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

Dixin Luo
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China

Yutong Wang
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China

Angxiao Yue
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China

Hongteng Xu*
Gaoling School of Artifical Intelligence,
Renmin University of China, Beijing Key Laboratory of
Big Data Management and Analysis Methods
Beijing, China

ABSTRACT

Temporal action alignment aims at segmenting videos into clips and tagging each clip with a textual description, which is an important task of video semantic analysis. Most existing methods, however, rely on supervised learning to train their alignment models, whose applications are limited because of the common insufficiency issue of labeled videos. To mitigate this issue, we propose a weakly-supervised temporal action alignment method based on a novel computational optimal transport technique called unbalanced spectral fused Gromov-Wasserstein (US-FGW) distance. Instead of using videos with known clips and corresponding textual tags, our method just needs each training video to be associated with a set of (unsorted) texts while does not require the fine-grained correspondence between the frames and the texts. Given such weakly-supervised video-text pairs, our method trains the representation models of the video frames and the texts jointly in a probabilistic or deterministic autoencoding architecture and penalizes the US-FGW distance between the distribution of visual latent codes and that of textual latent codes. We compute the US-FGW distance efficiently by leveraging the Bregman ADMM algorithm. Furthermore, we generalize classic contrastive learning framework and reformulate it based on the proposed US-FGW distance, which provides a new viewpoint of contrastive learning for our problem. Experimental results show that our method and its variants outperform state-of-the-art weakly-supervised temporal action alignment methods, whose results are even comparable to those derived by supervised learning methods on some specific evaluation measurements. The code is available at https://github.com/hhhh1138/Temporal-Action-Alignment-USFGW.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; **Matching**; **Video segmentation**.

KEYWORDS

Temporal action alignment, weakly-supervised learning, computational optimal transport, autoencoding, contrastive learning

ACM Reference Format:
Dixin Luo, Yutong Wang, Angxiao Yue, and Hongteng Xu. 2022. Weakly-Supervised Temporal Action Alignment Driven by Unbalanced Spectral Fused Gromov-Wasserstein Distance. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3503161.3548067

1 INTRODUCTION

Temporal action alignment is an important video understanding task, which aims at tagging video clips with textual descriptions according to the semantics of the corresponding scenarios. This task can be treated as a fine-grained video tagging problem — instead of tagging the whole video with a single textual description, temporal action alignment achieves clip/frame-level tagging, which assigns different texts to the clips of different action scenarios. Solving this task helps us understand the time-varying semantics of videos and thus is beneficial for many downstream applications, e.g, video segmentation [21, 52, 66], video retrieval [14, 18, 61], video classification [22, 38], anomaly detection [13, 46], video summary and the associated key frame detection [3, 10].

Generally, existing temporal action alignment methods can be coarsely categorized into two types. Given a set of videos and their associated texts, the methods of the first type, e.g., the works in [45, 69], treat the texts as the labels of the video frames and formulates the task as a classification problem. On the other hand, the methods of the second type, e.g., Moment Context Network (MCN) [1] and Taco [68], learn representation models for both video frames and their texts and formulate the task as a matching problem in a latent space. However, both of the above two methodologies depend on supervised learning and require fine-grained labeled videos (i.e., each frame is labeled by a textual description), while the frame-level video annotation is always time-consuming and

[*] Corresponding author. Email: hongtengxu@ruc.edu.cn

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3548067

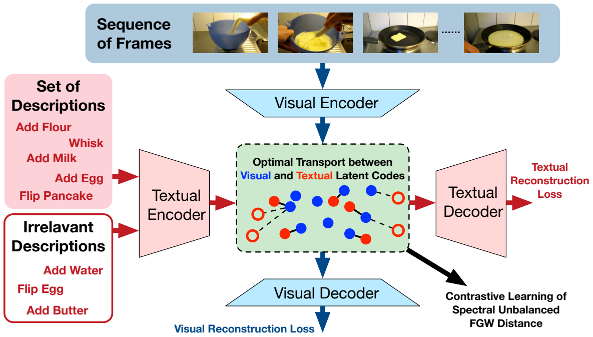

Figure 1: An illustration of our method. In the latent space (the green box), the latent codes of frames and positive texts are represented as blue and red dots, while the latent codes of irrelevant (negative) texts are represented as red circles.

expensive. As a result, the feasibility and the scalability of such supervised methods become questionable in practice.

Recently, some efforts have been made to reduce the requirement of fine-grained labels. For example, the *transcript-supervised* methods in [12, 33, 39, 49] only need each video to be associated with a sequence of texts while do not require the correspondence between the frames and the texts. In such a situation, the temporal action alignment problem is reformulated as a video segmentation problem, i.e., finding the clips corresponding to the texts. However, this labeling strategy requires to preserve the order information of the texts, so that its efficiency degrades a lot for the videos with repeated actions and frequent action changes. To overcome this challenge, the *set-supervised* methods [17, 34, 35, 48] are proposed, which just assign each video with a set of (unsorted) texts and aim at matching each frame with a text in the set. Such set-supervised methods, however, often lead to sub-optimal performance because of the lack of order information. Specifically, they often match the set of frames with that of texts empirically based on the pairwise distance between each frame and each text [6]. Without considering the global matching between the two sets, the matching results are often undesired, especially for the videos containing lots of semantically-meaningless background frames.

To improve the above set-supervised paradigm, in this work, we propose a new and solid weakly-supervised temporal action alignment method with the help of computational optimal transport. As illustrated in Figure 1, given a set of frames and a set of texts, our method applies two encoders to derive their latent codes and matches them in the latent space. Two decoders are associated with the encoders, which reconstruct/predict the raw features of the frames and the texts from the latent codes. In particular, the latent codes are learned via an autoencoding strategy — the encoders and the decoders are learned jointly via minimizing reconstruction errors. Besides learning two autoencoders, we compute the unbalanced spectral fused Gromov-Wasserstein (US-FGW) distance between the latent codes of frames and those of texts. Our US-FGW distance considers the point-wise similarity and the pair-wise similarity between the frames and the texts in the latent space. The optimal transport matrix associated with the US-FGW distance is a joint distribution of the latent codes and indicates the correspondence between the frames and the texts. Based on the US-FGW distance, we design a new contrastive learning loss to regularize the learning of the autoencoders, which minimizes the US-FGW distance of the frame set to its positive text set while penalizes the distance to its negative text set.

In summary, the contributions of our work include: (*i*) We provide a novel optimal transport-based solution to set-supervised temporal action alignment, and the regularizers based on the US-FGW distance lead to a new contrastive learning framework. (*ii*) Leveraging the Bregman ADMM algorithm, we compute the US-FGW distance efficiently. To our knowledge, it is the first attempt to solve unbalanced optimal transport problems via Bregman ADMM. (*iii*) We test our method on representative datasets and compare it to state-of-the-art methods. Experimental results demonstrate the effectiveness of our method.

2 RELATED WORK

2.1 Video action alignment and tagging

As aforementioned, most existing temporal action alignment methods are based on supervised learning [27, 53, 54, 70, 71], in which frame-level video annotations are used in the training. Because the correspondence between frames and texts is known, these methods are designed to encode the explicit correspondence via various sequential models, e.g., Markov chains [27, 30], local filters [23, 50], temporal convolutional networks (TCNs) [16, 29, 36, 55], and recurrent neural networks (RNNs) [54, 70, 71]. However, the insufficiency and the expensiveness of frame-level annotated videos limit the applications of these supervised methods. To solve this problem, many weakly-supervised methods learn their video action alignment models based on coarsely-labeled videos. For the transcript-supervised methods that require the sequence of actions (textual descriptions), they often treat the actions as the unknown latent codes of the corresponding video frame sequence and build hierarchical inference models, e.g., the hidden Markov models (HMMs) [28, 33, 49] and their neural network-based variants [12, 39, 47].

In a more challenging weakly-supervised setting, in which only a set, rather than a sequence of actions is available, the temporal action alignment problem becomes a highly-noisy matching problem across different domains. Facing the set-supervised learning problem, Richard et al. [48] considered three granularity models for video context, length, and frame information, respectively, which reduces the search space of alignment maps significantly. The work in [17] proposed a network to divide a video into small action-consistent clips and estimate its length and the corresponding action label accordingly. The work in [34] proposed a new set-constrained Viterbi algorithm, generating the action segmentation via maximizing the posterior (MAP) of the HMM model. More recently, the work in [35] improves the action alignment results by formulating a more effective differentiable approximation to the NP-hard matching problem. **Different from the above methods, we consider the set-supervised temporal action alignment problem in the perspective of computational optimal transport, which leads to a pair of autoencoders with a contrastive regularizer driven by our US-FGW distance.**

2.2 Optimal transport-based matching methods

The computational optimal transport (OT) methods have been widely used to solve various matching problems [43], e.g., distribution comparison [9], point cloud registration [5], sequence alignment [7, 72], and so on. Essentially, given two distributions, the optimal transport distance between them corresponds to learning an optimal transport plan to minimize the cost of changing one distribution to the other. When the cost is defined as norm-induced distance (e.g., Euclidean distance), the OT distance is specialized as the so-called "Wasserstein distance" [43].

Many challenging machine learning tasks can be solved via minimizing the OT distance. When training high-dimensional generative models, given the data distribution and the model distribution, minimizing their Wasserstein distance directly leads to the Wasserstein autoencoders [11, 58], while maximizing the Kantorovich dual form of the Wasserstein distance leads to the Wasserstein generative adversarial network (WGAN) [2]. For multi-modal data, e.g., the visual frames and the textual descriptions in our task, the work in [67] achieves optimal transport-based domain adaptation. The work in [37] establishes dense correspondences across semantically similar images based on an optimal transport-based framework.

