Temporal Visual Semantics-Induced Human Motion Understanding With Large Language Models

Zheng Xing, *Member, IEEE*, and Weibing Zhao, *Member, IEEE*

***Abstract*—Unsupervised human motion segmentation (HMS) can be effectively achieved using subspace clustering techniques. However, traditional methods overlook the role of temporal semantic exploration in HMS. This paper explores the use of temporal vision semantics (TVS) derived from human motion sequences, leveraging the image-to-text capabilities of a large language model (LLM) to enhance subspace clustering performance. The core idea is to extract textual motion information from consecutive frames via LLM and incorporate this learned information into the subspace clustering framework. The primary challenge lies in learning TVS from human motion sequences using LLM and incorporating this information into subspace clustering. To address this, we determine whether consecutive frames depict the same motion by querying the LLM and subsequently learn temporal neighboring information based on its response. We then develop a TVS-integrated subspace clustering approach, incorporating subspace embedding with a temporal regularizer that induces each frame to share similar subspace embeddings with its temporal neighbors. Additionally, segmentation is performed based on subspace embedding with a temporal constraint that induces the grouping of each frame with its temporal neighbors. We also introduce a feedback-enabled framework that continuously optimizes subspace embedding based on the segmentation output. Experimental results demonstrate that the proposed method outperforms existing state-of-the-art approaches on four benchmark human motion datasets.**

***Index Terms*—Human motion segmentation, temporal vision semantics, subspace embedding, temporal neighbors.**

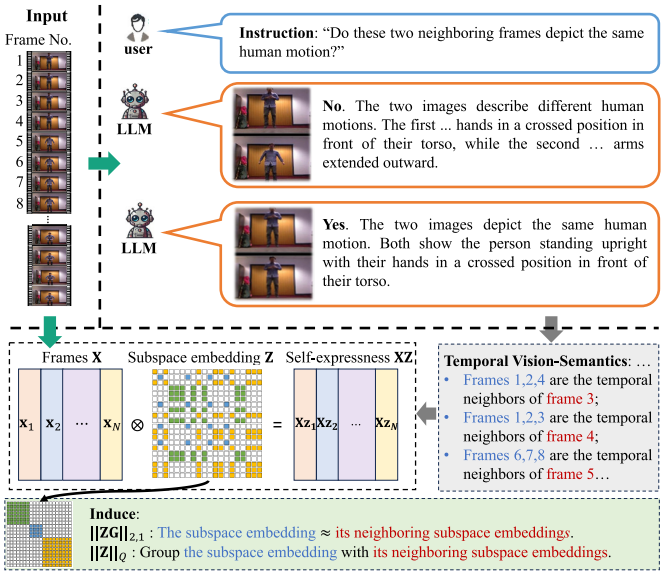


Fig. 1. Framework of the proposed method.

I. Introduction

HUMAN motion segmentation (HMS) has attracted significant attention in both industry and academic research due to its wide-ranging applications in video retrieval, virtual reality, and intelligent surveillance, particularly in human motion analysis [1], [2], [3], [4]. The primary goal of unsupervised HMS is to partition frame sequences depicting human actions into non-overlapping, internally consistent groups without the need for training, serving as a preprocessing step for motion-related analysis tasks [5]. However, unsupervised HMS faces the challenge of motion primitive ambiguity due to temporal variability across different actions [6], [7], [8], [9], [10].

Subspace clustering is a well-established strategy for HMS, aiming to partition a human motion sequence into distinct groups based on the assumption that the frames originate from multiple subspaces, with frames depicting the same motion belonging to the same subspace [11], [12], [13], [14], [15]. In recent years, a popular and effective approach is to first embed the human motion frames into multiple subspaces to learn the subspace structure of the data, and then apply traditional clustering algorithms to the subspace embeddings [16], [17], [18], [19], [20].

Human motion sequences inherently contain temporal information, which is crucial for HMS. As illustrated in Figure 1, a human motion sequence typically consists of multiple motion segments. For example, a person may first perform a motion with both hands clasped for a period, then extend their hands for another period, followed by squatting. However, due to the ambiguity of different motions and the complexity of temporal correlations, extracting temporal information from human motion sequences remains a significant challenge [20], [21].

Received 8 February 2025; revised 5 November 2025 and 24 December 2025; accepted 6 February 2026. Date of publication 19 February 2026; date of current version 24 February 2026. This work was supported in part by Shenzhen Science and Technology Program under Grant JCYJ20241202130548062 and in part by Guangdong Provincial Key Area Project of General Universities under Grant 2024ZDZX1017 and Grant 2025ZDZX3049. The associate editor coordinating the review of this article and approving it for publication was Prof. Nikos Deligiannis. *(Corresponding author: Weibing Zhao.)*

Zheng Xing is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: zhengxing@link.cuhk.edu.cn).

Weibing Zhao is with Guangdong Laboratory of Machine Perception and Intelligent Computing, Faculty of Engineering, Shenzhen MSU-BIT University, Shenzhen 518172, China (e-mail: weibingzhao@smbu.edu.cn).

Digital Object Identifier 10.1109/TIP.2026.3663857

Various subspace clustering algorithms have been proposed to achieve HMS by exploring the temporal information in the data. For instance, Wang et al. [22] eliminate redundant connections between adjacent motions in the subspace embedding to extract informative instance data and capture the compact structure of human motion videos. Bai et al. [10] extract informative features during subspace embedding to capture local temporal consistency in human motions. Zhou et al. [9] employ a multi-mutual consistency learning strategy to factorize source and target data into distinct multi-layer feature spaces, thereby learning the temporal information from the source domain. Despite efforts to explore and utilize temporal information, its inaccuracy and ambiguity may lead to incorrect segmentation results. While neural network algorithms have shown great promise in supervised tasks, they often fall short in unsupervised tasks due to the difficulty in exploring the underlying data structures.

This paper aims to learn temporal vision semantics (TVS) from human motion sequences by leveraging the image-to-text capabilities of a pre-trained large language model (LLM) to enhance HMS performance. The key idea is to extract textual motion information from consecutive frames using the LLM and integrate this learned information into the subspace clustering framework. By incorporating TVS into the subspace clustering, we aim to ensure that the segmentation output more effectively captures the temporal dynamics inherent in the human motion sequence.

Thus, we face the following *challenges*:

- *How to learn TVM using LLM?* Although LLMs are widely used in image-to-text tasks, no research has explored how to leverage LLMs to assist in unsupervised HMS. The challenge lies in learning the textual temporal information that can be converted into a mathematical form, which can be used to induce the HMS.
- *How to integrate TVM with HMS?* We aim to learn the subspace embedding and perform segmentation based on this embedding, all induced by TVM. However, integrating TVM into both the subspace embedding and segmentation presents significant challenges.

In this study, we determine whether consecutive frames represent the same motion by querying the LLM. Based on its response, we subsequently learn the temporal relationships between frames, as illustrated in Figure 1. Building upon this, we propose a subspace clustering approach integrated with TVS, which combines subspace embedding with a temporal regularizer. This regularizer ensures that each frame shares similar subspace embeddings with its temporal neighbors. Segmentation is then performed using these subspace embeddings, with a temporal constraint that encourages the grouping of each frame with its temporal neighbors. Furthermore, we introduce a feedback-enabled framework that iteratively optimizes the subspace embeddings based on the segmentation output, ensuring continuous refinement of the model.

In summary, the main *contributions* are as follows:

- *Exploring TVS via LLMs and Integrating TVS with HMS*: This paper introduces an approach that uses LLMs to learn TVS in human motion sequences. We develop a method that applies TVS to both subspace embedding and segmentation, ensuring neighboring consistency in both processes.
- *Feedback-enabled Subspace Embedding*: We propose a feedback-enabled strategy that allows the segmentation output to inform the subspace embedding. This is not merely a combination of two methods; rather, it enables the use of HMS output to induce subspace embedding, better capturing the underlying subspace structure in the human motion sequence.

We conduct extensive experiments on four benchmark datasets for HMS. The experimental results consistently demonstrate that our method outperforms existing state-of-the-art techniques, highlighting its superiority in HMS.

