The 1st-place solution for CVPR 2025 AVA Challenge: 3D Human Motion Generation Track

Team **MR-CAS**

Yang Liu, Qianqian Xu, Peisong Wen, Siran Dai, Qingming Huang

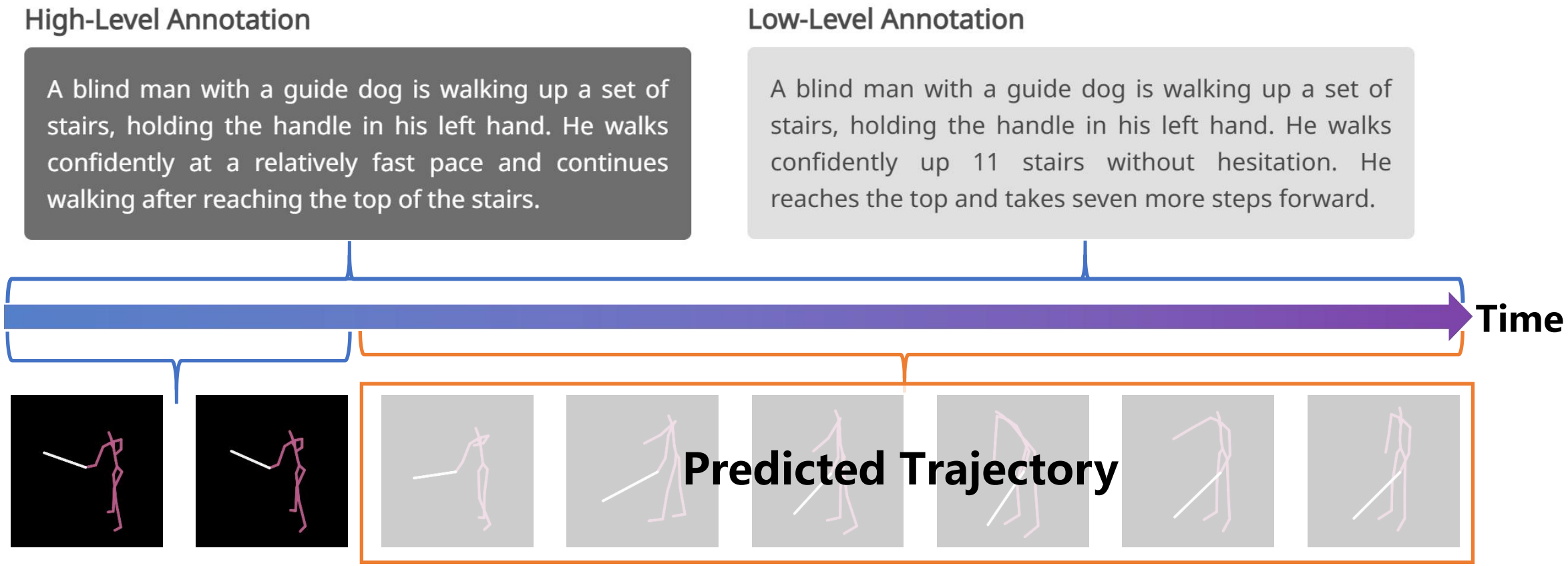




June 11, 2025

Background

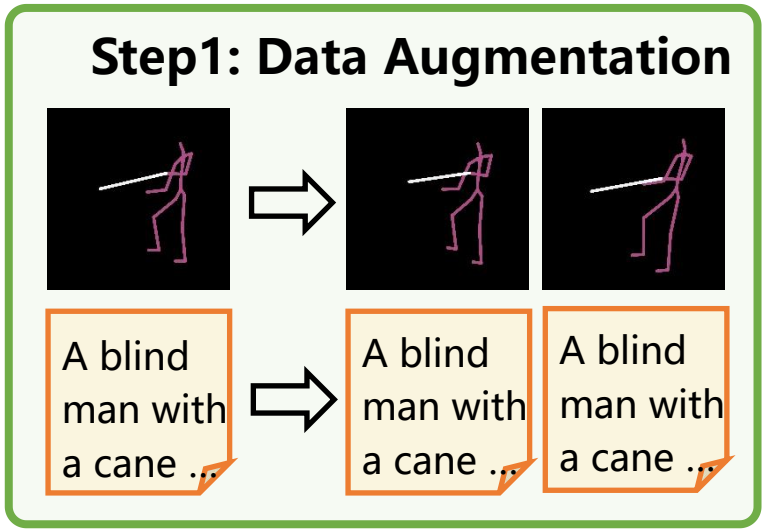
- Overview: Model 3D human motion in accessibility contexts with BlindWays dataset.
- Goal: Predict 3D motion over a 9.5s horizon with motion history and textual descriptions.