For the structured data like graphs [8] or the point clouds [44] with complex intrinsic structures, a kind of optimal transport distances called Gromov-Wasserstein (GW) distance [41] has shown its potentials in many matching problems. In particular, given sample pairs from different modalities (or sources), the GW distance compares the sample pairs and aims at finding an optimal transport between the samples such that the relational distance between the sample pairs is minimized. Based on the GW distance, the fused Gromov-Wasserstein distance (FGW) distance is proposed in [57], which combines the GW distance with the classic Wasserstein distance and achieves encouraging performance on graph and set matching [63–65]. **The benefits of the FGW distance motivates us to design its variant and propose an optimal transport-based framework for weakly-supervised temporal action alignment.**

3 PROPOSED METHOD

3.1 Problem statement

Suppose that we have a set of videos and associated textual descriptions of actions, denoted as $\mathcal{D} = \{(\mathcal{V}_n, \mathcal{W}_n)\}_{n=1}^{N}$. Here, $\mathcal{V}_n = \{v_{i,n}\}_{i=1}^{I_n}$ represents the I_n frames of the n-th video, which corresponds to different actions happening in the video, and $\mathcal{W}_n = \{w_{j,n}\}_{j=1}^{J_n}$ represents the J_n textual descriptions of the actions associated to the n-th video. As aforementioned, this dataset leads to a set-based weakly-supervised setting — although the video and the text set are paired, the correspondence between the frames (i.e., the $v_{i,n} \in \mathcal{V}_n$) and the texts (i.e., the $w_{j,n} \in \mathcal{W}_n$) is unknown.

Our weakly-supervised temporal action alignment task aims at not only learning the representation models for the frames and the texts, respectively, but also matching their representations in the latent space. Therefore, we need to find a new alignment method that is robust to the uncertainty of the visual-textual correspondence and the interferences caused by the meaningless background frames. We will show that this task can be achieved with the help of a new computational optimal transport distance called unbalanced spectral fused Gromov-Wasserstein distance.

3.2 Autoencoding with FGW regularization

Our alignment method matches the frames with the texts according to their representations in a latent space (a.k.a., the latent codes). We consider an autoencoding architecture to learn the latent codes. For the set of frames $\mathcal{V}$, we denote $f_v : \mathcal{V} \mapsto \mathbb{R}^D$ as the encoder representing each frame feature as a latent code, and $g_v : \mathbb{R}^D \mapsto \mathcal{V}$ as the decoder generating frames from latent codes. Similarly, for the texts in the set $\mathcal{W}$, we denote $f_w : \mathcal{W} \mapsto \mathbb{R}^D$ as the encoder representing each textual description as a latent code, and $g_w : \mathbb{R}^D \mapsto \mathcal{W}$ as the decoder predicting texts from latent codes.

Given $\mathcal{V}$ and $\mathcal{W}$, the above two encoders project them to the latent space and obtain two sets of latent codes as $\boldsymbol{V} = f_v(\mathcal{V}) \in \mathbb{R}^{I\times D}$ and $\boldsymbol{W} = f_w(\mathcal{W}) \in \mathbb{R}^{J\times D}$, respectively. For the latent codes, we would like to ensure that

1. **Representation power:** The latent codes should cover sufficient visual and textual information of the raw data.
2. **Semantic matching:** The latent codes of semantically similar frames and texts should be close to each other.

The first key point is often achieved via minimizing the reconstruction errors of the frames and the texts in the autoencoding framework, and this learning strategy has been commonly used in many generative modeling methods, e.g., the Wasserstein autoencoders in [11, 58, 64]. The second key point requires us to match the visual and textual latent codes, which can be achieved by optimal transport-based methods.

Wasserstein Distance. Given $\boldsymbol{V} \in \mathbb{R}^{I\times D}$ and $\boldsymbol{W} \in \mathbb{R}^{J\times D}$, a straightforward way to measure their difference is computing the sample-based Wasserstein distance [9] between them:

$$d_{\mathrm{w}}(\boldsymbol{V},\boldsymbol{W}) = \min_{\boldsymbol{T}\in\Pi(\boldsymbol{u},\boldsymbol{\mu})} \mathbb{E}_{(\boldsymbol{v},\boldsymbol{w})\sim\boldsymbol{T}}[d(\boldsymbol{v},\boldsymbol{w})] = \min_{\boldsymbol{T}\in\Pi(\boldsymbol{u},\boldsymbol{\mu})} \langle \boldsymbol{D}_{vw},\ \boldsymbol{T}\rangle. \tag{1}$$

Here, $\boldsymbol{D}_{vw} = [d(\boldsymbol{v}_i, \boldsymbol{w}_j)] \in \mathbb{R}^{I\times J}$ is a distance matrix, whose element $d(\boldsymbol{v}_i, \boldsymbol{w}_j)$ represents the distance between the latent code of the i-th frame and that of the j-th word. For the Wasserstein distance, we often apply the Euclidean distance matrix. $\Pi(\boldsymbol{u},\boldsymbol{\mu}) = \{\boldsymbol{T} \geq 0 | \boldsymbol{T}\boldsymbol{1}_J = \boldsymbol{u}, \boldsymbol{T}^T\boldsymbol{1}_I = \boldsymbol{\mu}\}$ is the set of doubly-stochastic matrix, whose marginals must be on the Simplex, i.e., $\boldsymbol{u} \in \Delta^{I-1}$ and $\boldsymbol{\mu} \in \Delta^{J-1}$. Generally, we set the marginals to be uniform, i.e., $\boldsymbol{u} = \frac{1}{I}\boldsymbol{1}_I$ and $\boldsymbol{\mu} = \frac{1}{J}\boldsymbol{1}_J$. The optimal transport matrix corresponding to $d_{\mathrm{w}}(\boldsymbol{V},\boldsymbol{W})$, denoted as $\boldsymbol{T}^* = [t_{ij}^*]$, is the optimal joint distribution of the frames and the texts that minimizing the expectation of the distance, as shown in (1). The elements of $\boldsymbol{T}^*$ can be explained as the coherency probability for each frame-text pair, and thus, we can match the the frames with the texts, i.e., for each frame $v_i \in \mathcal{V}$, the corresponding textual description $w_{\hat{j}} \in \mathcal{W}$ is determined by $\hat{j} = \arg\max_j t_{ij}^*$.

Gromov-Wasserstein Distance. Besides the above Wasserstein distance, the Gromov-Wasserstein (GW) distance is often applied to compute the optimal transport matrix as well. Given the latent codes $\boldsymbol{V}$ and $\boldsymbol{W}$, the GW distance between them is defined as

$$\begin{aligned} d_{\mathrm{gw}}(\boldsymbol{V},\boldsymbol{W}) = \min_{\boldsymbol{T}\in\Pi(\boldsymbol{u},\boldsymbol{\mu})} &\mathbb{E}_{(\boldsymbol{v},\boldsymbol{v}',\boldsymbol{w},\boldsymbol{w}')\sim\boldsymbol{T}\otimes\boldsymbol{T}}[|d(\boldsymbol{v},\boldsymbol{v}') - d(\boldsymbol{w},\boldsymbol{w}')|^2] \\ &\Rightarrow \min\nolimits_{\boldsymbol{T}\in\Pi(\boldsymbol{u},\boldsymbol{\mu})} -\langle \boldsymbol{D}_v \boldsymbol{T} \boldsymbol{D}_w^T,\ \boldsymbol{T}\rangle. \end{aligned} \tag{2}$$

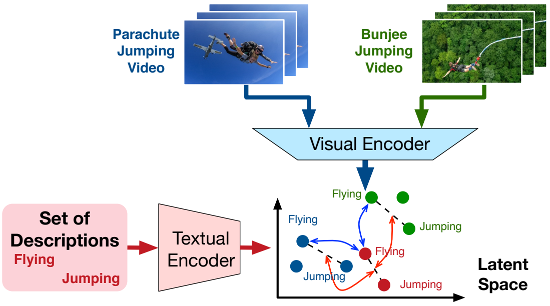

Figure 2: An illustration of the FGW distance. Given the textual and the visual latent codes, the FGW distance not only considers their point-wise distance (the blue arrows) but also considers the relational distance between "Flying" and related concept "Jumping" (the red arrows). Even if the point-wise distance is large, the relational distance can be small (i.e., the lengths of the black dotted lines are similar).

Here, the difference between different pairs $|d(\boldsymbol{v},\boldsymbol{v}')-d(\boldsymbol{w},\boldsymbol{w}')|^2$ is also called relational distance [62]. As shown in (2), the GW distance corresponds to an optimization problem of the transport matrix $\boldsymbol{T}$ that minimizes the expectation of the relational distance, where the Kronecker product $\boldsymbol{T}\otimes\boldsymbol{T}$ works as the joint distribution for the pair of frames and that of texts, i.e., $(\boldsymbol{v},\boldsymbol{v}'\in\boldsymbol{V}\times\boldsymbol{V})$ and $(\boldsymbol{w},\boldsymbol{w}'\in\boldsymbol{W}\times\boldsymbol{W})$. According to the work in [44, 62], the objective function can be reformulated in a matrix form $-\langle\boldsymbol{D}_v\boldsymbol{T}\boldsymbol{D}_w^T,\boldsymbol{T}\rangle$, where $\boldsymbol{D}_v=[d(\boldsymbol{v}_i,\boldsymbol{v}_j)]\in\mathbb{R}^{I\times I}$, and $\boldsymbol{D}_w=[d(\boldsymbol{w}_i,\boldsymbol{w}_j)]\in\mathbb{R}^{J\times J}$ represent the distance matrix of the visual latent codes and that of the textual latent codes, respectively. Note that for the video frames, we can impose their order information on $\boldsymbol{D}_v$: for each $d(\boldsymbol{v}_i,\boldsymbol{v}_j)\in\boldsymbol{D}_v$, we have $d(\boldsymbol{v}_i,\boldsymbol{v}_j)\leftarrow d(\boldsymbol{v}_i,\boldsymbol{v}_j)+\lambda\ell_1(\frac{i}{I},\frac{j}{I})$, where $\ell_1(a,b)=|a-b|$, and $\ell_1(\frac{i}{I},\frac{j}{I})$ defines the distance between the normalized frame indices, and $\lambda\geq 0$ controls its significance.