The remainder of this paper is structured as follows. Section II briefly introduces the related work including HMS and subspace clustering. Section III presents the details of the proposed method. Section IV provides the experimental settings, performance comparisons, and ablation study. Finally, Section V concludes the paper.

II. Related Works

A. Human Motion Segmentation

HMS is essential for accurately capturing human motion data, forming a basis for structural analysis, understanding, and practical applications. Significant research efforts have led to notable achievements in this area. For instance, Zhong et al. [23] proposed a bipartite graph co-clustering framework to segment unusual activities in videos. Fod et al. [24] utilized zero-velocity crossing frames of angular velocity to partition motion data streams into different sequences. Barbic et al. [25] employed probabilistic principal component analysis to decompose human motion into distinct motions. Additionally, Beaudoin et al. [26] introduced a framework for distilling a motion-motif graph from motion data collections. Spatio-temporal-based Convolutional Neural Networks (CNNs) [27] and clustering-based approaches [28] have been proposed for segmenting streams of human motion into multiple activities. Despite the capability of deep learning-based motion recognition models to complete HMS tasks by training with large datasets, the unsupervised model offers significant advantages in terms of interpretability and computational efficiency. Therefore, achieving HMS tasks through an unsupervised approach is highly beneficial.

However, these approaches may not fully exploit the temporal dynamics and semantic continuity inherent in human motion sequences, potentially limiting the accuracy and effectiveness of the segmentation results.

B. Subspace Clustering

Subspace clustering discerns and segregates distinct motion types into their respective subspaces, thereby addressing human motion segmentation tasks. Its capacity to manage intricate, high-dimensional motion data, combined with robustness to noise and data variability, ensures reliability in practical applications. Furthermore, by exploiting the inherent

low-dimensional structures within complex motion datasets, subspace clustering facilitates the further analysis and utilization of human motion data.

Subspace clustering serves to ascertain the low-dimensional embedding of a high-dimensional manifold. Specifically, assuming that vectorized frames corresponding to identical actions reside within the same subspace, the subspace embedding property inherent to the data can be harnessed to derive a representative subspace embedding [6], [9], [10], [22], [29], [30], [31]. Subspace clustering has recently gained significant attention due to its effectiveness in uncovering complex data structures and improving clustering performance in high-dimensional spaces. For instance, the *SIBMSC* method [16] extends the information bottleneck principle to learn view-common representations, removing redundant information and leveraging mutual information for view-specific clustering. Similarly, the *BTMSC* method [17] constructs a third-order tensor to capture high-order correlations, using the Bi-Nuclear Quasi-Norm for efficient tensor factorization. To improve robustness, the *FSMSC* method [18] integrates view-shared anchor learning with a self-guided discriminative feature selection approach, addressing noisy views and cross-view diversity. The *ARLRR* method [19] introduces affine and non-negative constraints in low-rank self-representation learning to manage affine subspaces and errors. The *DCTMSC* method [20] employs a two-step discrete cosine transform approach to simplify tensor nuclear norm calculations and enhance local structural representation. The *DCMVC* method [32] incorporates dynamic cluster diffusion and reliable neighbor-guided positive alignment to improve inter-cluster separation and within-cluster compactness. Subspace clustering methods incorporating temporal priors have proven effective in HMS tasks. For instance, the *OSC* method [6] applies a one-neighbor consistency constraint for closer representations of temporal data, while the *TSC* method [7] uses non-negative dictionary learning and temporal Laplacian regularization. The *LTS* method [8] captures temporal correlations in both source and target data with a graph regularizer and introduces a weighted low-rank constraint to reveal clustering structures [33]. The *CDMS* approach [9] leverages transfer subspace learning to capture multi-level information in videos. These methods, often formulated as unsupervised learning frameworks, typically adopt a self-representation strategy for motion segmentation. The *DSAE* method [34] enhances representation learning by considering temporal correlations, while the *VSDA* method [10] employs a multi-neighbor auto-encoder to extract temporal features and a long-short distance enembedding/deembedding strategy to maintain representation consistency, further enhanced by a velocity-sensitive guidance mechanism.

However, existing subspace clustering approaches typically divide the process into two independent stages and often overlook the potential of incorporating temporal semantics to simultaneously enhance both subspace embedding and clustering. This oversight limits the alignment of the HMS output with the true sequential dynamics of human motion.

C. Temporal Vision Semantics From Large Language Models

LLMs have progressed from text-only reasoning engines to unified multimodal systems capable of jointly understanding visual and linguistic information. Recent architectures such as GPT-4o, Gemini, Claude, DeepSeek, and Qwen3 integrate visual encoders with transformer-based text reasoning via large-scale contrastive pretraining, enabling them to perform complex cross-modal reasoning and semantic alignment between images and language. Unlike conventional convolutional or transformer-based visual encoders that rely on geometric or pixel-level similarity, multimodal LLMs exhibit *semantic reasoning capability*. They can compare two visual scenes and judge whether they convey the same conceptual meaning based on high-level world knowledge and contextual understanding.

Extensive research in vision–language modeling has demonstrated the strong semantic reasoning capabilities of LLMs when integrated with visual inputs. Early studies validated these capabilities in tasks such as zero-shot visual question answering and high-fidelity caption generation, where LLMs interpret visual entities, relationships, and contextual meanings directly from raw imagery [35], [36]. Building upon these foundations, subsequent works extended vision semantics to three-dimensional and dynamic scenes, leveraging language-guided scene understanding and position-aware video representations for 3D perception [37], [38]. Beyond visual applications, LLMs have also exhibited robust zero-shot reasoning and representational alignment abilities that enable general semantic understanding across modalities [39], [40].

Building upon this progress, a growing body of research has explored the integration of LLM-based visual semantics into temporal vision understanding. Recent work has examined whether video-oriented LLMs truly capture temporal reasoning or merely rely on knowledge and spatial perception [41], while subsequent studies have demonstrated that LLMs can effectively learn temporal dependencies and causal relations across video frames [42]. Further developments employ language-guided attention mechanisms to align visual dynamics with textual motion cues, thereby enhancing spatial–temporal object understanding and fine-grained temporal reasoning [43], [44]. Beyond short-term video grounding [45], recent efforts extend this semantic alignment to long-term sequence modeling, revealing that LLMs can encode cross-frame dependencies with human-level temporal abstraction [46], [47].

In the domain of human motion analysis, LLMs have been increasingly adopted to reason about human activities and their semantic transitions [48], [49], [50], [51]. These studies demonstrate that LLMs can interpret and describe motion in natural language, distinguish subtle phase changes, and correlate sensor or visual signals with linguistic motion descriptions. Recent studies have advanced from graph-based relational modeling to LLM-driven semantic reasoning in motion understanding. For example, some works employ LLMs to anticipate long-term actions by treating video frames as language-like tokens and enhancing vision–language interaction through cross-modal reasoning [52], while others utilize

graph attention mechanisms to capture individual–group interaction dynamics in collective activities [53]. Such findings inspire the present work, where we employ LLM-based reasoning to construct temporal semantics, serving as a high-level inductive prior for unsupervised human motion segmentation.

III. Methodology

In this section, we first introduce an LLM-based inference to identify the TVS. We then propose a feedback-enabled subspace embedding approach that incorporates TVS to efficiently determine the HMS with limited iterations.

A. LLM-Driven Temporal Semantics Inference

Human behavior unfolds as a continuous visual process, where adjacent frames in a motion sequence often exhibit high semantic correlation. To identify segments representing the same human action, we introduce the concept of TVS, which delineates the temporal neighborhood of each frame according to semantic consistency. Specifically, for a given frame $\mathbf{x}_i$, we aim to discover its left and right temporal neighbor bounds (l_i, r_i) that enclose all frames depicting the same motion as $\mathbf{x}_i$.