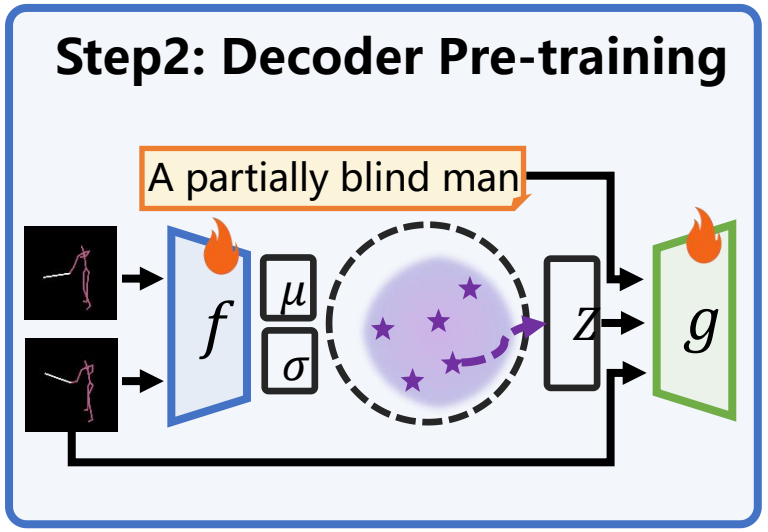


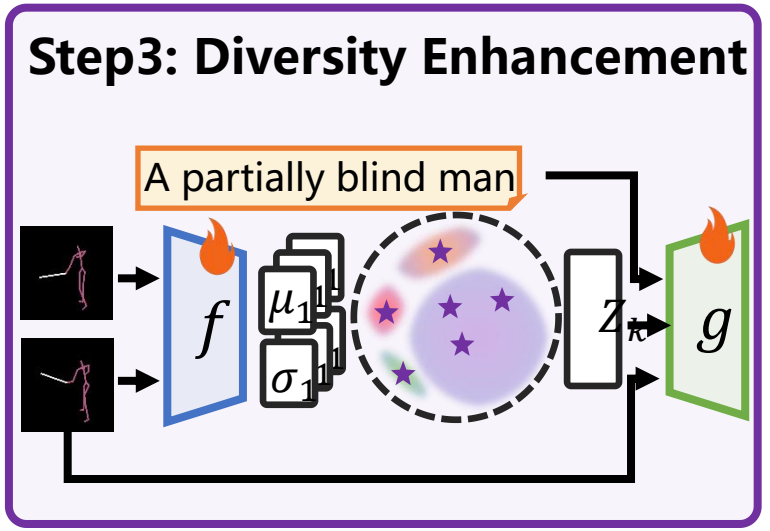
[1] Kim H J, Sengupta K, Kuribayashi M, et al. Text to Blind Motion[J]. In NeurIPS 2025.