Fused Gromov-Wasserstein Distance. Given two sets of latent codes, the Wasserstein distance derives the optimal transport matrix based on their point-wise comparisons, while the GW distance derives the optimal transport matrix based on their pair-wise comparisons. Therefore, a natural extension of these two optimal transport distances is considering them jointly, which leads to the fused Gromov-Wasserstein (FGW) distance proposed in [57]:

$$d_{\text{fgw}}(\boldsymbol{V},\boldsymbol{W};\beta)=\min_{\boldsymbol{T}\in\Pi(\boldsymbol{u},\boldsymbol{\mu})}\underbrace{(1-\beta)\langle\boldsymbol{D}_{vw},\ \boldsymbol{T}\rangle}_{\text{Wasserstein term}}+\underbrace{\beta\langle-\boldsymbol{D}_v\boldsymbol{T}\boldsymbol{D}_w^T,\ \boldsymbol{T}\rangle}_{\text{GW term}}.\qquad(3)$$

In particular, the FGW distance is an optimization problem consisting of a Wasserstein term and a GW term. It achieves a trade-off between point-wise comparison and pair-wise comparison when computing the optimal transport matrix, in which the significance of the two terms is controlled by the hyperparameter $\beta\in[0,1]$.

Note that this FGW distance is suitable for weakly-supervised temporal action alignment because the matching of frames and texts depends on not only the point-wise similarity between their latent codes but also the structural or relational similarity between them. Considering these two kinds of similarity jointly helps us to suppress the semantic gap between the visual and the textual information. As illustrated in Figure 2, the texts "Jumping" and "Flying" may correspond to different videos, e.g., "Bungee Jumping" and "Parachute Jumping". The visual latent codes of the "Bungee Jumping" video can be very different from those of the "Parachute Jumping" video, so that the textual latent codes of "Jumping" and "Flying" may not match well with both of them — when purely relying on the Wasserstein term in (3), the text "Flying" (in red) may be wrongly matched with the frame corresponding to "Jumping" (in blue) in the "Parachute Jumping". However, when considering the pairwise relations (based on the GW term in (3)), the distance between the texts "Jumping" and "Flying" in the latent space can be similar to both the distance between the corresponding frames of "Bungee Jumping" video and that between the frames of "Parachute Jumping" video. As a result, the matching driven by the FGW distance is more robust.

Given N paired frame-text sets $\{(\mathcal{V}_n,\mathcal{W}_n)\}_{n=1}^N$, a naïve strategy is learning the autoencoders with the FGW-based regularizer:

$$\begin{aligned}\min_{f_v,g_v,f_w,g_w}\sum_{(\mathcal{V}_n,\mathcal{W}_n)\in\mathcal{D}}&\underbrace{\ell_v(\mathcal{V}_n,g_v(f_v(\mathcal{V}_n)))}_{\text{Reconstruction loss of frames}}+\\&\underbrace{\ell_w(\mathcal{W}_n,g_w(f_w(\mathcal{W}_n)))}_{\text{Reconstruction loss of words}}+\gamma\underbrace{d_{\text{fgw}}(f_v(\mathcal{V}_n),f_w(\mathcal{W}_n);\beta)}_{\text{FGW distance}},\end{aligned}\qquad(4)$$

where $f_v(\mathcal{V}_n))=\boldsymbol{V}_n\in\mathbb{R}^{I_n\times D}$ derives the latent codes of the I_n frames in the video $\mathcal{V}_n$, and similarly, $f_w(\mathcal{W}_n))=\boldsymbol{W}_n\in\mathbb{R}^{J_n\times D}$ derives the latent codes of the J_n texts in the set $\mathcal{W}_n$. The ℓ_v and ℓ_w are the metrics used to quantify the reconstruction errors of frames and texts, respectively.

3.3 Unbalanced spectral FGW distance

According to our above analysis, (4) jointly considers the learning and the matching of visual and textual latent codes. However, this learning strategy still suffers from two potential issues: (*i*) For the videos with lots of meaningless background frames, we need to match texts with a part of frames, while (4) does not have any mechanisms to achieve such partial matching — for each d_{fgw} term in (4), it is often questionable to assume uniform marginal distributions $\boldsymbol{\mu}$ and $\boldsymbol{u}$ and make each $\boldsymbol{T}^*\in\Pi(\boldsymbol{\mu},\boldsymbol{u})$. (*ii*) For the latent codes with high dimensions, it is challenging to leverage the distance metric of the corresponding latent space because of the curse of dimensionality, and as a result, the distance matrices $\boldsymbol{D}_v$, $\boldsymbol{D}_w$, and $\boldsymbol{D}_{vw}$ might be unreliable or indistinguishable. To resolve these two issues, we improve the original FGW distance, proposing the following unbalanced spectral FGW distance.

For each video, each text is generally associated with multiple frames. However, on one hand, we do not know in advance that how many frames are matched with a single text. On the other hand, many frames are meaningless background frames, which should be not assigned to any texts. As a result, the marginal distributions of the frames and the texts (i.e., the $\boldsymbol{u}\in\Delta^{I-1}$ and $\boldsymbol{\mu}\in\Delta^{J-1}$ in (3)), which indicate the significance of different frames and how many frames are matched with a single text, respectively, are unavailable. In such a situation, the original FGW distance degrades to an unbalanced FGW distance — the $\boldsymbol{u}$ and $\boldsymbol{\mu}$ are just the prior of the marginals, and the marginals of the optimal transport are not equal to them strictly. Furthermore, given the latent codes $\boldsymbol{V}$'s and $\boldsymbol{W}$'s, instead of using the Euclidean distance to derive the corresponding distance matrices (i.e., $\boldsymbol{D}_{vw}$, $\boldsymbol{D}_v$, and $\boldsymbol{D}_w$), we design kernel matrices based on the latent codes and replace the distance matrices in the FGW distance with them.

Taking the above two modifications into account, we obtain the proposed unbalanced spectral FGW (US-FGW) distance:

$$\begin{aligned} d_{\text{us-fgw}}(\boldsymbol{V},\boldsymbol{W};\beta,\tau) = &\min\nolimits_{\boldsymbol{T}}\ (1-\beta)\langle -\boldsymbol{K}_{vw}, \boldsymbol{T}\rangle + \beta\langle -\boldsymbol{K}_v\boldsymbol{T}\boldsymbol{K}_w^T, \boldsymbol{T}\rangle \\ &+ \tau\Big(\text{KL}\big(\boldsymbol{T}\boldsymbol{1}_J \| \tfrac{1}{I}\boldsymbol{1}_I\big) + \text{KL}\big(\boldsymbol{T}^T\boldsymbol{1}_I \| \tfrac{1}{J}\boldsymbol{1}_J\big)\Big). \end{aligned} \tag{5}$$

Here, $\boldsymbol{K}_{vw}$, $\boldsymbol{K}_v$, and $\boldsymbol{K}_w$ can be arbitrary kernel matrices derived based on the latent codes, e.g. the RBF kernel, cosine similarity, and so on. For the marginals of the transport matrix, instead of imposing strict equality constraints, we add two regularizers to penalize the KL-divergences between them and uniform distributions ($\frac{1}{I}\boldsymbol{1}_I$ and $\frac{1}{J}\boldsymbol{1}_J$). The significance of the two regularizers is controlled by the hyperparameter τ. The regularizers allow us to learn the significance (i.e., $\boldsymbol{T}^*\boldsymbol{1}_J$) and the assignment of the frames (i.e., $\boldsymbol{T}^*$) while avoid trivial solutions (i.e., $\boldsymbol{T} = \boldsymbol{0}$).

Replacing the FGW distance in (4) with the US-FGW distance in (5), we obtain the following learning problem:

$$\begin{aligned} &\min\nolimits_{f_v,g_v,f_w,g_w} \sum\nolimits_{(\mathcal{V}_n,\mathcal{W}_n)\in\mathcal{D}} \ell_v(\mathcal{V}_n, g_v(f_v(\mathcal{V}_n)))+ \\ &\ell_w(\mathcal{W}_n, g_w(f_w(\mathcal{W}_n))) + \gamma d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}_n);\beta,\tau). \end{aligned} \tag{6}$$

3.4 Contrastive learning based on US-FGW

For each frame set $\mathcal{V}$, besides matching it with the corresponding (positive) text set $\mathcal{W}$, we can further consider its relation to a set of negative texts (e.g., the texts irrelevant to the video), denoted as $\mathcal{W}'$. In particular, besides penalizing the US-FGW distance between $f_v(\mathcal{V})$ and $f_w(\mathcal{W})$, we would like to encourage the US-FGW distance between the frame set and the negative word set in the latent space. The above analysis motivates us to propose a contrastive learning framework based on the US-FGW distance.