Machine learning techniques have been widely adopted across a broad range of application domains [54], [55], [56], [57], [58], [59]. However, unsupervised discovery of such temporal neighborhoods is challenging using traditional machine learning methods, as the semantic boundary between motions is difficult to define purely through pixel comparisons [60], [61], [62], [63], [64], [65], [66], [67], [68]. To address this challenge, we harness the visual reasoning capability of a LLM as a zero-shot semantic comparator. Instead of training an additional network, the LLM is instructed with a natural-language prompt to assess whether two consecutive frames represent the same human motion:

> *"Do these two neighboring frames depict the same human motion? Answer Yes or No."*

Given two adjacent frames $(\mathbf{x}_t, \mathbf{x}_{t+1})$, the LLM produces a binary response `Yes`/`No`, which is recorded as a Boolean variable $\text{eq}_t \in \{0, 1\}$ indicating whether the two frames belong to the same motion segment. The sequence $\{\text{eq}_1, \ldots, \text{eq}_{N-1}\}$ forms the adjacency pattern of temporal consistency across the sequence. In our implementation, this is achieved through an API call to a multimodal LLM (e.g., GPT-4o, Gemini-2.0, or Claude-4.5), where both frames are provided in base64-encoded format, allowing the model to reason directly over image content.

Using the response $\{\text{eq}_1, \ldots, \text{eq}_{N-1}\}$, we define the left and right temporal neighbor bounds for each frame $\mathbf{x}_i$ as follows:

$$l_i = \min\{j \mid j \le i,\ \mathbf{x}_j, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_i \text{describe the same motion}\},$$
$$r_i = \max\{j \mid j \ge i,\ \mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_j \text{describe the same motion}\}.$$

Accordingly, the temporal neighborhood set $\mathcal{N}_i$ of $\mathbf{x}_i$ is given by

$$\mathcal{N}_i = \{j \mid j \in \{l_i, l_i + 1, \ldots, r_i\},\ j \ne i, j \in \mathbb{Z}^+\}.$$

Each frame $\mathbf{x}_i$ is thus associated with a temporally and semantically coherent segment.

This structure serves as a foundation for constructing a TVS matrix $\mathbf{G} \in \mathbb{R}^{N \times N}$, whose entries encode the neighborhood connectivity as

$$G_{i,j} = \begin{cases} -|\mathcal{N}_i|, & \text{if } i = j, \\ 1, & \text{if } j \in \mathcal{N}_i, \\ 0, & \text{otherwise.} \end{cases}$$

In practice, the TVS matrix is implemented as a Laplacian-like structure, where the diagonal term penalizes the number of semantic neighbors, and the off-diagonal entries reflect temporal affinity.

Algorithm 1 Learning TVS via LLM

Input: Raw RGB frames sequence $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.
Output: TVS matrix $\mathbf{G} \in \mathbb{R}^{N \times N}$ and the set $\{\mathcal{N}_i\}_{i=1}^{N}$

Initialize eq $\leftarrow \varnothing$;
for $i = 1$ **to** $N - 1$ **do**
  Encode images $\mathbf{x}_i, \mathbf{x}_{i+1}$ to base64 formats separately;
  Query LLM with prompt: *"Do these two neighbouring frames depict the same human motion? Answer Yes or No."*;
  Parse response to $\text{eq}_i \in \{0, 1\}$ (`Yes`$\rightarrow$1, `No`$\rightarrow$0).
  if response ambiguous **then** re-query with stricter instruction (*"Answer strictly with a single token: YES or NO."*) and re-parse.
  Append eq_i to list eq.
end for
for $i = 1$ **to** N **do**
  Initialize $l_i \leftarrow i$; $r_i \leftarrow i$.
  while $l_i > 1$ **and** $\text{eq}_{l_i - 1} = 1$ **do** $l_i \leftarrow l_i - 1$.
  while $r_i < N$ **and** $\text{eq}_{r_i} = 1$ **do** $r_i \leftarrow r_i + 1$.
end for
Initialize $\mathbf{G} \leftarrow \mathbf{0}_{N \times N}$.
for $i = 1$ **to** N **do**
  $\mathcal{N}_i \leftarrow \{j \mid j \in [l_i, r_i],\ j \ne i\}$;
  $G_{ii} \leftarrow -|\mathcal{N}_i|$;
  for each $j \in \mathcal{N}_i$ **do**
    $G_{ij} \leftarrow 1$.
end for

Algorithm 1 summarizes the complete computational procedure. For each pair of adjacent frames $(i, i + 1)$, the LLM is queried once and the response recorded. Subsequently, left and right neighbor bounds (l_i, r_i) are determined through a recursive traversal of the Boolean adjacency list. Finally, the TVS matrix $\mathbf{G}$ is constructed and saved for downstream processing.

Remark 1: While the TVS introduces human-like temporal reasoning into motion segmentation, it only captures pairwise relationships between temporal consecutive frames, lacking global temporal dependencies. Therefore, a subsequent grouping stage is required to achieve globally consistent motion segmentation.

B. Subspace Embedding and Clustering Incorporating Vision Temporal Semantics

By leveraging matrix multiplication, we observe that the product $\mathbf{ZG}$ captures the similarity error between the

representation of a given sequential point and its neighbors. Specifically, the term

$$\mathbf{ZG} = \left[\sum_{l\in\mathcal{N}_1}(\mathbf{z}_1 - \mathbf{z}_l), \sum_{l\in\mathcal{N}_2}(\mathbf{z}_2 - \mathbf{z}_l), \ldots, \sum_{l\in\mathcal{N}_N}(\mathbf{z}_N - \mathbf{z}_l) \right].$$

measures the similarity of the ith data and its neighbors defined by $\mathcal{N}_i$. To encourage the subspace embedding of the current frame and the embedding of its neighbors to be as similar as possible, we introduce a structural regularization term $\|\mathbf{ZG}\|_{2,1}$, where $\|\cdot\|_{2,1}$ denotes the l_1 norm of the vector formed by the l_2 norms of each column of the matrix. This norm encourages the columns of $\mathbf{ZG}$ to exhibit consistent behavior across neighboring points, promoting smoothness in the subspace representation. Mathematically, we express it as

$$\|\mathbf{ZG}\|_{2,1} = \sum_{i=1}^{N} \left\| \sum_{l\in\mathcal{N}_i}(\mathbf{z}_i - \mathbf{z}_l) \right\|_2 = \sum_{i=1}^{N}\sum_{l\in\mathcal{N}_i} \|\mathbf{z}_i - \mathbf{z}_l\|_2.$$

We aim to minimize the subspace embedding error $\|\mathbf{ZG}\|_{2,1}$ to enhance the temporal consistency of the subspace embedding by promoting coherence between a subspace embedding and its neighboring subspace embeddings, which is crucial for capturing the dynamic temporal structure of the data.

Theorem 1 (Interpretation of the TVS Regularizer): The regularizer $\|\mathbf{ZG}\|_{2,1}$ represents an *isotropic graph total variation* over the temporal graph defined by $\{\mathcal{N}_i\}$. Minimizing it enforces local smoothness within temporal neighborhoods while preserving discontinuities at motion boundaries, thus producing piecewise-constant embeddings consistent with human motion transitions. (Proof See Appendix A.) □

Theorem 2: [Consistency under Noisy LLM Adjacency] If each LLM adjacency label is independently flipped with probability $p < \frac{1}{2}$ and each motion segment has length at least $L_{\min}$, then the expected number of erroneous TVS boundaries scales as $O(pN)$. Minimizing $\|\mathbf{ZG}\|_{2,1}$ yields piecewise-constant embeddings that smooth out isolated errors, ensuring segment-level consistency in expectation when p is small and $L_{\min}$ is sufficiently large. (Proof See Appendix B.) □

Theorem 2 implies that the proposed framework is robust to occasional LLM misjudgments: although local adjacency errors may occur, the TVS-induced regularizer preserves overall temporal coherence by enforcing smooth embeddings within segments. Thus, the method maintains consistent human motion segmentation under moderate annotation noise.

