Weakly-Supervised Temporal Action Alignment Driven by Unbalanced Spectral Fused Gromov-Wasserstein Distance

Dixin Luo[1] Yutong Wang[1] Angxiao Yue[1] Hongteng Xu[2]

[1]School of Computer Science and Technology, Beijing Institute of Technology [2]Gaoling School of Artificial Intelligence, Renmin University of China

Objectives

Predict frame labels, when the ground truth in training is limited and specifies only the actions appearing in the video, instead of their temporal ordering and frequency.

Introduction

How to characterize videos and textual labels?

How to match between two modalities?

How to train in the weakly-supervised setting?

Our contributions:

1. A novel optimal transport-based solution to set-supervised temporal action alignment.
2. A new contrastive learning paradigm based on US-FGW distance.
3. An algorithm for efficient calculation of US-FGW distance leveraging the Bregman ADMM algorithm.

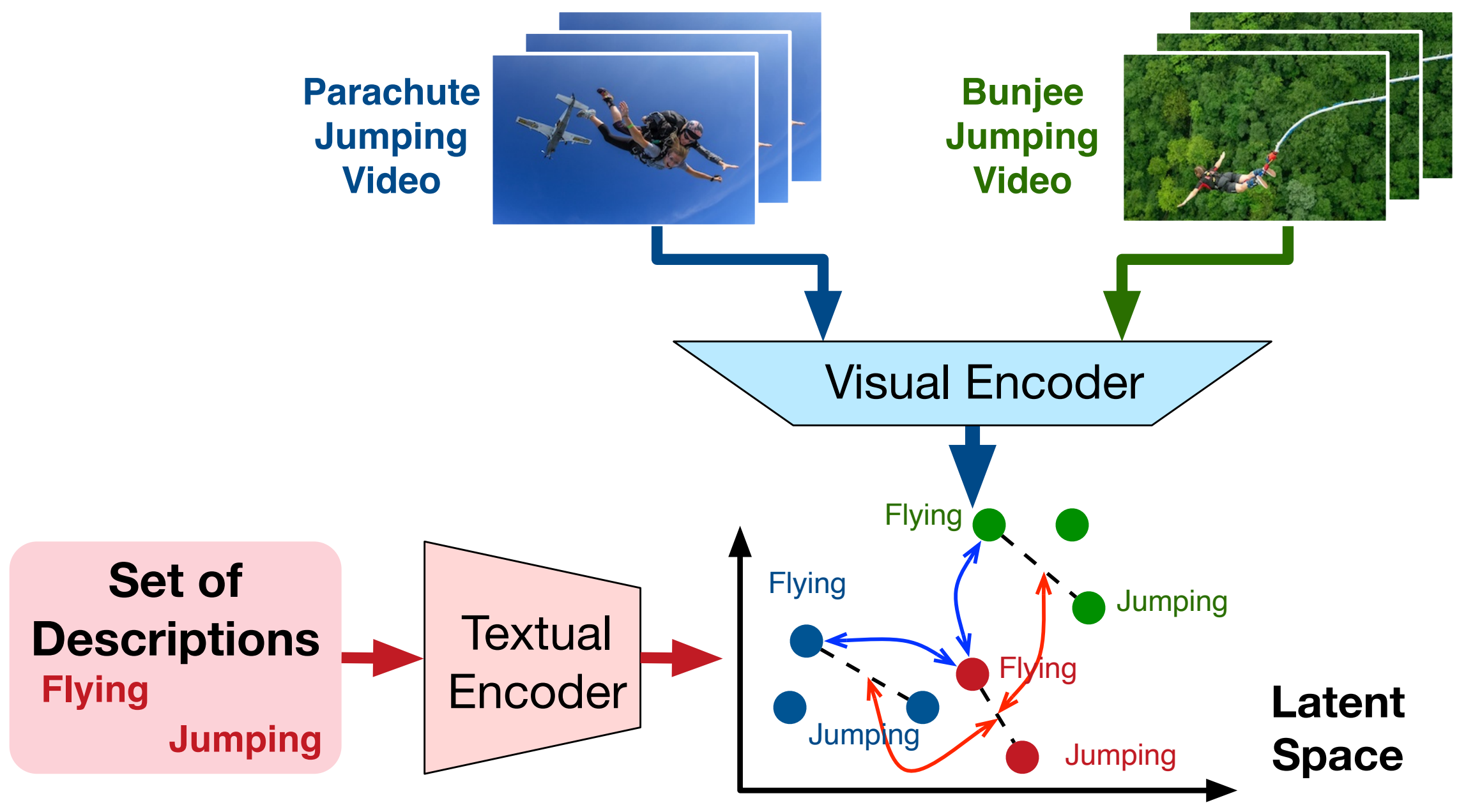
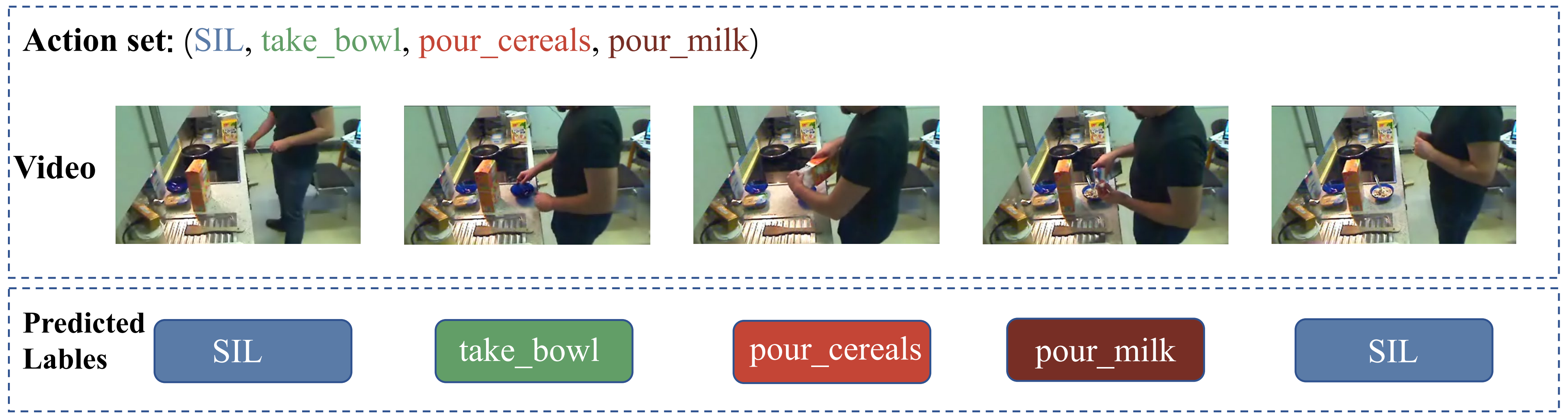