Generalization of classic contrastive learning. Take our task as an example. Given a set of frames, the classic contrastive learning methods like noise-contrastive estimation [19] maximizes the difference between the conditional distribution of positive texts and that of negative texts.

$$\begin{aligned} \max\nolimits_{f_v,f_w} \mathbb{E}_{\mathcal{V}\sim p_{\mathcal{D}}}[&\mathbb{E}_{\mathcal{W}\sim p_{\mathcal{P}|\mathcal{V}}}[s(f_w(\mathcal{W});f_v(\mathcal{V}))] - \\ &\mathbb{E}_{\mathcal{W}'\sim p_{\mathcal{N}|\mathcal{V}}}[h^*(s(f_w(\mathcal{W}');f_v(\mathcal{V})))]], \end{aligned} \tag{7}$$

where $p_{\mathcal{D}}$ represents the (empirical) distribution of training data, $p_{\mathcal{P}|\mathcal{V}}$ and $p_{\mathcal{N}|\mathcal{V}}$ represent the positive and the negative text distributions conditioned on the frame set $\mathcal{V}$. $s(\cdot;\cdot)$ is a score function indicating the similarity between the latent codes. $h^* : \mathbb{R} \mapsto \mathbb{R}$ is a conjugate function, whose formulation is determined by the final activation layer of the score function s.

Following the work in [42], we can explain the framework in (7) as maximizing the expectation of the f-divergence between $p_{\mathcal{P}|\mathcal{V}}$ and $p_{\mathcal{N}|\mathcal{V}}$. Specifically, when $s(\boldsymbol{a};\boldsymbol{b}) = -\log(1+\exp(-\boldsymbol{a}^\top\boldsymbol{b}))$, $h^*(t)$ becomes $-\log(1-\exp(t))$, and we can rewrite (7) as a mutual information maximization problem [20, 59]. When h^* is an identity function, *i.e.*, $h^*(t) = t$, (7) becomes a score matching framework corresponding to the maximum mean discrepancy (MMD) [15, 32]. This explanation provides a way to generalize the contrastive learning framework in (7) — leveraging a valid (pseudo) metric of distributions, we can achieve contrastive learning via maximizing the expectation of the discrepancy between the positive distribution and the negative one based on the metric, i.e.,

$$\max \mathbb{E}_{\mathcal{V}\sim p_{\mathcal{D}}}[d(p_{\mathcal{P}|\mathcal{V}}, p_{\mathcal{N}|\mathcal{V}})], \tag{8}$$

where d represents the (pseudo) metric of the distributions.

The proposed US-FGW contrastive learning. In this work, we take our US-FGW distance as the metric d in (8). Moreover, because our task involves two modalities, instead of maximizing the discrepancy between the conditional distributions of positive and negative texts, we shall focus more on the comparison between the distributions of texts and frames. Therefore, leveraging the triangle inequality of US-FGW distance,[1] we can relax the objective function in (8) and achieve contrastive learning via maximizing its lower bound, i.e.,

$$\begin{aligned} &\mathbb{E}_{\mathcal{V}\sim p_{\mathcal{D}}}[d_{\text{us-fgw}}(p_{\mathcal{P}|\mathcal{V}}, p_{\mathcal{N}|\mathcal{V}})] \\ \geq &\mathbb{E}_{\mathcal{V}\sim p_{\mathcal{D}}}[d_{\text{us-fgw}}(p_{\mathcal{N}|\mathcal{V}}, p_{\mathcal{V}}) - d_{\text{us-fgw}}(p_{\mathcal{P}|\mathcal{V}}, p_{\mathcal{V}})], \end{aligned} \tag{9}$$

where $p_{\mathcal{V}}$ is the distribution of video frames in a set. For the right side of the triangle inequality, we ignore the absolute operation because the distance of the frames to their negative texts should always be larger than the distance to the positive texts.

For the US-FGW distances in (9), we can obtain their sample-based estimation given $\mathcal{V} \sim p_{\mathcal{V}}$, $\mathcal{W} \sim p_{\mathcal{P}|\mathcal{V}}$, and $\mathcal{W}' \sim p_{\mathcal{N}|\mathcal{V}}$, e.g., the US-FGW distance shown in (6). Considering this contrastive learning strategy leads to the proposed learning task:

$$\begin{aligned} &\min\nolimits_{f_v,g_v,f_w,g_w} \sum\nolimits_{(\mathcal{V}_n,\mathcal{W}_n,\mathcal{W}_n')\in\mathcal{D}} \Big(\ell_v(\mathcal{V}_n, g_v(f_v(\mathcal{V}_n)))+ \\ &\ell_w(\mathcal{W}_n, g_w(f_w(\mathcal{W}_n))) + \gamma\big(d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}_n);\beta,\tau) - \\ &d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}_n');\beta,\tau)\big)\Big). \end{aligned} \tag{10}$$

Here, the last term weighted by the hyperparameter γ is the US-FGW contrastive learning regularizer, which minimizes the US-FGW distance of the frame set to the positive text set while maximizes its US-FGW distance to the negative text set.

4 LEARNING AND IMPLEMENTATIONS

4.1 Probabilistic or deterministic autoencoding

It should be noted that the learning framework in (10) is compatible with both probabilistic and deterministic autoencoders. For the probabilistic autoencoding model, like the variational autoencoder [25], the encoder outputs the mean and the logarithmic variance of the posterior distribution. In our model, given a frame $v_i \in \mathcal{V}$ or a text $w_j \in \mathcal{W}$, the probabilistic autoencoding model is

$$\begin{aligned} &\text{Encoding: } \boldsymbol{m}_{v_i},\ \log\boldsymbol{\sigma}_{v_i} = f_v(v_i), \quad \boldsymbol{m}_{w_j},\ \log\boldsymbol{\sigma}_{w_j} = f_w(w_j), \\ &\text{Decoding: } g_v(f_v(v_i)) = g_v(\boldsymbol{m}_{v_i} + \boldsymbol{\sigma}_{v_i} \odot \boldsymbol{\epsilon}), \\ &\qquad\qquad\quad g_w(f_w(w_j)) = g_w(\boldsymbol{m}_{w_j} + \boldsymbol{\sigma}_{w_j} \odot \boldsymbol{\epsilon}), \end{aligned} \tag{11}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$ is a random vector obeying to a normal distribution. Applying the reparameterization trick, we can sample the latent codes (i.e., $\boldsymbol{v}_i = \boldsymbol{m}_{v_i} + \boldsymbol{\sigma}_{v_i} \odot \boldsymbol{\epsilon}$ and $\boldsymbol{w}_j = \boldsymbol{m}_{w_j} + \boldsymbol{\sigma}_{w_j} \odot \boldsymbol{\epsilon}$)

[1] Although US-FGW is a pseudo-metric, it is easy to proof that it satisfies the triangle inequality based on the method shown in [51, 57].

and decode them. The reconstruction terms in (10) can be derived based on the data reconstructed from the decoding results.

As shown in (11), the probabilistic encoding results of I frames and J texts are two Gaussian mixture models, i.e., $\{\mathcal{N}(\boldsymbol{m}_{v_i}, \boldsymbol{\sigma}_{v_i}^2)\}_{i=1}^I$ and $\{\mathcal{N}(\boldsymbol{m}_{w_i}, \boldsymbol{\sigma}_{w_i}^2)\}_{j=1}^J$. Accordingly, the US-FGW distance in (5) can be derived as a hierarchical optimal transport problem [40], where the elements of the kernel matrices can be defined based on Wasserstein distance between arbitrary two Gaussian components, i.e., for $\boldsymbol{K}_{vw} = [k_{ij}]$,

$$\begin{aligned} k_{ij} &= \exp\left(-\frac{1}{b} d_{\mathrm{w}}^2(\mathcal{N}(\boldsymbol{m}_{v_i}, \boldsymbol{\sigma}_{v_i}^2), \mathcal{N}(\boldsymbol{m}_{w_j}, \boldsymbol{\sigma}_{w_j}^2))\right) \\ &= \exp\left(-\frac{1}{b}(\|\boldsymbol{m}_{v_i} - \boldsymbol{m}_{w_j}\|_2^2 + \|\boldsymbol{\sigma}_{v_i} - \boldsymbol{\sigma}_{w_j}\|_2^2)\right), \end{aligned} \tag{12}$$

where b is the bandwidth of the kernel. The elements of $\boldsymbol{K}_v$ and $\boldsymbol{K}_w$ can be defined in the same way.

For the deterministic autoencoding model, like the Wasserstein autoencoder [58], the encoder outputs latent codes of data directly (i.e., $\boldsymbol{v}_i = f_v(v_i)$ and $\boldsymbol{w}_i = f_w(w_i)$). In such a situation, we can reconstruct frames and texts directly by decoding the latent codes. For the kernel matrices, we can derive them based on the latent codes as well, e.g., for each $k_{ij} \in \boldsymbol{K}_{vw}$, $k_{ij} = \exp(-\frac{1}{b}\|\boldsymbol{v}_i - \boldsymbol{w}_j\|_2^2)$, and $\boldsymbol{K}_v$ and $\boldsymbol{K}_w$ are defined in the same way. In summary, we have $\boldsymbol{K} = \exp(-\frac{1}{b}\boldsymbol{D})$ in general, where $\boldsymbol{D}$ is the Wasserstein distance matrix of latent Gaussian components for the probabilistic autoencoder and the Euclidean distance matrix of latent codes for the deterministic autoencoder, respectively.

4.2 Nested alternating optimization

We solve the learning problem in (10) by a nested alternating optimization. In particular, the outer loop corresponds to computing the US-FGW distances and updating the autoencoders iteratively, while the inner loop corresponds to updating the optimal transport matrices and the corresponding auxiliary and dual variables iteratively when computing the US-FGW distances.