We also propose a TVS-integrated segmentation on the subspace embedding. Specifically, suppose the number of clusters is K. We introduce a cluster assignment indicator vector $\mathbf{q}_i \in \mathbb{R}^K$ for the i-th frame, where the k-th element is set to 1 if the i-th frame is assigned to the k-th cluster, and all other elements are set to zero. We then define a clustering regularizer based on the subspace embedding $\mathbf{Z}$ and the indicator matrix $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_N] \in \mathbb{R}^{N\times K}$:

$$\begin{aligned}\|\mathbf{Z}\|_{\mathbf{Q}} &= \frac{1}{2}\sum_{i,j}\frac{|Z_{i,j}| + |Z_{j,i}|}{2}\|\mathbf{q}_i - \mathbf{q}_j\|_2^2 \\ &= \frac{1}{2}\sum_{i,j}|Z_{i,j}|\cdot\|\mathbf{q}_i - \mathbf{q}_j\|_2^2 = \sum_{i,j}|Z_{i,j}\cdot\frac{\|\mathbf{q}_i - \mathbf{q}_j\|_2^2}{2}| \\ &= \sum_{i,j}|Z_{i,j}\cdot\Theta_{i,j}| = \sum_{i,j}|(\mathbf{\Theta}\odot\mathbf{Z})_{ij}| = \|\mathbf{\Theta}\odot\mathbf{Z}\|_1\end{aligned}$$

where $Z_{i,j}$ is the (i,j)-th element of $\mathbf{Z}$, and $\Theta_{i,j} = \frac{\|\mathbf{q}_i - \mathbf{q}_j\|_2^2}{2}$. The first equation ensures symmetry by combining both $|Z_{i,j}|$ and $|Z_{j,i}|$, accounting for the interactions between off-diagonal terms, as $Z_{i,j} \neq Z_{j,i}$ does not necessarily hold, thus incorporating these contributions into the final regularizer. The term $\Theta_{i,j}$ measures the squared Euclidean distance between the cluster indicators $\mathbf{q}_i$ and $\mathbf{q}_j$, normalized by a factor of $\frac{1}{2}$, capturing the dissimilarity between frames based on their clustering assignments and enforcing smoothness within clusters. The final expression $\|\mathbf{\Theta}\#\odot\mathbf{Z}\|_1$ represents the l_1-norm of the element-wise product between $\mathbf{\Theta}$ and $\mathbf{Z}$, which encourages a sparse representation of the subspace embedding while optimizing the clustering assignments, ensuring that frames assigned to the same cluster exhibit more similar representations. The term $\|\mathbf{Z}\|_{\mathbf{Q}}$ will be minimized to optimize the clustering assignment.

To ensure that $\mathbf{q}_i$ functions effectively as a cluster assignment indicator, we impose the condition that $\mathbf{Q}$ is a subset of

$$\mathcal{Q} = \left\{\mathbf{Q} \in \{0,1\}^{N\times K} : \mathbf{Q}\mathbf{1}_{K\times 1} = \mathbf{1}_{N\times 1}, \mathbf{q}_i = \mathbf{q}_j \,\forall\, j \in \mathcal{N}_i\right\}.$$

This constraint ensures that the clustering assignment for the i-th frame and its neighbors are identical, i.e., $\mathbf{q}_i = \mathbf{q}_j \,\forall\, j \in \mathcal{N}_i$, which enforces temporal consistency within cluster assignments.

Building on the traditional subspace clustering formulation in [69], we develop a feedback-enabled framework that integrates the proposed subspace embedding, which incorporates temporal vision semantics, with the proposed clustering method. The optimization problem is formulated as follows:

$$\begin{aligned}&\underset{\mathbf{Z},\mathbf{Q}}{\text{minimize}}\ \|\mathbf{X} - \mathbf{XZ}\|_{\mathrm{F}}^2 + \|\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\|_1 + \|\mathbf{ZG}\|_{2,1} + \|\mathbf{Z}\|_{\mathbf{Q}} \\ &\text{subject to } \operatorname{diag}(\mathbf{Z}) = 0, \mathbf{Z} \geq 0, \mathbf{Q} \in \mathcal{Q}. \end{aligned} \tag{1}$$

Proposition 1: Let $\mathbf{X} \in \mathbb{R}^{D\times N}$ be a matrix whose columns are drawn from a union of K distinct subspaces, with the subspace assignment indicated by $\mathbf{Q}^*$. The optimal solution to the problem in (1) is given by $\mathbf{Q}^*$ and $\mathbf{Z}^*$, where $\mathbf{Z}^*$ is block-diagonal after permuted according to $\mathbf{Q}^*$. (Proof See Appendix C.) □

Proposition 1 demonstrates that, under the assumption that frames are distributed across distinct subspaces, the optimal solution to problem (1) will align with the true segmentation. However, due to the influence of noise, the human motion data may not lie perfectly within the subspaces. Thus, there are inherent trade-offs in (1) due to practical dataset challenges such as image noise and subtle motions, which may cause frames not to align precisely with K subspaces. Specifically, the term $\|\mathbf{X} - \mathbf{XZ}\|_{\mathrm{F}}^2 + \|\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\|_1$ ensures sparsity and accurate data subspace embedding in the outputs $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$, promoting distinctiveness among them. In contrast, the term $\|\mathbf{ZG}\|_{2,1}$ requires these outputs to align with their neighbors' coefficients $\{\mathbf{z}_l\}_{l\in\mathcal{N}_i}$. This necessitates a balance between representing data across K clusters and maintaining temporal,

aiming to segment the data sequence into smaller segments where, for instance, in a segment $[\mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_j]$ with $j > i$, the subspace embedding are identical, i.e., $\mathbf{z}_i = \mathbf{z}_{i+1} = \ldots = \mathbf{z}_j$. Additionally, the term $\|\mathbf{Z}\|_{\mathbf{Q}}$ promotes effective grouping based on $\mathbf{Z}$ while adhering to the constraint $\mathbf{Q} \in \mathcal{Q}$, which stipulates that the cluster assignment of the ith frame must match that of its neighbors, introducing a further trade-off between dependent clustering and TVS considerations.

Theorem 3: [Impact of TVS on Segmentation]

Assume each motion segment generates data lying in one of K linear subspaces with within-segment variance σ^2 and between-subspace separation $\Delta^2_{\text{sub}} > 0$. With independent LLM adjacency errors of rate $p < \frac{1}{2}$, the expected segmentation error satisfies $\mathbb{E}[\text{Err}_{\text{HMS}}] \leq C_1 \frac{p}{L_{\min}} + C_2 \frac{\sigma^2}{\Delta^2_{\text{sub}}}$. When TVS boundaries align with true actions, the optimal solution $\mathbf{Z}^*$ becomes block-diagonal, achieving exact segmentation. (Proof See Appendix D.) □

Theorem 3 establishes that TVS improves segmentation robustness by suppressing random adjacency errors and stabilizing intra-segment embeddings. The first term $C_1 \frac{p}{L_{\min}}$ reflects the resilience to LLM-induced boundary noise, which diminishes as segment length increases, while the second term $C_2 \frac{\sigma^2}{\Delta^2_{\text{sub}}}$ captures the dependence on subspace separability. Perfectly aligned TVS boundaries yield theoretically exact segmentation, confirming the effectiveness of LLM-guided temporal reasoning in enhancing motion boundary localization.

C. A Feedback-Enabled Optimization Algorithm

We employ the ADMM method [70] to solve the optimization problem formulated in (1). In order to separate the third term in (1) from the other three terms, we introduce an additional variable $\mathbf{H} = \mathbf{ZG}$. By incorporating an augmented Lagrangian multiplier to handle the introduced linear constraint, we can reformulate (1) as the following problem:

$$\begin{aligned} \underset{\mathbf{Z},\mathbf{H},\mathbf{Q}}{\text{minimize}} \quad & \|\mathbf{X} - \mathbf{XZ}\|_{\text{F}}^2 + \|\mathbf{Z}^{\text{T}}\mathbf{Z}\|_1 + \|\mathbf{H}\|_{2,1} \\ & + \langle \mathbf{F}, \mathbf{H} - \mathbf{ZG} \rangle + \frac{\gamma}{2}\|\mathbf{H} - \mathbf{ZG}\|_{\text{F}}^2 + \|\mathbf{Z}\|_{\mathbf{Q}} \\ \text{subject to} \quad & \text{diag}(\mathbf{Z}) = 0, \mathbf{Z} \geq 0, \mathbf{Q} \in \mathcal{Q} \end{aligned} \tag{2}$$

where $\mathbf{F} \in \mathbb{R}^{N \times N}$ is the Lagrangian multiplier and γ is an adaptive weight parameter for enforcing the condition $\mathbf{H} = \mathbf{ZG}$. To solve (2), we adopt a feedback-enabled optimization strategy, where we iteratively solve three sub-problems for $\mathbf{Z}$, $\mathbf{H}$, and $\mathbf{Q}$ while keeping the other fixed, respectively.