Figure 1:An illustration of the FGW distance, which not only considers the pointwise comparison between the latent codes of different modalities, but also considers the pairwise comparison between the latent relations of different modalities.



Action set: (SIL, take_bowl, pour_cereals, pour_milk)

Video

Predicted Lables SIL take_bowl pour_cereals pour_milk SIL

Learning Strategy

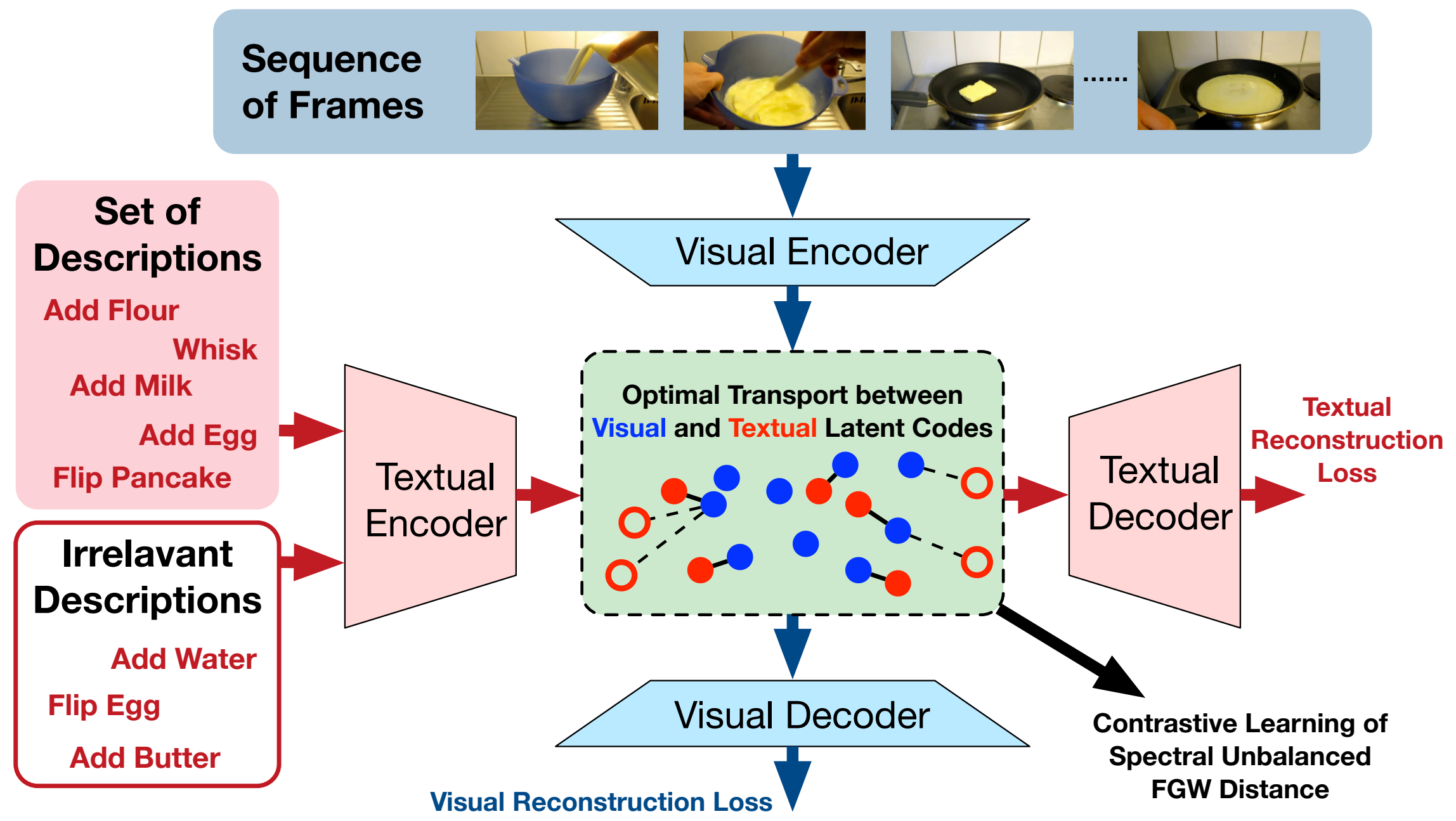


Figure 2:An illustration of the proposed method.

Considering the input video $\mathcal{V}$, the text set $\mathcal{W}$, we obtain the proposed unbalanced spectral FGW (US-FGW) distance:

$$d_{\text{us-fgw}}(\boldsymbol{V},\boldsymbol{W};\beta,\tau) = \min_{\boldsymbol{T}} \underbrace{(1-\beta)\langle -\boldsymbol{K}_{vw},\boldsymbol{T}\rangle}_{\text{Wassertein term}} + \underbrace{\beta\langle -\boldsymbol{K}_v\boldsymbol{T}\boldsymbol{K}_w^T,\boldsymbol{T}\rangle}_{\text{GW term}} + \tau\Big(\text{KL}(\boldsymbol{T}\boldsymbol{1}_J\|\frac{1}{I}\boldsymbol{1}_I) + \text{KL}(\boldsymbol{T}^T\boldsymbol{1}_I\|\frac{1}{J}\boldsymbol{1}_J)\Big).$$

Taking the negative text set $\mathcal{W}_n'$ into account, the overall learning strategy is as follows:

$$\min_{\underbrace{f_v,g_v,f_w,g_w}_{\text{Auto-Encoders}}} \Sigma_{(\mathcal{V}_n,\mathcal{W}_n,\mathcal{W}_n')\in\mathcal{D}}\Big(\underbrace{\ell_v(\mathcal{V}_n,g_v(f_v(\mathcal{V}_n)))}_{\text{Reconstruction loss of frames}} + \underbrace{\ell_w(\mathcal{W}_n,g_w(f_w(\mathcal{W}_n)))}_{\text{Reconstruction loss of words}} + \gamma\big(\underbrace{d_{\text{us-fgw}}(f_v(\mathcal{V}_n),f_w(\mathcal{W}_n);\beta,\tau)}_{\text{Positive US-FGW distance}} - \underbrace{d_{\text{us-fgw}}(f_v(\mathcal{V}_n),f_w(\mathcal{W}_n')}_{\text{Negative US-FGW distance}};\beta,\tau)\big)\Big).$$

