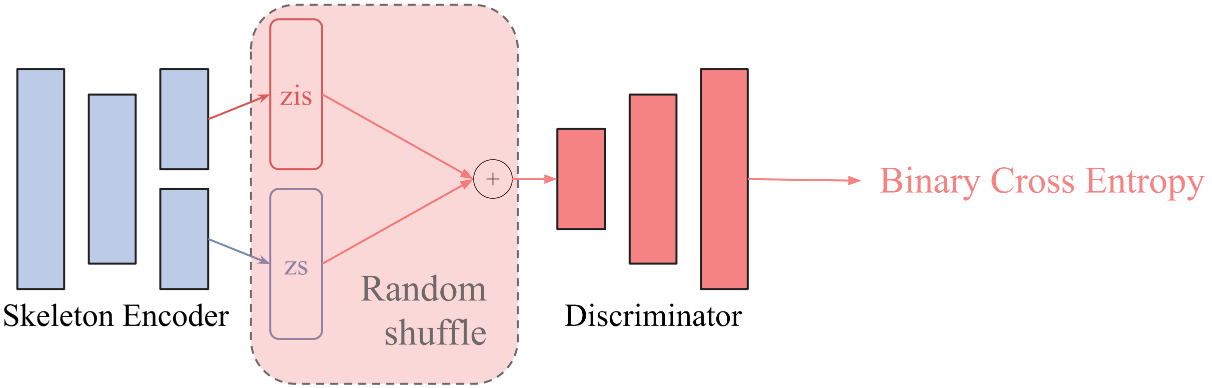


## Summary

Introduced feature disentanglement for a more generalized representation:




Encourages disentanglement by applying adversarial total correlation penalty:



# 4 Experiments

Datasets, Evaluation Protocols, and State-of-the-Arts

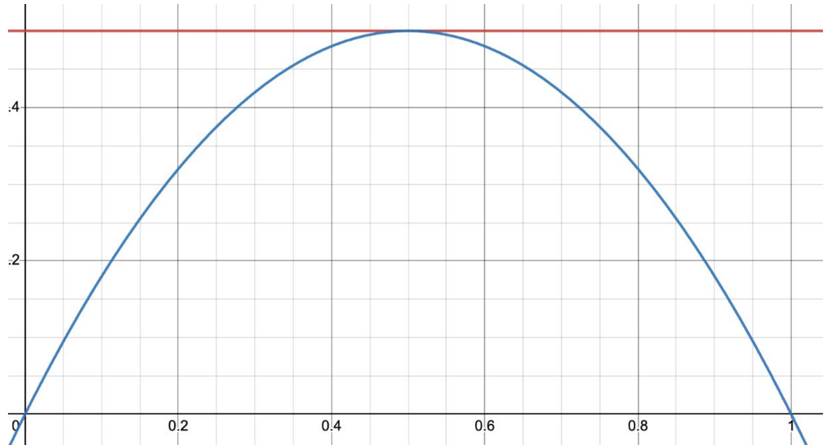
# Evaluation Protocol

**Evaluation Metric:**

ZSL: Accuracy

GZSL: Harmonic mean of *seen class accuracy* and *unseen class accuracy*.

$$H = \frac{2 \cdot seen\ accuracy \cdot unseen\ accuracy}{seen\ accuracy + unseen\ accuracy}$$




Arithmetic Mean VS Harmonic Mean of x and (1-x)

# State-of-the-Arts

Direct Mapping

- **ReViSE** (ICCV 2017):
  Uses a maximum mean discrepancy to align the embedding spaces.
- **JPoSE** (ICCV 2019):
  Performs fine-grained text-to-skeleton retrieval using PoS tags.

Generative Embedding Space

- **CADA-VAE** (CVPR 2019):
  Learns a shared latent space for both modalities via aligned VAEs.
- **SynSE** (ICIP 2021):
  Following CADA-VAE, infuses the latent space with PoS syntactic info.
- **MSF** (ICIG 2023):
  Augments the semantic text descriptions with human annotators.

Contrastive Learning

- **SMIE** (ACM MM 2023):
  Optimizes a shared multi-modal latent space using contrastive learning.