1) Z-Solution: Fixing $\mathbf{H}$ and $\mathbf{Q}$, solve for $\mathbf{Z}$ by

$$\begin{aligned} \underset{\mathbf{Z}}{\text{minimize}} \quad & \|\mathbf{X} - \mathbf{XZ}\|_{\text{F}}^2 + \|\mathbf{Z}^{\text{T}}\mathbf{Z}\|_1 + \langle \mathbf{F}, \mathbf{H} - \mathbf{ZG} \rangle \\ & + \frac{\gamma}{2}\|\mathbf{H} - \mathbf{ZG}\|_{\text{F}}^2 + \|\mathbf{Z}\|_{\mathbf{Q}} \\ \text{subject to} \quad & \text{diag}(\mathbf{Z}) = 0, \mathbf{Z} \geq 0 \end{aligned} \tag{3}$$

Since $\mathbf{Z}$ consists of non-negative elements, we can rewrite the objective function in (3) as a function:

$$\begin{aligned} \mathcal{J}(\mathbf{Z}) = & \|\mathbf{X} - \mathbf{XZ}\|_{\text{F}}^2 + \mathbf{e}^{\text{T}}\mathbf{Z}^{\text{T}}\mathbf{Z}\mathbf{e} + \langle \mathbf{F}, \mathbf{H} - \mathbf{ZG} \rangle \\ & + \frac{\gamma}{2}\|\mathbf{H} - \mathbf{ZG}\|_{\text{F}}^2 + \|\mathbf{\Theta} \odot \mathbf{Z}\|_1 \end{aligned} \tag{4}$$

The sub-problem defined in (3) can be formulated as a convex quadratic programming problem with specific constraints for the variable $\mathbf{Z}$, involving the function $\mathcal{J}(\mathbf{Z})$ from (4). In this problem, we aim to minimize $\mathcal{J}(\mathbf{Z})$ while satisfying the given constraints. To tackle this, we employ the projected gradient method, which is a well-established approach known for its simplicity and effectiveness in solving such problems. This method is chosen as our preferred solution due to its suitability for our problem's requirements.

Consider the partial derivative of $\|\mathbf{\Theta} \odot \mathbf{Z}\|_1$ with respect to each element Z_{ij}:

$$\frac{\partial}{\partial Z_{ij}} \left(\sum_{k,l} |\Theta_{kl} Z_{kl}| \right) = \frac{\partial}{\partial Z_{ij}} |\Theta_{ij} Z_{ij}|$$

Using the properties of the absolute value function, we get:

$$\frac{\partial}{\partial Z_{ij}} |\Theta_{ij} Z_{ij}| = \Theta_{ij} \cdot \text{sign}(\Theta_{ij} Z_{ij})$$

where $\text{sign}(x)$ is the sign function, defined as:

$$\text{sign}(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < 0 \end{cases}$$

Therefore, the partial derivative for each element Z_{ij} is:

$$\frac{\partial \|\mathbf{\Theta} \odot \mathbf{Z}\|_1}{\partial Z_{ij}} = \Theta_{ij} \cdot \text{sign}(\Theta_{ij} Z_{ij})$$

Combining all the partial derivatives into matrix form, the gradient of $\|\mathbf{\Theta} \odot \mathbf{Z}\|_1$ with respect to $\mathbf{Z}$ is:

$$\frac{\partial \|\mathbf{\Theta} \odot \mathbf{Z}\|_1}{\partial \mathbf{Z}} = \mathbf{\Theta} \odot \text{sign}(\mathbf{\Theta} \odot \mathbf{Z})$$

where $\text{sign}(\mathbf{\Theta} \odot \mathbf{Z})$ is the matrix obtained by applying the sign function element-wise to $\mathbf{\Theta} \odot \mathbf{Z}$.

The derivative of $\mathcal{J}(\mathbf{Z})$ with respect to $\mathbf{Z}$ can be expressed as: $\partial \mathcal{J}(\mathbf{Z}) = -2\mathbf{X}^{\text{T}}(\mathbf{X} - \mathbf{XZ}) + 2\mathbf{ZE} - \mathbf{FG}^{\text{T}} - \gamma(\mathbf{H} - \mathbf{ZG})\mathbf{G}^{\text{T}} + \mathbf{\Theta} \odot \text{sign}(\mathbf{\Theta} \odot \mathbf{Z})$ where $\mathbf{E} \in \mathbb{R}^{N \times N}$ is an all-one matrix.

Setting the derivative to zero gives

$$\begin{aligned} & 2\mathbf{X}^{\text{T}}\mathbf{XZ} + \mathbf{Z}(2\mathbf{E} + \gamma\mathbf{GG}^{\text{T}}) \\ & = \mathbf{FG}^{\text{T}} + \gamma\mathbf{HG}^{\text{T}} + 2\mathbf{X}^{T}\mathbf{X} - \mathbf{\Theta} \odot \text{sign}(\mathbf{\Theta} \odot \mathbf{Z}). \end{aligned} \tag{5}$$

The equation presented is a well-known Sylvester equation in the form $\mathbf{AZ} + \mathbf{ZB} = \mathbf{C}$, where $\mathbf{A} = 2\mathbf{X}^{\text{T}}\mathbf{X}$, $\mathbf{B} = 2\mathbf{E} + \gamma\mathbf{GG}^{\text{T}}$, and $\mathbf{C} = \mathbf{FG}^{\text{T}} + \gamma\mathbf{HG}^{\text{T}} + 2\mathbf{X}^{T}\mathbf{X} - \mathbf{\Theta} \odot \text{sign}(\mathbf{\Theta} \odot \mathbf{Z})$.

We adopt Bartels-Stewart algorithm [71] to solve $\mathbf{Z}$. Specifically, we first perform Schur decomposition on $\mathbf{A}$ and $\mathbf{B}$. The Schur decomposition of $\mathbf{A}$ and $\mathbf{B}$ is given by $\mathbf{A} = \mathbf{Q}_A\mathbf{T}_A\mathbf{Q}_A^H$ and $\mathbf{B} = \mathbf{Q}_B\mathbf{T}_B\mathbf{Q}_B^H$, where $\mathbf{Q}_A$ and $\mathbf{Q}_B$ are unitary matrices and $\mathbf{T}_A$ and $\mathbf{T}_B$ are upper triangular matrices. Then, we use the unitary matrices from the Schur decomposition to transform $\mathbf{C}$ to $\mathbf{C}' = \mathbf{Q}_A^H\mathbf{CQ}_B$. Next, we solve the simplified equation $\mathbf{T}_A\mathbf{Z}' + \mathbf{Z}'\mathbf{T}_B = \mathbf{C}'$. This can be done using a back-substitution method since $\mathbf{T}_A$ and $\mathbf{T}_B$ are upper triangular matrices. Finally, we transform $\mathbf{Z}'$ back to $\mathbf{Z}$ by $\mathbf{Z} = \mathbf{Q}_A\mathbf{Z}'\mathbf{Q}_B^H$.