In the inner loop, given current visual and textual autoencoders, we first learn the optimal transport matrices for each pair of the frame set and the text set. Different from most existing methods, which solve optimal transport problems via the Sinkhorn scaling algorithm [9, 56, 65], we apply the Bregman alternating direction method of multipliers (B-ADMM) [60, 62] to compute the US-FGW distance. In particular, we rewrite (5) in an equivalent format by introducing three auxiliary variables $\boldsymbol{S}$, $\boldsymbol{u}$ and $\boldsymbol{\mu}$:

$$\begin{aligned} &\min\nolimits_{\boldsymbol{T},\boldsymbol{S},\boldsymbol{u},\boldsymbol{\mu}} (1-\beta)\langle -\boldsymbol{K}_{vw}, \boldsymbol{T}\rangle + \beta\langle -\boldsymbol{K}_v \boldsymbol{S} \boldsymbol{K}_w^T, \boldsymbol{T}\rangle + \\ &\qquad \tau\Big(\mathrm{KL}(\boldsymbol{u}\|\frac{1}{I}\boldsymbol{1}_I) + \mathrm{KL}(\boldsymbol{\mu}\|\frac{1}{J}\boldsymbol{1}_J)\Big) \\ &s.t.\ \boldsymbol{T} = \boldsymbol{S},\ \boldsymbol{T}\boldsymbol{1}_J = \boldsymbol{u},\ \boldsymbol{S}^T\boldsymbol{1}_I = \boldsymbol{\mu}. \end{aligned} \tag{13}$$

These three auxiliary variables correspond to the optimal transport matrix $\boldsymbol{T}$ and its two marginals. This problem can be further rewritten in a Bregman-augmented Lagrangian form by introducing three dual variables $\boldsymbol{Z}$, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ for the three constraints in (13), respectively. Applying an alternating optimization strategy, we update the primal, the auxiliary, and the dual variables iteratively. The detailed algorithm is shown in Appendix A.

5 EXPERIMENTS

5.1 Implementation details

Baselines and variants of our US-FGW method. To demonstrate the feasibility and effectiveness of our US-FGW method, we compare it with state-of-the-art set-based weakly-supervised action alignment methods, including Actionset [48], SCT [17], SCV [34], ACV [35]. For a comprehensive study, we also use representative transcript-supervised methods (*i.e.*, ISBA [12], NNV [49], CDFL [33], TASL [39]) and the state-of-the-art attention based action localization method (*i.e.*, UM [31]) as our baselines. To achieve ablation study, we further consider the two variants of our US-FGW method that solve (4) and (6), respectively.

Datasets. For each method, we consider three representative temporal action alignment datasets in our experiments, including Breakfast [26], Hollywood Extended [4] and CrossTask [73]. The Breakfast dataset consists of 1,712 video sequences and 48 cooking actions. Each video contains 6.9 actions on average, and has 7.3% background frames. The Hollywood Extended dataset contains 937 Hollywood movie clips with 16 actions. Each video contains 2.5 actions on average, while 60.9% of frames are background. The CrossTask dataset consists of 2,552 videos and 80 actions. Each video contains 14.4 actions on average, while 74.8% of frames are background. For a fair comparison, we following the feature engineering and the data splitting used in previous works. For Breakfast and Hollywood Extended, we use the 64-dimensional features provided by [49]. For CrossTask dataset, we use the 64-dimensional features provided by [39]. For Breakfast, we use 80% data for training and the remaining 20% for testing. For Hollywood and CrossTask, we follow the TASL [39]'s training/testing splitting manner, using 90% for training and 10% for testing.

Model architecture and hyperparameter setting. For the visual autoencoder, we apply a GRU as a feature extraction module of frame sequences, and pass the frame features through a MLP with two fully-connected (FC) layers. The dimension of the latent code is 8. Given the visual latent codes, we reconstruct frame features through a decoder consisting of two three FC layers. For the textual autoencoder, the encoder is an embedding layer converting action indices to latent codes, and the decoder is a MLP with two FC layers, which predicts the probabilities of the actions. For the two autoencoders, the adjacent FC layers are connected through the ReLU activation layers, and their encoders are deterministic.[2] Accordingly, the visual autoencoder takes the mean-square-error (MSE) as its reconstruction loss, while the textual autoencoder takes the cross-entropy loss as its reconstruction loss. The hyperparameters of our algorithm are set as follows: the number of epochs is 2; the learning rate is 0.0003; each batch contains the frames sampled from one video, and the sampling rate is 1 frame per second; the optimizer is Adam [24] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$; the weight γ of the contrastive regularizer is 1.0; the bandwidth b of the kernel is 0.1; the weight β of the GW term is set to be 0.1; the weight τ of the unbalanced regularizer is 0.01; the maximum number of B-ADMM iterations is 3,000, and the weight ρ is 1.0.

Evaluation. Following the work in [12, 33], we use the following four measurements to evaluate each method. The mean-over-frames

[2] We found that using probabilistic encoders provides similar learning results, so we only report the results of using deterministic encoders in the following experiments.

Table 1: Comparisons for various methods on three datasets. The baselines are categorized according to the level of supervision. For the methods with "∗", their source codes are unavailable so that we quote their results from the references, and "-" means that the corresponding results are not provided by the references.

Methods and Categories		Breakfast				Hollywood Extended				CrossTask			
		Mof	Mof-bg	IoU	IoD	Mof	Mof-bg	IoU	IoD	Mof	Mof-bg	IoU	IoD
Transcript-Supervised	ISBA$_{\text{ED-TCN}}$ [12]	0.4548	0.2653	0.2790	0.4775	0.5553	0.4865	0.2228	0.3878	0.5174	0.5645	0.1869	0.2722
	ISBA$_{\text{TCFPN}}$ [12]	0.4825	0.2560	0.2911	0.4972	0.5598	0.4826	0.2348	0.3958	0.5308	0.5739	0.1914	0.2706
	NNV [49]	0.5404	0.5131	0.4415	0.5998	0.5656	0.4737	0.3195	0.4692	0.4132	0.1991	0.1154	0.1888
	CDFL [33]	0.5940	0.5692	0.4391	0.6041	0.5982	0.6200	0.3514	0.5009	0.4248	0.2039	0.1256	0.2006
	TASL [39]	0.6042	0.5842	0.4880	0.6405	0.6066	0.5551	0.3546	0.4959	0.5148	0.4130	0.1859	0.2841
Set-Supervised	Actionset [48]	0.2137	0.1614	0.0504	0.1272	0.3743	0.2111	0.0966	0.1833	0.2993	0.1144	0.0268	0.0476
	SCT [17]	0.122	0.130	0.036	0.078	0.055	0.220	0.008	0.026	0.064	0.150	0.021	0.049
	*SCT [17]	0.2660	-	-	-	-	-	-	0.1770	-	-	-	-
	*SCV [34]	0.3020	-	-	-	-	-	-	0.1770	-	-	-	-
	*ACV [35]	0.3340	-	-	-	-	-	-	0.2090	-	-	-	-
	UM [31]	0.0383	0.0104	0.0315	0.0649	0.4053	0.2427	0.1845	0.2945	0.2053	0.1855	0.0437	0.1498
	Proposed US-FGW	0.3357	0.3577	0.1160	0.1530	0.3840	0.4082	0.2307	0.4001	0.1853	0.1907	0.0720	0.1929

(Mof) is the average percentage of correctly-labeled frames. For the video dominated by background frames, e.g., those in the Hollywood Extended and the CrossTask datasets, the mean-over-frames without background (Mof-bg) records the average percentage of correctly-labeled non-background frames, which is a more reasonable measurement. Given the ground truth labels of the frames GT and their detected actions D, we further compute the intersection over union (IoU) and the intersection over detection (IoD) as $IoU = |GT \cap D| / |GT \cup D|$, and $IoD = |GT \cap D| / |D|$, respectively.

5.2 Comparison experiments

For each dataset, we compare our US-FGW method and its variants with the baselines, including the set-supervised and the transcript-supervised methods, on above four metrics. Since SCT [17] uses I3D features as input to the model, for a fair comparison, we also show the results of our replication of SCT by its released codes when using IDT features, denoted as SCT. We list the averaged performance of various methods in 5 trials in Table 1. The standard deviation of each method's result is smaller than 0.005. As a new set-supervised method, our US-FGW method outperforms existing set-supervised methods in most situations. Especially for the datasets with lots of meaningless background frames (i.e., Hollywood Extended and CrossTask), our method achieves obvious improvements on Mof-bg, IoU, and IoD, which means that it effectively filters out background frames and thus predicts the actions of the semantically-meaningful frames with higher accuracy. These results demonstrate the rationality of using unbalanced optimal transport to achieve partial matching between texts and frames.

It should be noted that our US-FGW method shrinks the gap between the set-supervised methods and the transcript-supervised methods. As shown in Table 1, the transcript-supervised methods always work better than the set-supervised methods (including ours) because they leverage the order information of texts in both training and testing phases. However, for the challenging datasets with lots of background frames, our US-FGW method is comparable to the transcript-supervised methods on some evaluation measurements. For example, for the Hollywood Extendd dataset, our US-FGW method is comparable to the ISBA methods [12] on Mof-bg and IoU, and it even outperforms the ISBA$_{\text{ED-TCN}}$ on IoD. Similarly, for the CrossTask dataset, our method is comparable to the NNV [49] on Mof-bg and outperforms it on IoD. Figure 3 further qualitatively compares our proposed method's output with the ground truth and other referenced methods. Although our method may fail to detect the background tag (i.e., SIL), it indeed detects actions and localizes their changes with higher accuracy than both transcript-supervised and set-supervised baselines.