Experiments

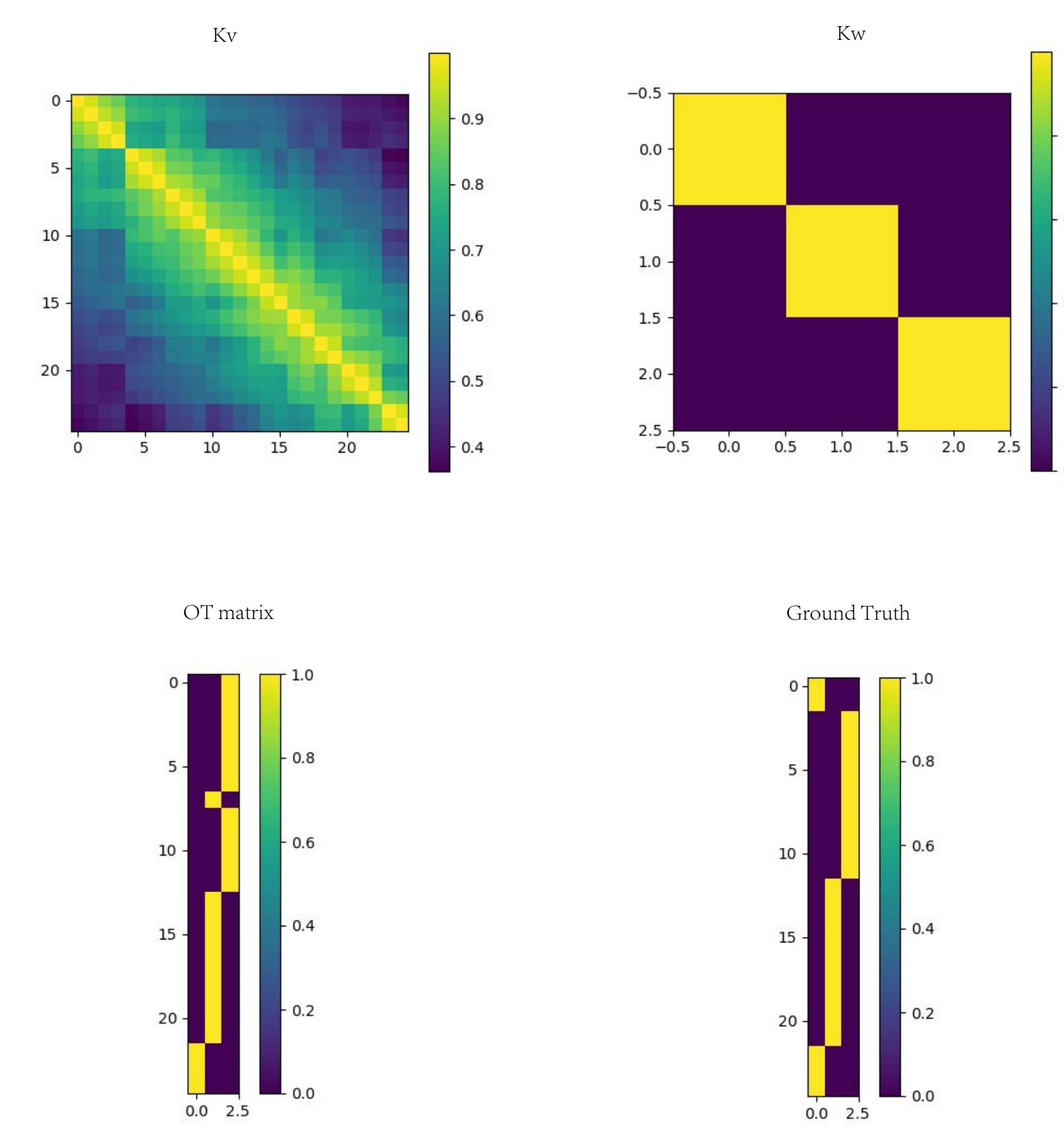


Figure 3:Visualization on distance matrix $\boldsymbol{K}_v$, $\boldsymbol{K}_w$, optimal transport matrix $\boldsymbol{T}$ and ground truth matrix.