However, it's worth noting that the critical frame $\mathbf{Z}$ of the objective function may not necessarily lie within the feasible set defined in (3). To address this, we can employ a projection

operator to find a feasible frame starting from the critical frame $\mathbf{Z}$.

$$\prod\nolimits_{\mathcal{Z}}(\mathbf{Z}) = \arg\min_{\tilde{\mathbf{Z}}\in\mathcal{Z}} \|\tilde{\mathbf{Z}} - \mathbf{Z}\|_{\mathrm{F}}^2 \tag{6}$$

where $\mathcal{Z} = \{\mathbf{Z}|\mathrm{diag}(\mathbf{Z}) = 0, \mathbf{Z} \geq 0\}$. For a simple and quick solution to (6), we implement the projection operator $\prod_{\mathcal{Z}}(\mathbf{Z})$ as follows:

$$z_{ij}^* = \begin{cases} z_{ij} & \text{if } z_{ij} \geq 0 \text{ and } i \neq j \\ 0 & \text{if } z_{ij} < 0 \text{ or } i = j \end{cases}$$

where z_{ij} and z_{ij}^* are the elements of $\mathbf{Z}$ and its projection $\prod_{\mathcal{Z}}(Z)$, respectively.

2) H-Solution: Fixing $\mathbf{Z}$ and $\mathbf{Q}$, solve for $\mathbf{H}$ by

$$\begin{aligned}\underset{\mathbf{H}}{\text{minimize}} \quad & \|\mathbf{H}\|_{2,1} + \langle \mathbf{F}, \mathbf{H} - \mathbf{Z}\mathbf{G}\rangle \\ & + \frac{\mathbf{H} - \mathbf{Z}\mathbf{G}}{2}\|\mathbf{H} - \mathbf{Z}\mathbf{G}\|_{\mathrm{F}}^2 \end{aligned} \tag{7}$$

which is equivalent to minimizing $\|\mathbf{H}\|_{2,1} + \frac{\gamma}{2}\|\mathbf{H} - (\mathbf{Z}\mathbf{G} - (1/\gamma)\mathbf{F})\|_{\mathrm{F}}^2$ with respect to $\mathbf{H}$. Denote $\mathbf{P} = \mathbf{Z}\mathbf{G} - (1/\gamma)\mathbf{F}$. Then the closed-form solution to (7) will be given as follows [72]:

$$\mathbf{h}_i = \begin{cases} \dfrac{\|\mathbf{p}_i\| - (/\gamma)}{\|\mathbf{p}_i\|}\mathbf{p}_i & \text{if } \|\mathbf{p}_i\| > /\gamma \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{h}_i$, $\mathbf{p}_i$ are the ith column of $\mathbf{H}$, $\mathbf{P}$, respectively.

Proposition 2 (Optimality): For fixed $(\mathbf{Z}, \mathbf{Q}, \mathbf{F}, \gamma)$ the sub-problem (7) is the proximal operator of the $\ell_{2,1}$ norm and admits the closed-form group-shrinkage solution. Hence the $\mathbf{H}$-update attains the global minimizer of (7) at every iteration. (Proof See Appendix E.) □

3) Q-Solution: Fixing $\mathbf{Z}$ and $\mathbf{H}$, solve for $\mathbf{Q}$ by

$$\min_{\mathbf{Q}} \|\mathbf{Z}\|_{\mathbf{Q}}, \quad \text{subject to} \quad \mathbf{Q} \in \mathcal{Q} \tag{8}$$

Proposition 3: We have the following equivalent problem

$$\min_{\mathbf{Q}} \|\mathbf{Z}\|_{\mathbf{Q}} \iff \min_{\mathbf{Q}} \mathrm{Trace}(\mathbf{Q}^\top(\mathbf{D} - (|\mathbf{Z}| + |\mathbf{Z}^{\mathrm{T}}|)/2)\mathbf{Q})$$

where the matrix $\mathbf{D}$ is known as the degree matrix. The degree matrix $\mathbf{D}$ is defined as: $D_{ii} = \sum_{j=1}^{N}[(|\mathbf{Z}| + |\mathbf{Z}^{\mathrm{T}}|)/2]_{ij}$. For all off-diagonal elements $i \neq j$, $D_{ij} = 0$. (Proof See Appendix G.) □

The objective function in (8) is the traditional normalized cut clustering problem [73] with a TVS constraint. The Laplacian matrix $\mathbf{L}$ can be computed using the formula $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}[(|\mathbf{Z}| + |\mathbf{Z}^{\mathrm{T}}|)/2]\mathbf{D}^{-1/2}$, where $\mathbf{I}$ is an identity matrix. Consequently, the eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K$ corresponding to the first K smallest eigenvalues $\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_K$ of $\mathbf{L}$ are computed, satisfying $\mathbf{L}\mathbf{v}_k = \tilde{\lambda}_k\mathbf{v}_k$. These eigenvectors are arranged as columns in a matrix $\mathbf{V} \in \mathbb{R}^{N\times K}$.

Let $\mathbf{u}_i \in \mathbb{R}^K$ represent the vector of the ith row of $\mathbf{V}$, where $i = 1, \ldots, N$. The problem (8) can be relaxed to the following form:

$$\underset{\{\mathcal{C}_k, \boldsymbol{\mu}_k\}_{k=1}^K}{\text{minimize}} \sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k} \|\mathbf{u}_i - \boldsymbol{\mu}_k\|_2^2, \quad \text{subject to} \quad \mathbf{Q} \in \mathcal{Q} \tag{9}$$

where $i \in \mathcal{C}_k$ if $q_{i,k} = 1$. However, the requirement $\mathbf{q}_i = \mathbf{q}_j \, \forall j \in \mathcal{N}_i$ in the constraint $\mathbf{Q} \in \mathcal{Q}$ makes solving problem (9) highly challenging. Since the constraint mandates that the clustering assignments of the ith frame and its neighbors remain consistent, we relax the constraint to that the clustering center corresponding to the ith frame should coincide with the center of its neighbors. This leads us to the formulation of the following problem:

$$\underset{\{\mathcal{C}_k, \boldsymbol{\mu}_k\}_{k=1}^K}{\text{minimize}} \sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k} \left(\|\mathbf{u}_i - \boldsymbol{\mu}_k\|_2^2 + \eta \sum_{j\in\mathcal{N}_i} \|\mathbf{u}_j - \boldsymbol{\mu}_k\|_2^2 \right) \tag{10}$$

where the penalty coefficient η is set to $1/\mathcal{N}_i$ for weight balance.

The term $\sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k} \|\mathbf{u}_i - \boldsymbol{\mu}_k\|_2^2$ aims to independently fit all the data with the center $\{\boldsymbol{\mu}_k\}$. However, the term $\eta\sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k}\sum_{j\in\mathcal{N}_i} \|\mathbf{u}_j - \boldsymbol{\mu}_k\|_2^2$ desires the $\boldsymbol{\mu}_k$ to be identical temporally, i.e., the center of $\mathbf{u}_i$ is the same as the center of $\mathbf{u}_j$ for any $j \in \mathcal{N}_i$. Consequently, minimizing these two terms simultaneously leads to a trade-off between fitting data to K centers and maintaining temporal of the center assignment, where the desired outcome is to divide the data sequence into multiple small segments.

It is still challenging to solve problem (10) directly due to its NP-hard nature. We first propose the following proposition, which will be utilized to adapt problem (10) into a new form.

Proposition 4: The term $\sum_{i\in\mathcal{C}_k}\left(\|\mathbf{u}_i - \boldsymbol{\mu}_k\|_2^2 + \eta\sum_{j\in\mathcal{N}_i}\|\mathbf{u}_j - \boldsymbol{\mu}_k\|_2^2\right)$ in (10) is equivalent to $\sum_{i=1}^{N}\|\mathbf{u}_i - \boldsymbol{\mu}_k\|_2^2(\mathbb{1}(i \in \mathcal{C}_k) + \eta n_k(i))$ where $n_k(i)$ is the number of times the frame $\mathbf{u}_i$ appears as a sequential neighbor of a frame in the k-th cluster, i.e., $n_k(i) = \sum_{j\in\mathcal{C}_k}\mathbb{1}(i \in \mathcal{N}_j)$ and the indicator function $\mathbb{1}(s) = 1$ if s is true and zero otherwise. (Proof See Appendix F.) □