Method

- **Overview**
  - Scale of datasets: 365 training samples and 644 testing samples
  - Methods: Cross-modal 3D motion prediction via diversity-accuracy trade-off





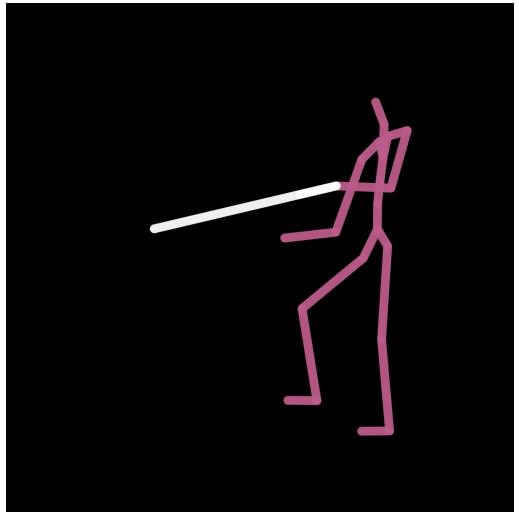


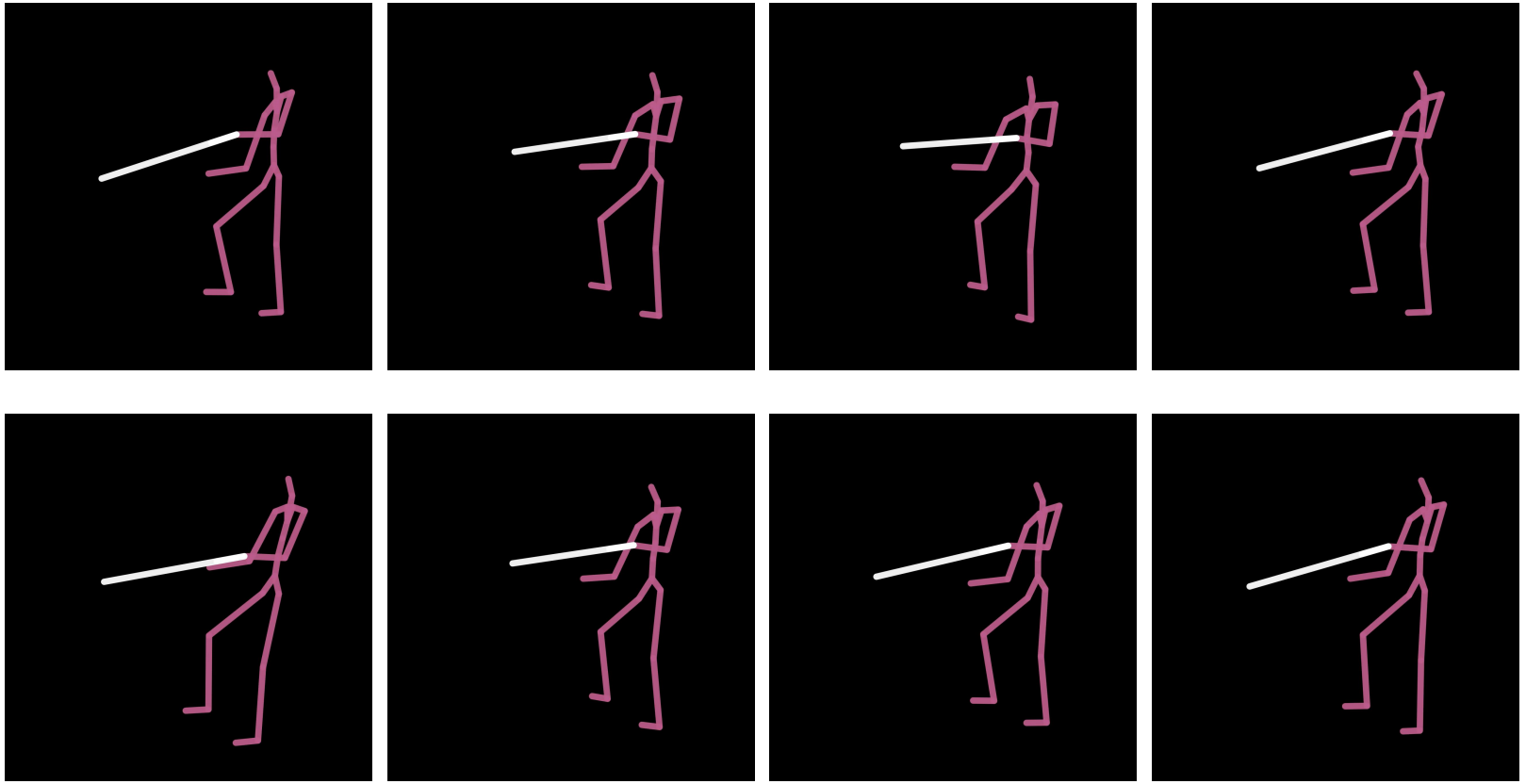
Method

➢ Step1: Data Augmentation (Motion)

- ➢ Perform data enhancement on motion to obtain 20 times more training data.