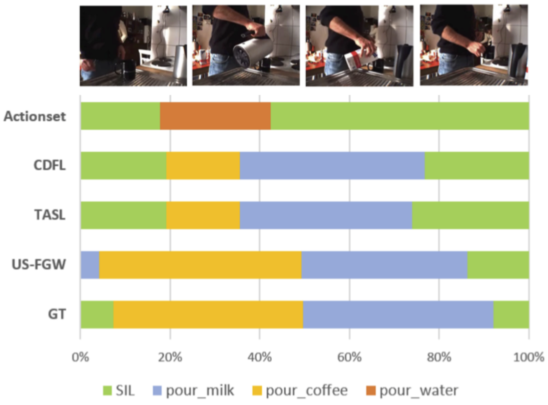

Figure 3: Alignment results of Actionset, CDFL, TASL and our US-FGW (i.e., deterministic autoencoders learned by solving (10)) against the ground-truth of the video "P14_stereo_P14_coffee" in the Breakfast dataset.

5.3 Ablation study and robustness analysis

To demonstrate the usefulness of our US-FGW distance and the corresponding contrastive learning framework, we compare our US-FGW method with its variants on the Hollywood Extended dataset, as shown in Table 2. In particular, when we learn our autoencoders via solving (4), we just consider the unbalanced FGW distance

Table 2: Ablation study of our US-FGW method on the Hollywood Extended dataset.

Method	Mof	Mof-bg	IoU	IoD
Solving (4)	0.3422	0.3685	0.1957	0.3666
Solving (6)	0.3732	0.4067	0.2101	0.3780
Solving (10) (**Proposed**)	**0.3840**	**0.4082**	**0.2307**	**0.4001**

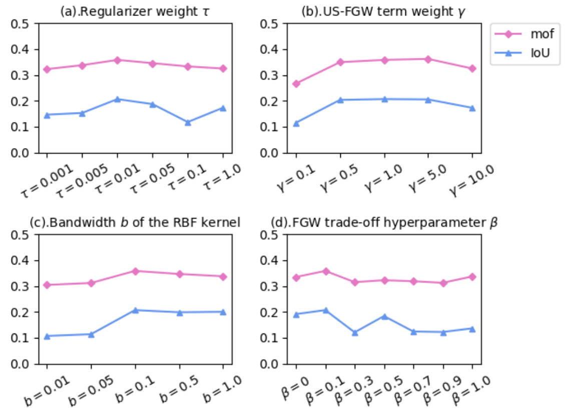

Figure 4: The influences of four key hyperparameters.

defined on the distance matrices and do not use the contrastive learning framework. This variant simplifies our method but leads to degraded performance at the same time because the distance matrices are not so robust as the kernel matrices and the influence of negative text sets is not considered. When solving (6), we use the proposed US-FGW distance defined on the kernel matrices. It works better than solving (4), which demonstrates the necessity of using kernel matrices. The proposed method solves (10), which further considers the contrastive learning framework based on the US-FGW distance. As shown in Table 2, the proposed method outperforms the above two variants consistently. The above analytic results verify the rationality and the superiority of our method.

Besides the above ablation study, we further test the influence of four key hyperparameters on our US-FGW method: *i*) the weight τ of the unbalanced regularizer, *ii*) the weight γ of the contrastive US-FGW term, *iii*) the bandwidth b of the RBF kernel, and *iv*) the hyperparameter β controlling the trade-off between the Wasserstein term and the GW term in the US-FGW distance.

We visualize the Mof and IoU achieved by our method for the Hollywood Extended dataset with respect to these four hyperparameters, as shown in Figure 4, and demonstrate the robustness of our method to them. Firstly, the τ indicates the significance of the unbalanced term in the US-FGW distance. Figure 4(a) shows that our method achieves the best performance when $\tau = 0.01$ and its performance is relatively stable when $\tau \in [0.001, 1]$. Secondly, the γ balances the reconstruction loss term and the US-FGW contrastive term. When we set $\gamma \in [0.5, 5]$, our method performs well, as shown in Figure 4(b). When γ is too small, i.e., $\gamma = 0.1$, the contrastive term will be too weak to regularize the learning of the autoencoders, which leads to the over-fitting issue. Thirdly, for the bandwidth b of the RBF kernel, we find that when $b \in [0.1, 1]$, our method works well in general, as shown in Figure 4(c). When the bandwidth is too small, the similarity between arbitrary two latent codes will tend to zero, which does harm to the discriminative power of the corresponding kernel matrix. As a result, the optimal transport (and accordingly, the matching result) learned by our method becomes unreliable. Finally, the $\beta \in [0, 1]$ controls the proportions between Wasserstein term and GW term in the US-FGW distance. Figure 4(d) shows that our method achieves the best performance when $\beta = 0.1$. Note that, when $\beta = 0$ or 1, our US-FGW distance degrades to the unbalanced spectral Wasserstein distance or the unbalance spectral GW distance, which just considers either the point-wise or the pair-wise similarity between the visual and the textual latent codes. As shown in Figure 4, these two variants are inferior to our US-FGW method.

6 CONCLUSION

In this work, we have proposed a new weakly-supervised temporal action alignment method based on computational optimal transport methods. The proposed method leverages a novel US-FGW distance to capture the correspondence between frames and textual descriptions in their latent space and formulates a new contrastive learning paradigm based on this distance. Experimental results show that our method achieves encouraging performance for set-supervised temporal action alignment. In the aspect of computation, the computational complexity of US-FGW distance is $O\left(I^2J + J^2I\right)$, and the learning process involves nested alternating optimization. Therefore, when dealing with long videos with hundreds of actions, we need to further improve the efficiency of our method. In the aspect of modeling, the current framework treats the visual and textual modalities evenly. However, for temporal action alignment, the visual information of video frames is much more representative than the textual tags, and the tags are very sparse. Taking such imbalance into account may help us to train a better model. In the future, we plan to explore the theoretical property of the new contrastive learning paradigm and apply our method to other applications, e.g., large-scale multi-modal learning and other video tagging tasks.

ACKNOWLEDGMENTS

Dr. Dixin Luo was supported in part by the National Natural Science Foundation of China (No. 62102031); the Young Scholar Programme (No. XSQD-202107001) and the Technology Innovation Programme (No. 3070012212116) from Beijing Institute of Technology, and the National Key Research and Development Program of China funded by the Ministry of Science and Technology (No. 2020YFF0305200). Dr. Hongteng Xu was supported by the National Natural Science Foundation of China (No. 62106271), Beijing Outstanding Young Scientist Program (No. BJJWZYJH012019100020098), and the Research Funds of Renmin University of China (the Fundamental Research Funds for the Central Universities). He also would like to thank the supports from the Beijing Key Laboratory of Big Data Management and Analysis Methods, the Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, and the Public Policy and Decision-making Research Lab of RUC.

REFERENCES

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*. 5803–5812.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 214–223.

[3] Himadri Bhuyan, Partha Pratim Das, Jatindra Kumar Dash, and Jagadeesh Killi. 2021. An Automated Method for Identification of Key frames in Bharatanatyam Dance Videos. *IEEE Access* 9 (2021), 72670–72680.

[4] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. 2014. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*. Springer, 628–643.

[5] Nicolas Bonneel and David Coeurjolly. 2019. Spot: sliced partial optimal transport. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–13.

[6] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. 2021. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8012–8021.

[7] Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2018. Improving Sequence-to-Sequence Learning via Optimal Transport. In *International Conference on Learning Representations*.

[8] Samir Chowdhury and Facundo Mémoli. 2019. The gromov–wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA* 8, 4 (2019), 757–787.

[9] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* 26 (2013).

[10] Chinh Dang and Hayder Radha. 2015. RPCA-KFE: Key frame extraction for video using robust principal component analysis. *IEEE Transactions on Image Processing* 24, 11 (2015), 3742–3753.

[11] Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. 2018. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3483–3491.

[12] Li Ding and Chenliang Xu. 2018. Weakly-supervised action segmentation with iterative soft boundary assignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6508–6516.

[13] Keval Doshi and Yasin Yilmaz. 2022. Rethinking Video Anomaly Detection-A Continual Learning Approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3961–3970.

[14] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. 2021. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3354–3363.

[15] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. 2015. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*. 258–267.

[16] Yazan Abu Farha and Jurgen Gall. 2019. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3575–3584.

[17] Mohsen Fayyaz and Jurgen Gall. 2020. Sct: Set constrained temporal transformer for set supervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 501–510.

[18] Valentin Gabeur, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2022. Masking Modalities for Cross-modal Video Retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1766–1775.

[19] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 297–304.

[20] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

[21] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 603–612.

[22] Yu-Gang Jiang, Zuxuan Wu, Jinhui Tang, Zechao Li, Xiangyang Xue, and Shih-Fu Chang. 2018. Modeling multimodal clues in a hybrid deep learning framework for video classification. *IEEE Transactions on Multimedia* 20, 11 (2018), 3137–3147.

[23] Svebor Karaman, Lorenzo Seidenari, and Alberto Del Bimbo. 2014. Fast saliency based pooling of fisher encoded dense trajectories. In *ECCV THUMOS Workshop*, Vol. 1. 5.

[24] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[25] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[26] Hilde Kuehne, Ali Arslan, and Thomas Serre. 2014. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 780–787.

[27] Hilde Kuehne, Juergen Gall, and Thomas Serre. 2016. An end-to-end generative framework for video segmentation and recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–8.

[28] Hilde Kuehne, Alexander Richard, and Juergen Gall. 2017. Weakly supervised learning of actions from transcripts. *Computer Vision and Image Understanding* 163 (2017), 78–89.

[29] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. 2017. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 156–165.

[30] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager. 2016. Segmental spatiotemporal cnns for fine-grained action segmentation. In *European Conference on Computer Vision*. Springer, 36–52.