According to proposition 4, problem (10) can be rewritten as the following new weighted problem: $\underset{\{\mathcal{C}_k, \boldsymbol{\mu}_k\}_{k=1}^K}{\text{minimize}} \sum_{k=1}^{K}\sum_{i=1}^{N}\|\mathbf{u}_i - \boldsymbol{\mu}_k\|_2^2 w_{k,i}$, where $w_{k,i} = \mathbb{1}(i \in \mathcal{C}_k) + \eta n_k(i)$. Observing that the new weighted problem can be solved by addressing two sub-problems for $\mathcal{C}_k$ and $\boldsymbol{\mu}_k$ in an alternating manner when one is fixed, respectively, we first focus on solving the new problem with the given cluster assignment $\{\mathcal{C}_k\}_{k=1}^K$. Denote the objective function of the new weighted problem as $\mathcal{J}_1(\{\boldsymbol{\mu}_k\}_{k=1}^K)$. If we take the derivative of $\mathcal{J}_1(\{\boldsymbol{\mu}_k\}_{k=1}^K)$ with respect to $\boldsymbol{\mu}_k$ and set it to zero, i.e., $\frac{\partial\mathcal{J}_1(\{\boldsymbol{\mu}_k\}_{k=1}^K)}{\partial\boldsymbol{\mu}_k} = 0$, we obtain $\boldsymbol{\mu}_k = \frac{1}{\sum_{i=1}^{N} w_{k,i}}\sum_{i=1}^{N} w_{k,i}\mathbf{u}_i$. We then solve the cluster assignment with the given cluster center. This is done by evaluating the weighted combination of the residual from the frame to a given center, as well as the residuals of its sequential neighbors, so that the estimated cluster label for the frame $\mathbf{u}_i$ is assigned to the l-th cluster, where $l = \underset{k\in\{1,2,\ldots,K\}}{\arg\min} \|\mathbf{u}_i - \boldsymbol{\mu}_k\|_2^2 w_{k,i}$.

The algorithm alternates between center update and cluster assignment steps until convergence. In the center update step, the resulting center represents the global optimum given a cluster assignment. This step learns the center that minimizes the distance to all frames in the cluster, including their sequential neighbors. Therefore, the center update step cannot increase the overall objective function. Similarly, in the cluster assignment step, each frame is assigned to the cluster that minimizes the distance to itself and its sequential neighbors, which also cannot increase the overall objective function. Since there is a finite number of ways the frames can be assigned, and the objective function in the new weighted

problem is bounded below by zero, the proposed alternating algorithm must terminate at a locally optimal clustering result. To determine the number of clusters K, we use the silhouette score, which measures the similarity of a sample point to its own cluster in comparison to the nearest cluster. By calculating the silhouette score for different values of K, the optimal number of clusters is chosen as the value of K that maximizes the silhouette score.

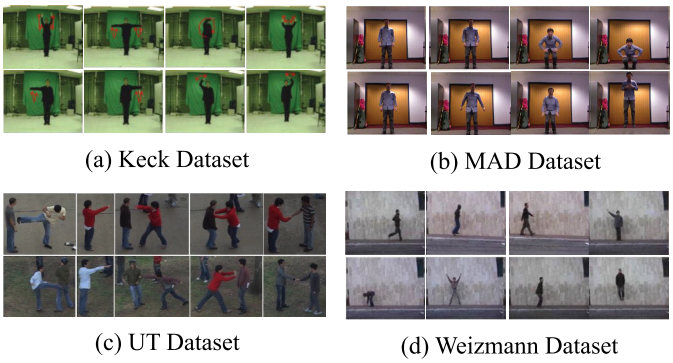
Algorithm 2 TVSH Method

Input: $\mathbf{X}$.
Output: $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K\}$
Initialize $\mathbf{G}$, $\mathbf{F} = \mathbf{1}$, $\rho = 1.1, \gamma = 0.1$. $\mathbf{H} = \mathbf{ZG}$ where $\mathbf{Z}$ is the similarity matrix given by cosine measurement. $\mathbf{Q}$ is initialized by K-means [13].
repeat
  Find $\mathbf{Z}$ by solving (5).
  Calculate the projection $\mathbf{Z} \leftarrow \prod_{\mathcal{Z}}(\mathbf{Z})$ by solving (6).
  Find $\mathbf{H}$ by solving (7);
  Update $\mathbf{F} \leftarrow \mathbf{F} + \gamma(\mathbf{H} - \mathbf{ZG})$, $\gamma \leftarrow \rho\gamma$.
  Update $\mathbf{Q}$ by the following steps:
  Calculate $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}[(|\mathbf{Z}| + |\mathbf{Z}^{\mathrm{T}}|)/2]\mathbf{D}^{-1/2}$.
  Compute the smallest K eigenvectors of $\mathbf{L}$ denoted by $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K]$. Denote the row of $\mathbf{V}$ as $\{\mathbf{u}_i\}_{i=1}^{N}$.
  repeat
    for $k = 1$ **to** K **do**
      Form weight matrix $\mathbf{W}_k$ and calculate $\boldsymbol{\mu}_k$.
    end for
    for $k = 1$ **to** K **do**
      $\mathcal{C}_k \leftarrow \{i \in \{1, 2, ..., N\} : k = \arg\min_{k \in \{1,2,...,K\}} ||\mathbf{u}_i - \boldsymbol{\mu}_k||_2^2 w_{k,i}\}$
    end for
  until $\mathbf{Q}$ can not be changed.
until The objective function value of (2) can not be decreased.

By iteratively solving (3), (7), and (8), we can obtain a solution to (2). During this process, we group the subspace embeddings by solving (8) and update the embeddings based on feedback from the HMS solution of (8). The convergence of sub-problem (3), the closed-form solution of sub-problem (7), and the convergence of solving (8) ensure the overall convergence of the algorithm for (2). Algorithm 2 presents the pseudocode for our clustering method.

Theorem 4 (Convergence): Under bounded and lower-semicontinuous augmented Lagrangian, nondecreasing $\gamma_t \to \gamma_\infty \in (0, \infty)$, and bounded $\rho > 1$, the proposed ADMM-based alternating scheme ensures monotonic decrease of the objective and convergence of $(\mathbf{Z}^{(t)}, \mathbf{H}^{(t)}, \mathbf{Q}^{(t)})$ to a first-order stationary point. If each $\mathbf{Q}$-update reaches its relaxed global optimum, every accumulation point satisfies the KKT conditions. (Proof See Appendix H) .□

This theorem confirms that the alternating optimization is theoretically stable and convergent: the objective value decreases monotonically, the iterates approach a stationary solution, and, with exact subproblem updates, the algorithm attains KKT-level optimality, guaranteeing reliable convergence behaviour in practice.



(a) Keck Dataset (b) MAD Dataset (c) UT Dataset (d) Weizmann Dataset

Fig. 2. Sampling frames from four human motion benchmark datasets, *i.e.*, (a) Keck [75], (b) MAD [76], (c) UT [77], and (d) Weiz [78].

IV. Experimental Results and Analysis

In this section, we first introduce the human motion datasets used in our experiments (Section IV-A). We then present a comparison of our method with state-of-the-art techniques (Section IV-B). Finally, we show the effective analysis of the LLM-based TVS (Section IV-C).

A. Human Motion Datasets and Experimental Setup

To provide a comprehensive evaluation of the proposed model, we perform experiments on four well-established benchmark human motion datasets. Some example frames from these datasets can be seen in Figure 2.

- *Keck Gesture Dataset (Keck)* [75] consists of 14 different motions from military signals, in which each subject is carried out 14 motions and gestures. Besides, the videos in this dataset were obtained by a fixed camera when these subjects stand out in a static background.
- *Multi-Modal Action Detection Dataset (MAD)* [76] consists of motions captured from various modalities using a Microsoft Kinect V2 system, which includes RGB images, depth cues, and skeleton formats. Specifically, the RGB images and 3D depth cues are of a size of 240 × 320. Moreover, each subject performs 35 different motions within two indoor scenes.
- *UT-Intermotion Dataset (UT)* [77] is composed of 20 videos, each of which includes six different motion types of human-human intermotions (such as punching, pushing, pointing, hugging, kicking, and handshaking).
- *Weizmann Dataset (Weiz)* [78] is composed of 90 video sequences with 10 motions (running, walking, skipping, bending, etc.) captured by nine subjects in an outdoor environment. All videos have a size of 180 × 144 with 50 fps.