- Coordinate shift
- Position disturbance
- Rotation around an axis





Method

➢ Step1: Data Augmentation (Text)

➢ Perform data enhancement on text to obtain 20 times more training data.

- Synonym substitution
- Tense changes
- Parentheses Insertion

A blind man with a cane is walking up stairs, holding the railing in his left hand. He reaches the top and finds the door.

A blind man with a cane is **strolling** up stairs, **carrying** the railing in his left hand. He reaches the top and finds the door.

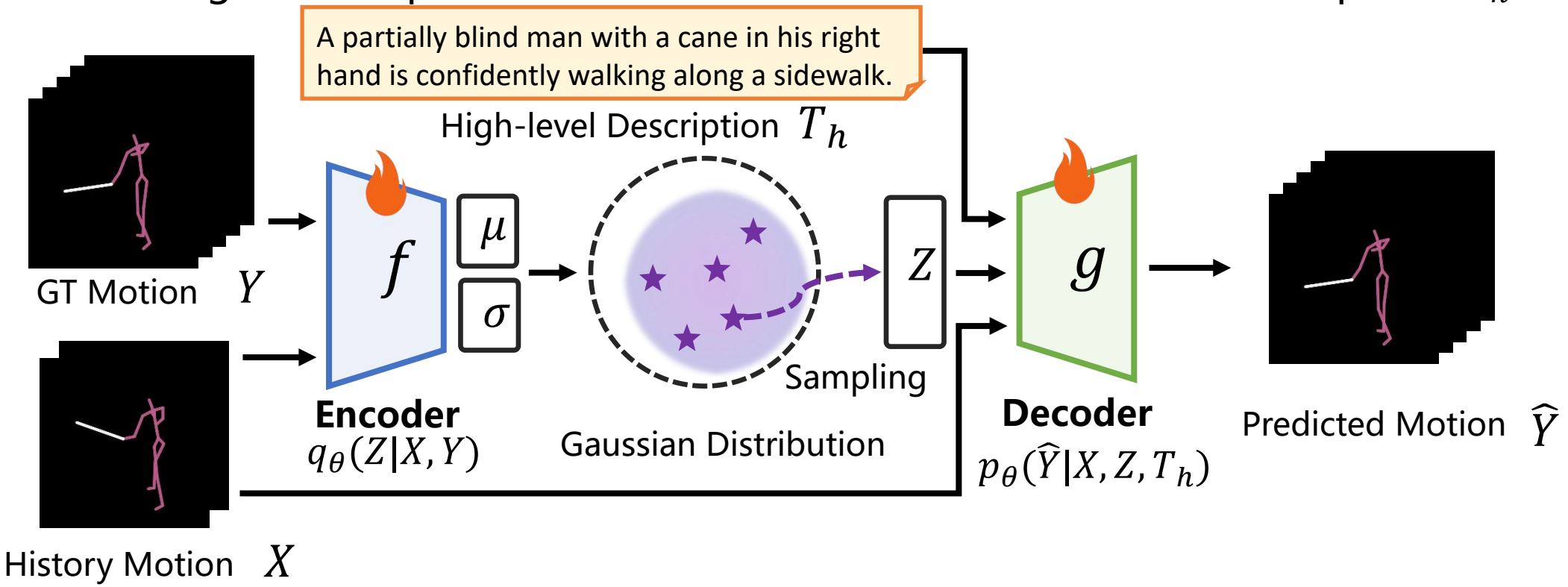
A blind man with a cane is walking up stairs, holding the railing in his left hand. He **reached** the top and **found** the door.