[31] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. 2021. Weakly-supervised temporal action localization by uncertainty modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1854–1862.

[32] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. 2017. MMD GAN: towards deeper understanding of moment matching network. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2200–2210.

[33] Jun Li, Peng Lei, and Sinisa Todorovic. 2019. Weakly supervised energy-based learning for action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6243–6251.

[34] Jun Li and Sinisa Todorovic. 2020. Set-constrained viterbi for set-supervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10820–10829.

[35] Jun Li and Sinisa Todorovic. 2021. Anchor-constrained viterbi for set-supervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9806–9815.

[36] Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. 2020. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE transactions on pattern analysis and machine intelligence* (2020).

[37] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. 2020. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4463–4472.

[38] Xiang Long, Chuang Gan, Gerard Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen. 2018. Multimodal keyless attention fusion for video classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[39] Zijia Lu and Ehsan Elhamifar. 2021. Weakly-Supervised Action Segmentation and Alignment via Transcript-Aware Union-of-Subspaces Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8085–8095.

[40] Dixin Luo, Hongteng Xu, and Lawrence Carin. 2020. Hierarchical optimal transport for robust multi-view learning. *arXiv preprint arXiv:2006.03160* (2020).

[41] Facundo Mémoli. 2011. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics* 11, 4 (2011), 417–487.

[42] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 271–279.

[43] Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning* 11, 5-6 (2019), 355–607.

[44] Gabriel Peyré, Marco Cuturi, and Justin Solomon. 2016. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*. PMLR, 2664–2672.

[45] Lionel Pigou, Aäron Van Den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. 2018. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision* 126, 2 (2018), 430–439.

[46] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. 2021. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2806–2814.

[47] Alexander Richard, Hilde Kuehne, and Juergen Gall. 2017. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 754–763.

[48] Alexander Richard, Hilde Kuehne, and Juergen Gall. 2018. Action sets: Weakly supervised action segmentation without ordering constraints. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 5987–5996.

[49] Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. 2018. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

7386–7395.

[50] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012. A database for fine grained activity detection of cooking activities. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 1194–1201.

[51] Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. 2021. The unbalanced Gromov Wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems* 34 (2021).

[52] Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. 2021. Generic event boundary detection: A benchmark for event segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8075–8084.

[53] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*. Springer, 510–526.

[54] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. 2016. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1961–1970.

[55] Dipika Singhania, Rahul Rahaman, and Angela Yao. 2021. Coarse to fine multi-resolution temporal convolutional network. *arXiv preprint arXiv:2105.10859* (2021).

[56] Richard Sinkhorn and Paul Knopp. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.* 21, 2 (1967), 343–348.

[57] Vayer Titouan, Nicolas Courty, Romain Tavenard, and Rémi Flamary. 2019. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*. PMLR, 6275–6284.

[58] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. 2018. Wasserstein Auto-Encoders. In *International Conference on Learning Representations*.

[59] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2018. Deep Graph Infomax. In *International Conference on Learning Representations*.

[60] Huahua Wang and Arindam Banerjee. 2014. Bregman alternating direction method of multipliers. *Advances in Neural Information Processing Systems* 27 (2014).

[61] Michael Wray, Hazel Doughty, and Dima Damen. 2021. On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3650–3660.

[62] Hongteng Xu. 2020. Gromov-Wasserstein factorization models for graph clustering. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 04 (2020), 6478–6485.

[63] Hongteng Xu, Dixin Luo, and Lawrence Carin. 2019. Scalable gromov-wasserstein learning for graph partitioning and matching. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 3052–3062.

[64] Hongteng Xu, Dixin Luo, Ricardo Henao, Svati Shah, and Lawrence Carin. 2020. Learning autoencoders with relational regularization. In *International Conference on Machine Learning*. PMLR, 10576–10586.

[65] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. 2019. Gromov-wasserstein learning for graph matching and node embedding. In *International conference on machine learning*. PMLR, 6932–6941.

[66] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. 2018. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 585–601.

[67] Renjun Xu, Pelen Liu, Yin Zhang, Fang Cai, Jindong Wang, Shuoying Liang, Heting Ying, and Jianwei Yin. 2020. Joint Partial Optimal Transport for Open Set Domain Adaptation.. In *IJCAI*. 2540–2546.

[68] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. 2021. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11562–11572.

[69] Weilong Yang and George Toderici. 2011. Discriminative tag learning on youtube videos with latent sub-tags. In *CVPR 2011*. IEEE, 3217–3224.

[70] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. 2018. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision* 126, 2 (2018), 375–389.

[71] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. 2016. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2678–2687.

[72] Ruiyi Zhang, Changyou Chen, Xinyuan Zhang, Ke Bai, and Lawrence Carin. 2020. Semantic matching via optimal partial transport. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 212–222.

[73] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3537–3545.

Algorithm 1 US-FGW Temporal Action Alignment

```
Require: Data $\mathcal{D} = \{(\mathcal{V}_n, \mathcal{W}_n)\}_{n=1}^{N}$, a vocabulary set $\mathcal{W}_{\text{all}}$, and
    hyper-parameters $\beta, \tau, \gamma$.
for $m = 0, ..., M-1$ (Outer Loop) do
    Sample $\{\mathcal{V}_n, \mathcal{W}_n\}$ randomly from $\mathcal{D}$.
    Construct a negative set $\mathcal{W}_n'$ randomly from $\mathcal{W}_{\text{all}} \setminus \mathcal{W}_n$.
    Compute US-FGW distances (Inner Loop):
        Obtain $T_n^+ \leftarrow d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}_n))$ via (14)-(18).
        Obtain $T_n^- \leftarrow d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}_n'))$ via (14)-(18).
    Update autoencoders:
        Plug $\{T_n^+, T_n^-\}$ into (10) and update $f_v, f_w, g_v, g_w$ via Adam.
end for
```

A BREGMAN ADMM ALGORITHM

In the beginning, the dual variables are initialized as a zero matrix and two zero vectors, while $\boldsymbol{T} = \boldsymbol{S} = \frac{1}{IJ}\mathbf{1}_{I\times J}$, $\boldsymbol{u}$ and $\boldsymbol{\mu}$ are random vectors on the Simplex. At the k-th iteration, we rewrite (13) in the following the Bregman-augmented Lagrangian form for $\boldsymbol{T}$ and update $\boldsymbol{T}$ in a closed form:

$$\begin{aligned}\boldsymbol{T}^{(k+1)} &= \arg\min\nolimits_{\boldsymbol{T}\in\Pi(\boldsymbol{u}^{(k)},\cdot)} (\beta-1)\langle \boldsymbol{K}_{vw}, \boldsymbol{T}\rangle - \beta\langle \boldsymbol{K}_v \boldsymbol{S}^{(k)} \boldsymbol{K}_w^T, \boldsymbol{T}\rangle \\ &\qquad + \langle \boldsymbol{Z}^{(k)}, \boldsymbol{T} - \boldsymbol{S}^{(k)}\rangle + \rho \mathrm{KL}(\boldsymbol{T}\|\boldsymbol{S}^{(k)}) \\ &= \mathrm{diag}(\boldsymbol{u}^{(k)})\sigma_{\mathrm{r}}\left(\frac{(1-\beta)\boldsymbol{K}_{vw} + \beta\boldsymbol{K}_v\boldsymbol{S}^{(k)}\boldsymbol{K}_w^T - \boldsymbol{Z}^{(k)}}{\rho} + \log \boldsymbol{S}^{(k)}\right),\end{aligned} \tag{14}$$

where $\Pi(\boldsymbol{u}^{(k)}, \cdot)$ represents a one-side marginal constraint, and σ_{r} applies the softmax operation to each row of matrix.

Similarly, we can consider the Bregman-augmented Lagrangian form for $\boldsymbol{S}$ and update $\boldsymbol{S}$ in a closed form:

$$\begin{aligned}\boldsymbol{S}^{(k+1)} &= \arg\min\nolimits_{\boldsymbol{S}\in\Pi(\cdot,\boldsymbol{\mu}^{(k)})} -\beta\langle \boldsymbol{K}_v^T \boldsymbol{T}^{(k+1)} \boldsymbol{K}_w, \boldsymbol{S}\rangle \\ &\qquad + \langle \boldsymbol{Z}^{(k)}, \boldsymbol{T}^{(k+1)} - \boldsymbol{S}\rangle + \rho\mathrm{KL}(\boldsymbol{S}\|\boldsymbol{T}^{(k+1)}) \\ &= \sigma_{\mathrm{c}}\left(\frac{\beta\boldsymbol{K}_v^T\boldsymbol{T}^{(k+1)}\boldsymbol{K}_w + \boldsymbol{Z}^{(k)}}{\rho} + \log \boldsymbol{T}^{(k+1)}\right)\mathrm{diag}(\boldsymbol{\mu}^{(k)}),\end{aligned} \tag{15}$$

where σ_{c} is the column-wise softmax operation.