# Performance Comparison to SOTAs

Goal:

To have a system-level comparison with state-of-the-arts.

Scenario:

To have a direct compare to SynSE[1], we use the **pre-extracted** skeleton features as supplied in their codebase.

The skeleton feature extractor employed in the study is Shift-GCN[2], while the text feature extractor utilized is CLIP[3].

Reference:
[1] Syntactically Guided Generative Embeddings for Zero-Shot Skeleton Action Recognition.
[2] Skeleton-Based Action Recognition with Shift Graph Convolutional Network.
[3] Learning Transferable Visual Models From Natural Language Supervision.

# Performance Comparison to SOTAs

**Table 3:** ZSL accuracy (%) on the NTU RGB+D datasets.

| Method | NTU-60 | | NTU-120 | |
|---|---|---|---|---|
| | 55/5 split | 48/12 split | 110/10 split | 96/24 split |
| ReViSE [13] | 53.91 | 17.49 | 55.04 | 32.38 |
| JPoSE [25] | 64.82 | 28.75 | 51.93 | 32.44 |
| CADA-VAE [22] | 76.84 | 28.96 | 59.53 | 35.77 |
| SynSE [8] | 75.81 | 33.30 | 62.69 | 38.70 |
| SMIE [29] | 77.98 | 40.18 | 65.74 | 45.30 |
| SA-DVAE | **82.37** | **41.38** | **68.77** | **46.12** |

# Performance Comparison to SOTAs

**Table 4:** GZSL metrics: seen class accuracy $Acc_s$, unseen class accuracy $Acc_u$, and their harmonic mean $H$ (%) on the NTU RGB+D datasets. *: SynSE paper reports 29.22, but it is a miscalculation.

| Method | NTU-60 | | | | | | NTU-120 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 55/5 split | | | 48/12 split | | | 110/10 split | | | 96/24 split | | |
| | $Acc_s$ | $Acc_u$ | $H$ | $Acc_s$ | $Acc_u$ | $H$ | $Acc_s$ | $Acc_u$ | $H$ | $Acc_s$ | $Acc_u$ | $H$ |
| ReViSE [13] | 74.22 | 34.73 | 47.32* | 62.36 | 20.77 | 31.16 | 48.69 | 44.84 | 46.68 | 49.66 | 25.06 | 33.31 |
| JPoSE [25] | 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 [22] | 69.38 | 61.79 | 65.37 | 51.32 | 27.03 | 35.41 | 47.16 | 49.78 | 48.44 | 41.11 | 34.14 | 37.31 |
| SynSE [8] | 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 | 62.28 | 70.80 | **66.27** | 50.20 | 36.94 | **42.56** | 61.10 | 59.75 | **60.42** | 58.82 | 35.79 | **44.50** |

# Performance Comparison to SOTAs

We generate 3 random sets of unseen classes and report the average performance.

Ablations:

- Naive Alignment: Disables the style head.
- +FD: Enable feature disentanglement to the skeleton VAE.
- SA-DVAE (+FD +TC): Combined FD with total correlation penalty.

# ZSL Performance

| Method | NTU-60 55/5 split | NTU-120 110/10 split | PKU-MMD 46/5 split |
|---|---|---|---|
| ReViSE [13] | 60.94 | 44.90 | 59.34 |
| JPoSE† [25] | 59.44 | 46.69 | 57.17 |
| CADA-VAE [22] | 61.84 | 45.15 | 60.74 |
| SynSE† [8] | 64.19 | 47.28 | 53.85 |
| SMIE [29] | 65.08 | 46.40 | 60.83 |
| Naive alignment | 69.26 | 39.73 | 60.13 |
| FD | 82.21 | 49.18 | 60.97 |
| SA-DVAE (FD+TC) | **84.20** | **50.67** | **66.54** |

Ablations:

- Naive Alignment: Disables the "style" head.
- +FD: Enable feature disentanglement to the skeleton VAE.
- SA-DVAE (+FD +TC): Combined FD with total correlation penalty.