A blind man with a cane is walking up stairs, **grasping** the railing in his left hand. He reaches the top and **then** found the door.

A blind man with a cane is walking up stairs, holding the railing in his left hand. He **arrives at** the top and **seeks** the door.

Method

- **Step2: CVAE Decoder Pre-training (Process)**
  - Jointly encodes historical motion X and GT motion Y to latent space
  - Reparameterization samples from Gaussian distribution
  - Decoder generates predicted motion $\hat{Y}$ conditioned on textual descriptions T_h

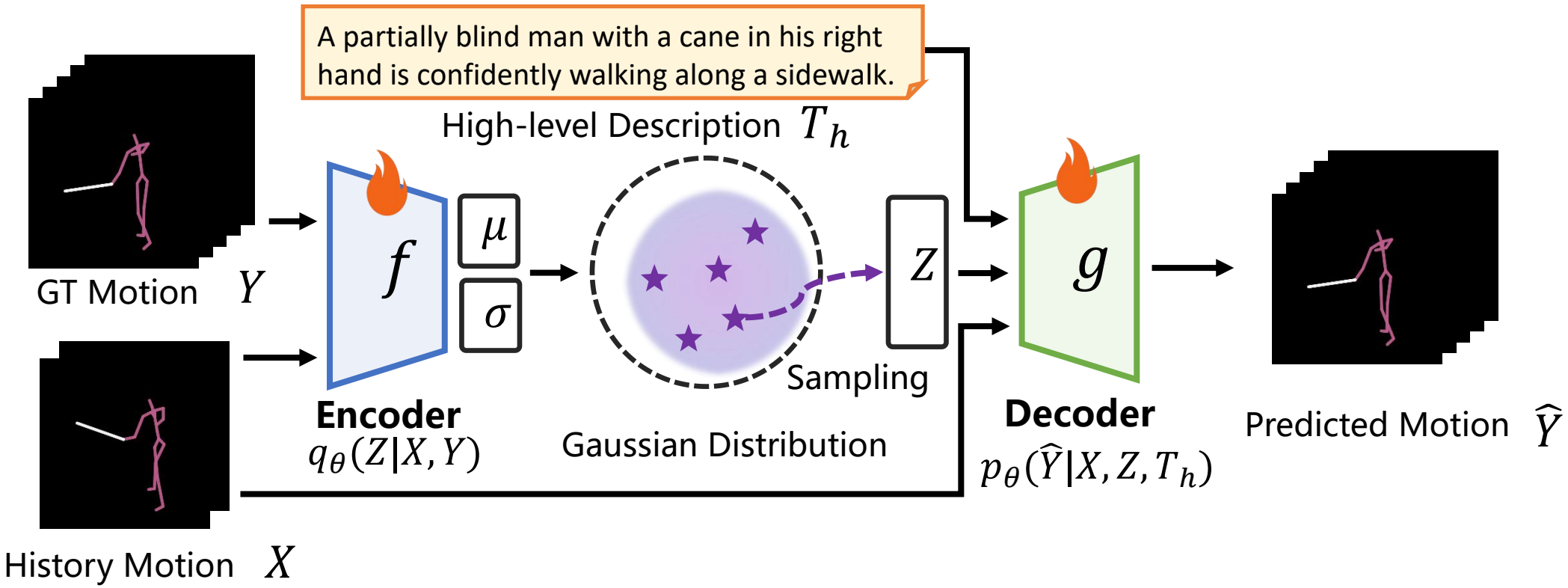


[1] Kingma D P, Welling M. Auto-encoding variational bayes[EB/OL].(2013-12-20)