Then, the auxiliary variables corresponding to the marginals of $\boldsymbol{T}$ and $\boldsymbol{S}$ are updated as follows:

$$\begin{aligned}&\min\nolimits_{\boldsymbol{u}\in\Delta^{I-1}} \tau\mathrm{KL}(\boldsymbol{u}\|\frac{1}{I}\mathbf{1}_I) + \langle \boldsymbol{z}_1^{(k)}, \boldsymbol{u}\rangle + \rho\mathrm{KL}(\boldsymbol{u}\|\boldsymbol{T}^{(k+1)}\mathbf{1}_J) \\ &\Rightarrow\ \boldsymbol{u}^{(k+1)} = \sigma\left(\frac{\rho\log(\boldsymbol{T}^{(k+1)}\mathbf{1}_J) + \tau\log\frac{1}{I}\mathbf{1}_I - \boldsymbol{z}_1^{(k)}}{\rho+\tau}\right),\end{aligned} \tag{16}$$

$$\begin{aligned}&\min\nolimits_{\boldsymbol{\mu}\in\Delta^{J-1}} \tau\mathrm{KL}(\boldsymbol{\mu}\|\frac{1}{J}\mathbf{1}_J) + \langle \boldsymbol{z}_2^{(k)}, \boldsymbol{\mu}\rangle + \rho\mathrm{KL}(\boldsymbol{\mu}\|(\boldsymbol{S}^{(k+1)})^T\mathbf{1}_I) \\ &\Rightarrow\ \boldsymbol{\mu}^{(k+1)} = \sigma\left(\frac{\rho\log((\boldsymbol{S}^{(k+1)})^T\mathbf{1}_I) + \tau\log\frac{1}{J}\mathbf{1}_J - \boldsymbol{z}_2^{(k)}}{\rho+\tau}\right),\end{aligned} \tag{17}$$

where σ is the softmax operation of vectors.

Finally, we update the dual variables via the ADMM manner:

$$\begin{aligned}\boldsymbol{Z}^{(k+1)} &= \boldsymbol{Z}^{(k)} + \rho(\boldsymbol{T}^{(k+1)} - \boldsymbol{S}^{(k+1)}), \\ \boldsymbol{z}_1^{(k+1)} &= \boldsymbol{z}_1^{(k)} + \rho(\boldsymbol{u}^{(k+1)} - \boldsymbol{T}^{(k+1)}\mathbf{1}_J), \\ \boldsymbol{z}_2^{(k+1)} &= \boldsymbol{z}_2^{(k)} + \rho(\boldsymbol{\mu}^{(k+1)} - (\boldsymbol{S}^{(k+1)})^T\mathbf{1}_I).\end{aligned} \tag{18}$$

Repeating the above steps till $\boldsymbol{T}^{(k)}$ converges, we obtain the optimal transport matrix $\boldsymbol{T}^* = \boldsymbol{T}^{(K)}$. Plugging the optimal transport matrices of all US-FGW distance terms into (10), we then update the autoencoders via stochastic gradient descent (SGD) algorithm like Adam [24]. As a result, our scheme of our learning algorithm is shown in Algorithm 1, where $\boldsymbol{T}^+$ and $\boldsymbol{T}^-$ are the optimal transport matrices from the frame set to its positive text set and to its negative text set, respectively.

B IMPLEMENTATION DETAILS

B.1 The architecture of autoencoders

For all three datasets, the architecture of the visual autoencoder is Encoder:

$$x \in \mathbb{R}^{64} \rightarrow \mathrm{GRU}_{64} \rightarrow \mathrm{FC}_{32} + \mathrm{ReLU} \rightarrow \mathrm{FC}_{16} \rightarrow \mathrm{FC}_{8} \rightarrow z \in \mathbb{R}^{8}$$

Decoder:

$$z \in \mathbb{R}^{8} \rightarrow \mathrm{FC}_{16} + \mathrm{ReLU} \rightarrow \mathrm{FC}_{32} + \mathrm{ReLU} \rightarrow \mathrm{FC}_{64} \rightarrow x \in \mathbb{R}^{64},$$

For Breakfast, the architecture of the textual autoencoder is Encoder:

$$x \in \mathbb{R}^{48} \rightarrow \mathrm{FC}_{32} + \mathrm{ReLU} \rightarrow \mathrm{FC}_{16} \rightarrow \mathrm{FC}_{8} \rightarrow z \in \mathbb{R}^{8}$$

Decoder:

$$z \in \mathbb{R}^{8} \rightarrow \mathrm{FC}_{16} + \mathrm{ReLU} \rightarrow \mathrm{FC}_{32} + \mathrm{ReLU} \rightarrow \mathrm{FC}_{48} \rightarrow x \in \mathbb{R}^{48},$$

For Hollywood Extended, the architecture of the textual autoencoder is

$$\text{Encoder:}\quad x \in \mathbb{R}^{16} \rightarrow \mathrm{FC}_{10} + \mathrm{ReLU} \rightarrow \mathrm{FC}_{8} \rightarrow z \in \mathbb{R}^{8}$$

$$\text{Decoder:}\quad z \in \mathbb{R}^{8} \rightarrow \mathrm{FC}_{10} + \mathrm{ReLU} \rightarrow \mathrm{FC}_{16} \rightarrow x \in \mathbb{R}^{16},$$

For Crosstask, the architecture of the textual autoencoder is Encoder:

$$x \in \mathbb{R}^{80} \rightarrow \mathrm{FC}_{40} + \mathrm{ReLU} \rightarrow \mathrm{FC}_{16} \rightarrow \mathrm{FC}_{8} \rightarrow z \in \mathbb{R}^{8}$$

Decoder:

$$z \in \mathbb{R}^{8} \rightarrow \mathrm{FC}_{16} + \mathrm{ReLU} \rightarrow \mathrm{FC}_{40} + \mathrm{ReLU} \rightarrow \mathrm{FC}_{80} \rightarrow x \in \mathbb{R}^{80},$$

where GRU_k stands for a one-layer gated recurrent unit (GRU) whose hidden state dimension is k, FC_k stands for the fully-connected layer mapping to $\mathbb{R}^k$. When the autoencoders are probabilistic, for each of their encoders, the last layer contains two FC layers, outputting the mean and the logarithmic variance accordingly. Note that, for the dimension of the latent codes, we consider multiple configurations, *i.e.*, choosing it from $\{2, 3, 4, 6, 8, 10, 16, 32, 64\}$. Experimental results show that our US-FGW method achieves the best performance when setting the latent dimension to be 8.

B.2 Length model

In (5), we set a uniform distribution ($\frac{1}{J}\mathbf{1}_J$) to as the prior distribution of texts, which works to regularize the marginal distribution of the optimal transport corresponding to texts. To describe the distribution more accurately, we adopt a simple but effective approach to build a length model for the texts and update the texts' distribution accordingly. In particular, during the training process, we maintain a buffer for each video, whose length is same as the size of the corresponding action set. We initialize the buffer via

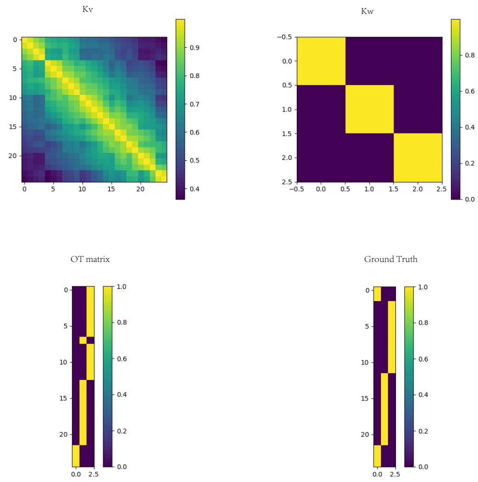

Figure 5: Visualization on distance matrix K_v, K_w, optimal transport matrix T and ground truth matrix of the video "P14_cam01_P14_coffee" in the Breakfast dataset.

$\frac{1}{J}\mathbf{1}_J$. In each epoch, we tag each frame in the video according to current model, and then, obtain the number of frames belonging to each action (a.k.a., the length of the clip of each action). As a result, these numbers provide us with a histogram estimation of the texts' distribution, and thus, we store the normalized numbers in the buffer and use them as the texts' distribution in the next epoch. Since one action usually lasts a similar time, thus with the update of the length model, the prior length distribution can get closer and closer to the real distribution.

B.3 The specific testing methods

In the training process, we reduce the distance between the corresponding frames and words by learning the visual and textual autoencoders. In the testing phase, after mapping the frames and words to the latent space, we derive the matching results by computing an optimal transport matrix between two latent codes. Based on the optimal transport matrix, we align the frames to their corresponding words. It should be noted that besides this testing strategy, the following two testing strategies can be applied to our method as well.

- Taking advantage of the two well-learning autoencoders, we can derive the matching results by the distance matrix $\boldsymbol{K}_{vw}$ directly, whose element $d(\boldsymbol{v}_i, \boldsymbol{w}_j)$ represents the distance between the latent code of the i-th frame and that of the j-th word.
- Additionally, we can feed the visual latent codes into the textual decoder part to get a classification probability for each frame, and then assigning the maximum one as the predicted label to the frame.

Experimental results show that using the OT matrix to infer the actions helps to achieve the best results.

B.4 Inference method

In the testing phase, we sample one frame every 10 frames of the video and compute the OT matrix accordingly. Then, we expand the tags inferred from the OT matrix by a factor of 10. Additionally, to improve the continuity of the predicted actions, we set a threshold value in the inference process. For each predicted clips (a sequence of frames tagged with the same textual description/action), if the number of predicted actions is less than the threshold value, we replace it with the one with more predicted frames among the preceding and following actions.

C VISUALIZATION

Figure 5 displays distance matrix $\boldsymbol{K}_v$, $\boldsymbol{K}_w$, optimal transport matrix $\boldsymbol{T}$ and ground truth matrix of the video "P14_cam01_P14_coffee" in the Breakfast dataset. For distance matrix $\boldsymbol{K}_v$, we can find that it not only reflects the similarity of frames' content but shows the order information between frames. As to distance matrix $\boldsymbol{K}_w$, it only reflects the difference in the symbolical meanings. From the second row of Figure 5, we can compare the optimal transport matrix with the ground truth matrix intuitively. For all three action labels appearing in the video, we can recognize them well and finally achieve the alignment task between frames and words.