## GZSL Performance

| Method | NTU-60 55/5 splits | | | NTU-120 110/10 split | | | PKU-MMD 46/5 split | | |
|---|---|---|---|---|---|---|---|---|---|
| | $Acc_s$ | $Acc_u$ | $H$ | $Acc_s$ | $Acc_u$ | $H$ | $Acc_s$ | $Acc_u$ | $H$ |
| ReViSE [13] | 71.75 | 52.06 | 60.34 | 48.29 | 34.64 | 40.34 | 60.89 | 42.16 | 49.82 |
| JPoSE † [25] | 66.25 | 54.92 | 60.05 | 49.43 | 39.14 | 43.69 | 60.26 | 45.18 | 51.64 |
| CADA-VAE [22] | 77.35 | 58.14 | 66.38 | 51.09 | 41.24 | 45.64 | 63.17 | 35.86 | 45.75 |
| SynSE † [8] | 75.84 | 60.77 | 67.47 | 41.73 | 45.36 | 43.47 | 63.09 | 40.69 | 49.47 |
| Naive alignment | 82.11 | 47.99 | 60.58 | 57.01 | 31.62 | 40.68 | 58.76 | 43.14 | 49.75 |
| FD | 82.31 | 61.98 | 70.71 | 58.57 | 37.83 | 45.97 | 58.11 | 48.15 | 52.66 |
| SA-DVAE (FD+TC) | 78.16 | 72.60 | **75.27** | 58.09 | 40.23 | **47.54** | 58.49 | 51.40 | **54.72** |

Ablations:

- Naive Alignment: Disables the “style” head.
- +FD: Enable feature disentanglement to the skeleton VAE.
- SA-DVAE (+FD +TC): Combined FD with total correlation penalty.

# Conclusion

Summary, Contributions, and Future Work

# Conclusion

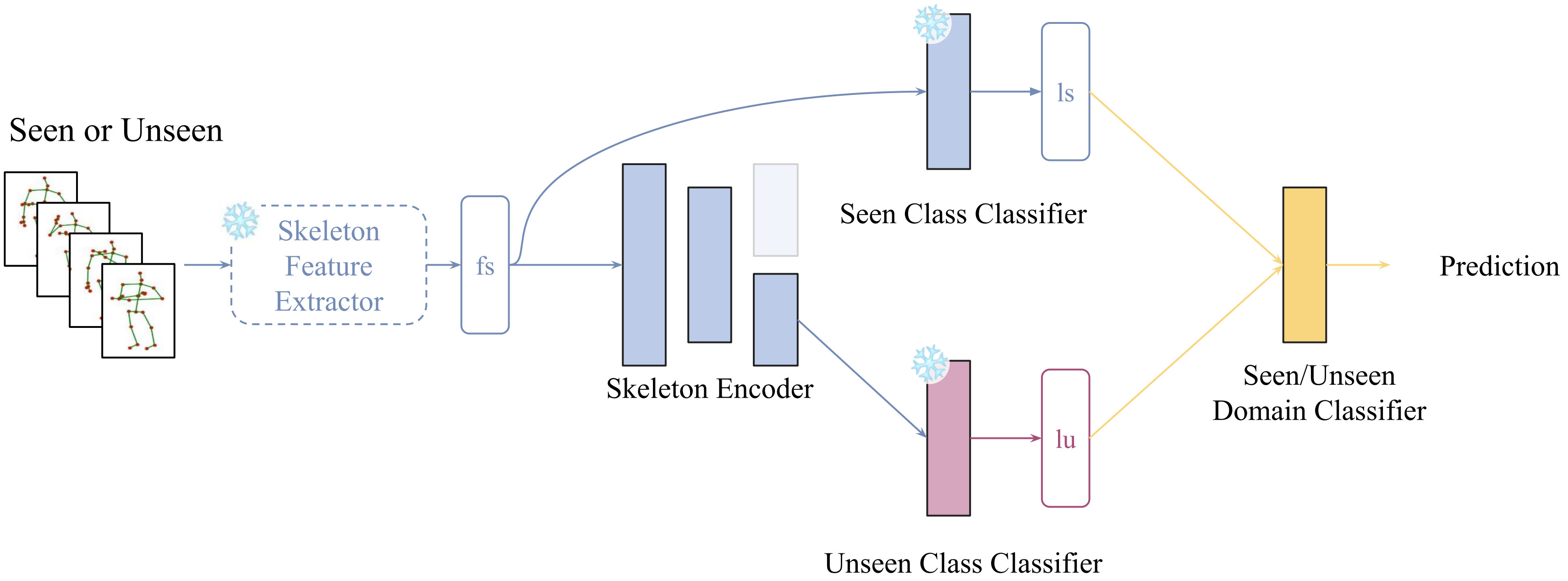
Learning a generalized representation from only seen classes persists as a challenge.