Method

➢ **Step2: CVAE Decoder Pre-training (Loss)**

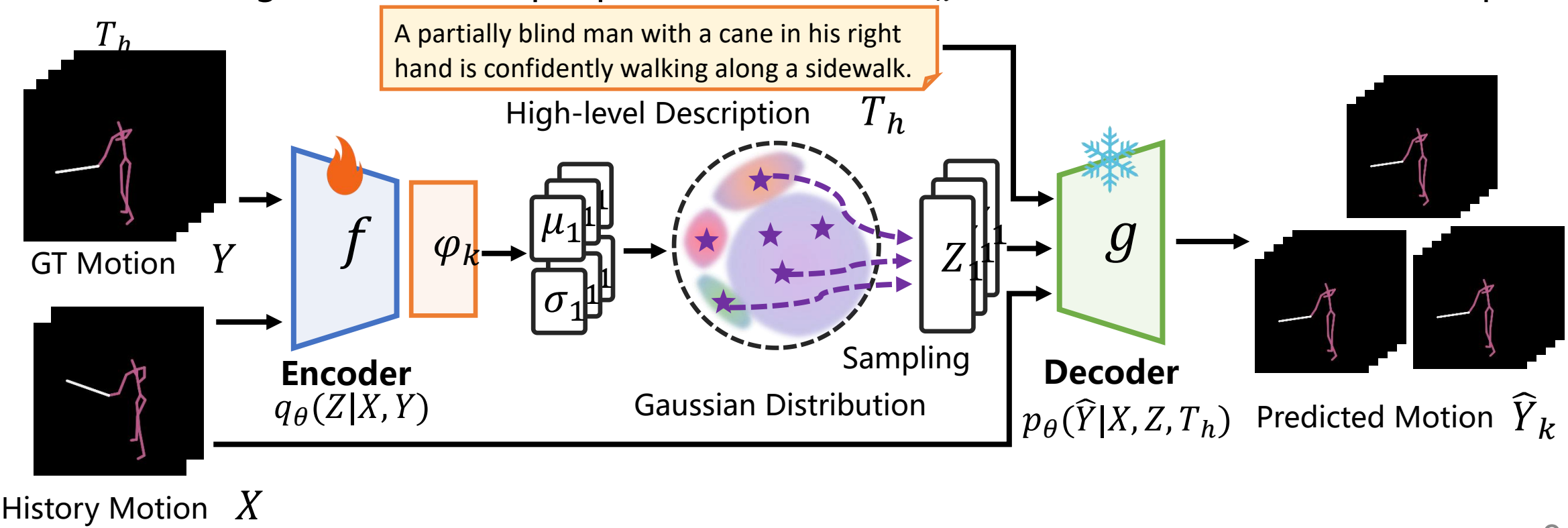
$$\mathcal{L} = \lambda_1 \underbrace{MSE(Y, \hat{Y})}_{\text{Prediction Alignment}} + \lambda_2 \underbrace{MSE(X_t, \hat{Y}_0)}_{\text{Trajectory Smoothing}} + \lambda_3 \underbrace{KLD(\mu, \sigma)}_{\text{Spatial Regularization}}$$



[1] Kingma D P, Welling M. Auto-encoding variational bayes[EB/OL].(2013-12-20)

Method

- **Step3: Diversity Enhancement via DLow (Process)**
  - Jointly encodes history X and GT motion Y to latent space via mapping functions φ_k
  - Reparameterization samples from diverse Gaussian distributions
  - Decoder generates multiple predicted motion $\widehat{Y}_k$ conditioned on textual descriptions