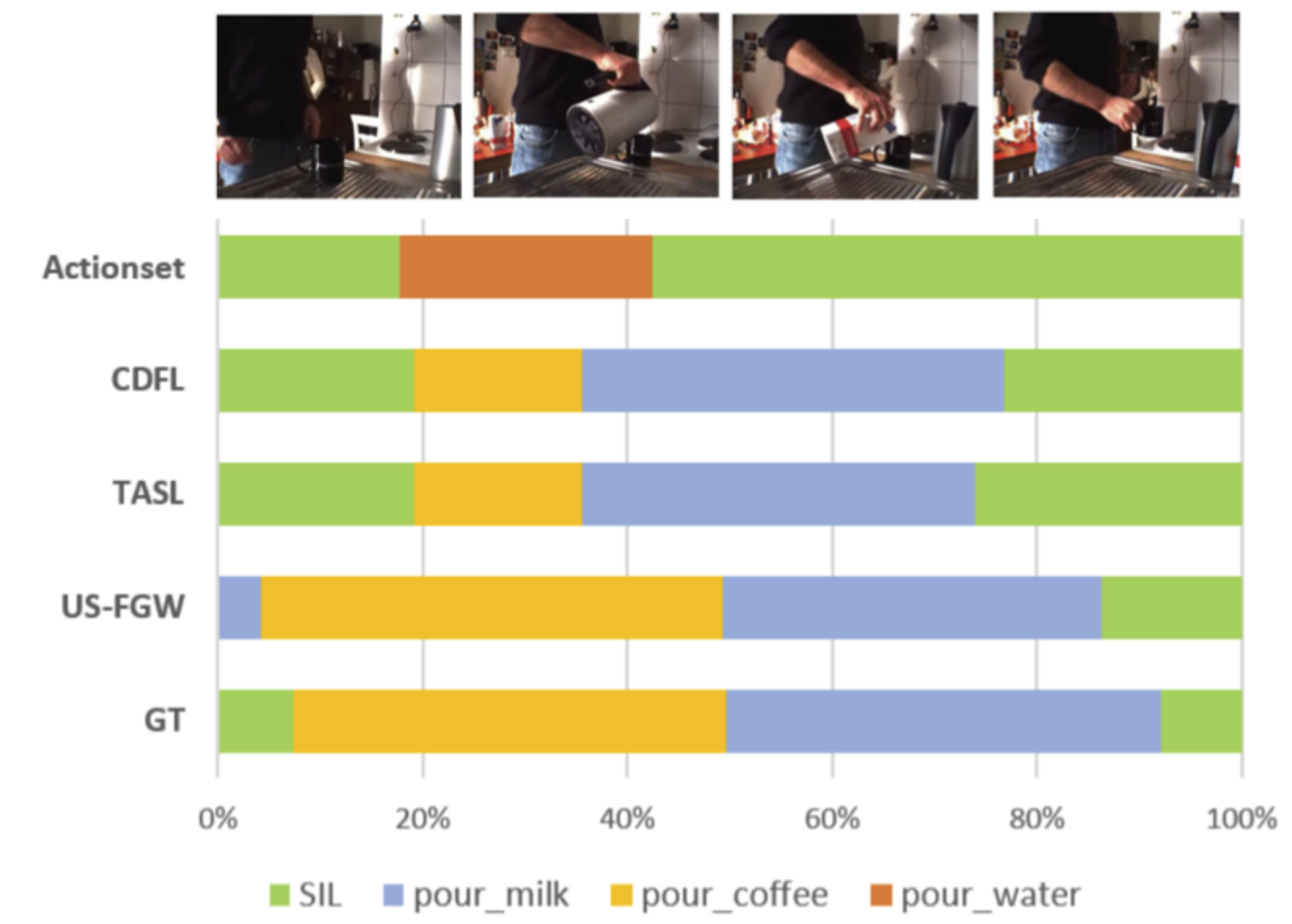


Figure 4:A qualitative alignment result comparison between our US-FGW and Actionset, CDFL and TASL.

Methods	Breakfast (MoF)	Hollywood (IoD)
Transcript-Supervised		
ISBA$_{\text{ED-TCN}}$	0.4548	0.3878
ISBA$_{\text{TCFPN}}$	0.4825	0.3958
NNV	0.5404	0.4692
CDFL	0.5940	0.5009
TASL	0.6042	0.4959
Set-Supervised		
Actionset	0.2137	0.1833
*SCT	0.2660	0.1770
*SCV	0.3020	0.3020
*ACV	0.3340	0.2090
US-FGW(Ours)	0.3357	0.4001

Conclusion

- A new set-supervised paradigm for temporal action alignment based on optimal transport.
- Unified learning of the visual and textual auto-encoders within the new contrastive learning framework.
- Experimental results show that our method achieves encouraging performance for set-supervised temporal action alignment.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62102031, 62106271); the Young Scholar Programme (No. XSQD-202107001) and the Technology Innovation Programme (No. 3070012212116) from Beijing Institute of Technology, and the Research Funds of Renmin University of China (the Fundamental Research Funds for the Central Universities).

Contact Information

- Code and data: https://github.com/hhhh1138/Temporal-Action-Alignment-USFGW
- Email: dixin.luo@bit.edu.cn