## Contributions:

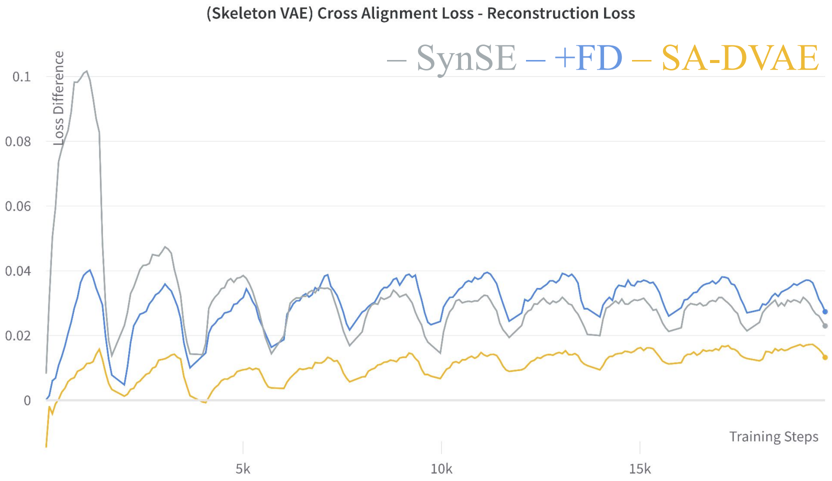
- Proposed a new method, SA-DVAE, to address the asymmetry in action recognition datasets and improve generalizability of the model.
- We show through experiments that our proposed feature disentanglement and adversarial total correlation penalty are effective on different datasets, class labels, and feature extractors.
- Sets new benchmarks for the NTU-60, NTU-120, and PKU-51 datasets.

# Appendix

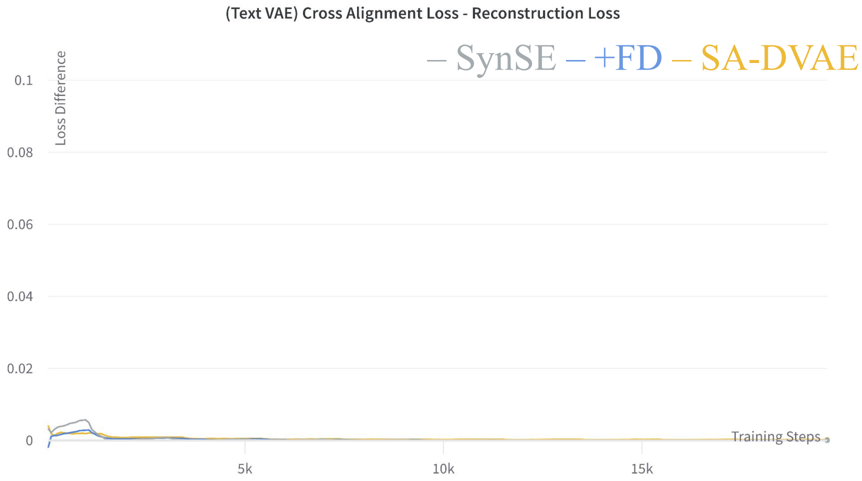
# Inference Dataflow



# Cross-Alignment Difficulty




Loss Value of Reconstructing Skeletons from Text




Loss Value of Reconstructing Text from Skeletons

+FD and SA-DVAE actually makes cross-reconstruction from text to skeleton much easier when compared to SynSE.