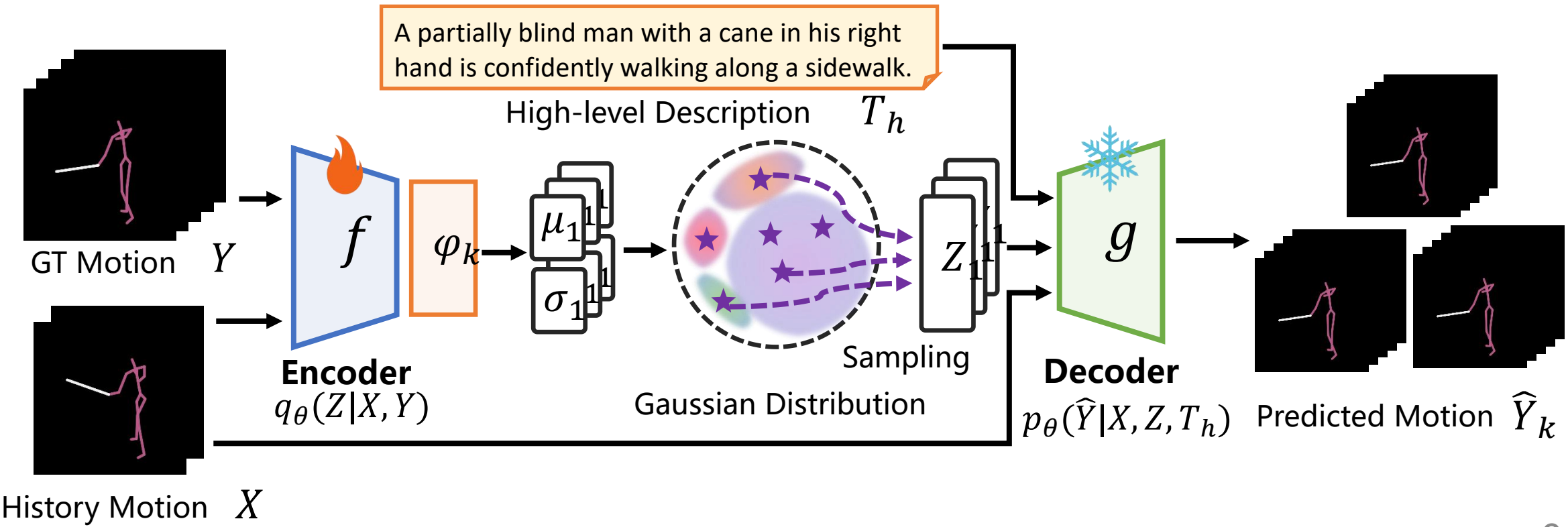


[1] Yuan Y, Kitani K. Dlow: Diversifying latent flows for diverse human motion prediction[C]. In ECCV 2020.

Method

➢ **Step3: Diversity Enhancement via DLow (Loss)**

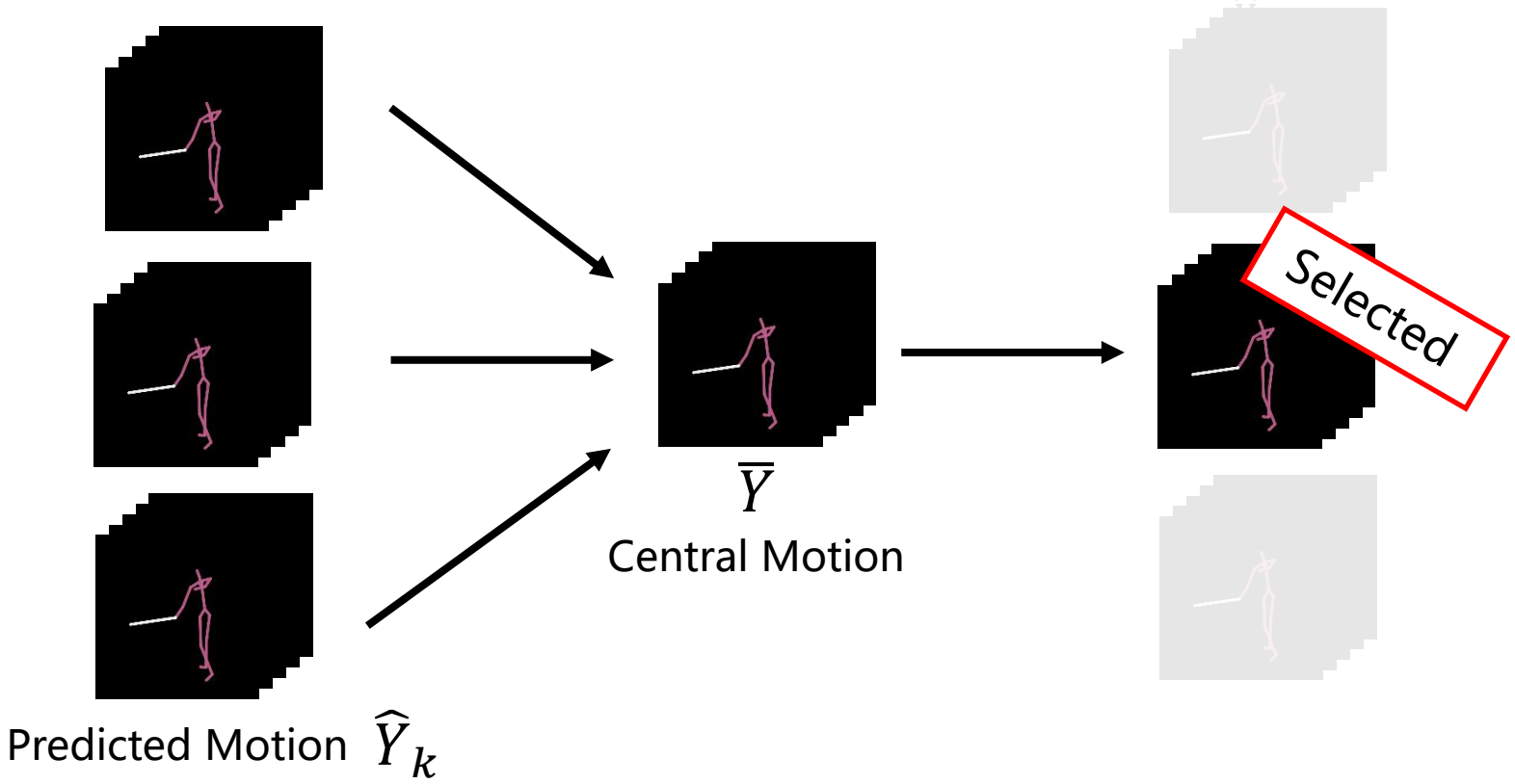
$$\mathcal{L} = \lambda_1 \underbrace{MSE(Y, \hat{Y}_k)}_{\text{Prediction Alignment}} + \lambda_2 \underbrace{d_{i,j}(\hat{Y}_i, \hat{Y}_j)}_{\text{Diversity Enhancement}} + \lambda_3 \underbrace{KLD(\mu_k, \sigma_k)}_{\text{Spatial Regularization}}$$



[1] Yuan Y, Kitani K. Dlow: Diversifying latent flows for diverse human motion prediction[C]. In ECCV 2020.

Method

- **Step3: Diversity Enhancement via DLow (Evaluation)**
  - Ensemble the multiple prediction $\widehat{Y}_k$ to generate a central motion trajectory $\overline{Y}$
  - Calculate the distance $d_k(\widehat{Y}_k, \overline{Y})$ of each prediction $\widehat{Y}_k$ to the central trajectory $\overline{Y}$
  - Select the trajectory closest to the central trajectory: $\widehat{Y}_o = \text{argmin}\, d_k(\widehat{Y}_i, \overline{Y})$



Results

➢ Evaluation Results

- ➢ Each step can further improve the model performance

No.	Step	ADE	FDE
1	Baseline	0.7257	0.8946
2	Data Augmentation	0.6304	0.7565
3	Decoder Pre-training	0.5728	0.6772
4	Diversity Enhancement	**0.5560**	**0.6472**

Further Exploration

- **Additional Attempts:**
  - Dynamic schedules of loss to balance diversity and accuracy in CVAE (ADE=0.5769)
  - Test-time augmentation and model ensemble based on CVAE (ADE=0.5620)
  - Optimization of reconstruction loss distance calculation in DLow (ADE=**0.5540**)

- **Future Work:**
  - Fusion of coarse-grained and fine-grained visual-textual representations
  - Full utilization of visual and textual conditions in encoding process
  - Temporal cycle consistency optimization for improved prediction accuracy
  - …

Thanks for your listening!

E-mail: liuyang232@mails.ucas.ac.cn

Personal Website