We evaluate clustering performance using four metrics: accuracy (Acc), normalized mutual information (NMI), precision (Pr), and adjusted rand index (ARI). These metrics assess the consistency between learned and true labels, with higher values indicating better performance. Let $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$ and $\hat{\mathcal{L}} = \{\hat{l}_1, \hat{l}_2, \ldots, \hat{l}_N\}$ represent the ground-truth and predicted labels, respectively, where l_i and $\hat{l}_i$ denote the true and predicted labels for the ith sample. Acc is defined as the proportion of correctly clustered samples:

TABLE I

CLUSTERING PERFORMANCE OF COMPARED METHODS IN TERMS OF ACC AND NMI ON FOUR HUMAN MOTION VIDEOSETS. THE BEST RESULT IS HIGHLIGHTED IN BOLD. THE IMPROVEMENT RELATIVE TO THE SECOND-BEST METHOD IS DEPICTED BY ↑. (M) DENOTES THE NEED FOR LABELED MAD DATASET ASSISTANCE, AND (K) INDICATES THE REQUIREMENT FOR LABELED KECK DATASET ASSISTANCE

(a) Results on Keck dataset

Method	Acc ↑	NMI ↑
SSC [74]	0.3137	0.3858
OSC [6]	0.4393	0.5931
TSC(M) [7]	0.4653	0.6935
LTS [8]	0.4924	0.6213
DSAE [34]	0.5136	0.5100
VSDA [10]	0.5804	0.7397
CDMS(M) [9]	0.6044	0.7891
ARLRR [19]	0.5010	0.5270
TVSH	**0.8048**	**0.8090**

(b) Results on MAD dataset

Method	Acc ↑	NMI ↑
SSC [74]	0.3817	0.4758
OSC [6]	0.4327	0.5589
TSC(K) [7]	0.5473	0.7691
LTS [8]	0.5466	0.6547
DSAE [34]	0.5898	0.6309
VSDA [10]	0.5606	0.7770
CDMS(K) [9]	0.6536	0.8251
ARLRR [19]	0.5125	0.5099
TVSH	**0.7829**	**0.8438**

(c) Results on UT dataset

Method	Acc ↑	NMI ↑
SSC [74]	0.4389	0.4998
OSC [6]	0.5846	0.6877
TSC(K) [7]	0.5213	0.7216
LTS [8]	0.6724	0.7435
DSAE [34]	0.7323	0.6717
VSDA [10]	0.6203	0.8226
CDMS(K) [9]	0.6547	0.8267
ARLRR [19]	0.5148	0.5121
TVSH	**0.8123**	**0.8488**

(d) Results on Weiz dataset

Method	Acc ↑	NMI ↑
SSC [74]	0.4576	0.6009
OSC [6]	0.5216	0.7047
TSC(K) [7]	0.5931	0.7971
LTS [8]	0.5674	0.6959
DSAE [34]	0.6120	0.6627
VSDA [10]	0.6287	0.7992
CDMS(K) [9]	0.6465	0.8601
ARLRR [19]	0.5436	0.5371
TVSH	**0.8045**	**0.8716**

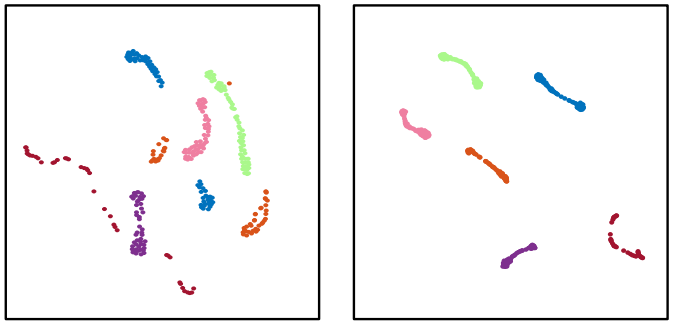
Acc $= \frac{1}{N}\sum_{i=1}^{N}\delta(l_i, \text{map}(\hat{l}_i))$, where $\delta(a,b)$ is the indicator function ($\delta(a,b) = 1$ if $a = b$, and 0 otherwise), and map$(\cdot)$ maps predicted labels to the best matching true labels using the Hungarian algorithm [79]. NMI quantifies the coherence between two sets. Let $H(\mathcal{L})$ and $H(\hat{\mathcal{L}})$ represent the entropies of the sets $\mathcal{L}$ and $\hat{\mathcal{L}}$, respectively. NMI is defined as: $\text{NMI}(\mathcal{L},\hat{\mathcal{L}}) = \text{MI}(\mathcal{L},\hat{\mathcal{L}})/\sqrt{H(\mathcal{L})H(\hat{\mathcal{L}})}$, where $\text{MI}(\mathcal{L},\hat{\mathcal{L}})$ measures the mutual information between the sets. Higher mutual information and lower uncertainty result in a higher NMI. If the sets are randomly distributed, NMI equals 0. Pr calculates the percentage of correctly clustered pairs among all pairs with the same clustering label. True positive (TP), false positive (FP), and false negative (FN) represent the numbers of correctly labeled samples in the positive class, misclassified samples in the positive cluster, and misclassified samples in the negative cluster, respectively. Precision is defined as: $\text{Pr} = \frac{\text{TP}}{\text{TP+FP}}$. ARI [80] quantifies the similarity between two sets: $\text{ARI} = \left(\sum_{i,j=1}^{K} C_{n_{ij}}^2 - \mathbb{E}[\text{RI}]\right)/(C_0 - \mathbb{E}[\text{RI}])$, where $C_0 = \frac{1}{2}\left(\sum_{i=1}^{K} C_{|\hat{\mathcal{L}}^{(i)}|}^2 + \sum_{i=1}^{K} C_{|\mathcal{L}^{(i)}|}^2\right)$ and $\mathbb{E}[\text{RI}] = \sum_{i=1}^{K} C_{|\mathcal{L}^{(i)}|}^2 \sum_{i=1}^{K} C_{|\hat{\mathcal{L}}^{(i)}|}^2 / C_N^2$. Here, $|\mathcal{L}^{(i)}|$ and $|\hat{\mathcal{L}}^{(i)}|$ represent the number of samples in the ith cluster of the ground-truth and predicted labels, respectively. The value n_{ij} denotes the number of samples in the ith true cluster grouped into the jth predicted cluster. The notation C_n^m represents the number of ways to choose m items from n.

We evaluate the performance of our method through a comparative analysis with thirteen approaches, as outlined in Section II. Each method was independently tested ten times, and the average results were reported. For the proposed scheme, the TVS learning was performed using the following LLMs: GPT-o1, DeepSeek-v3-2-exp, Claude-Sonnet-4-5-20250929, Gemini-2.0-Flash-exp, Grok-4, and Qwen3-235B-a22b.

B. HMS Performance Comparison

Tables I–II summarize results on four benchmarks (Keck, MAD, UT, Weiz) using Acc, NMI, Pr, and **ARI**. TVSH attains the best performance across all datasets and metrics.

On the Keck dataset, TVSH improves accuracy from 0.6044 (CDMS) to 0.8048, representing a significant improvement over the best baseline. On MAD, the accuracy increases from 0.6536 to 0.7829, while NMI rises from 0.8286 to 0.8438. On UT, TVSH achieves 0.8123 accuracy and 0.8488 NMI, both higher than those of existing methods. On the more diverse Weiz dataset, TVSH reaches 0.8045 accuracy and 0.8716 NMI. Precision and ARI exhibit consistent improvement trends, confirming the robustness of TVSH in maintaining temporal coherence and enhancing motion discriminability.



(a) CDMS (b) TVSH

Fig. 3. Visualization of the two-dimensional t-SNE of the extracted features from six motions in the Weiz dataset. Points in different colors depict frames of different motions.

1) Superiority of the Proposed TVSH Method:

a) Superiority 1: temporal modeling: Conventional methods often fail to explicitly capture temporal dependencies in human motion sequences, making it difficult to identify gradually transitioning motions. Specifically, when a single motion contains multiple stages with significant amplitude variations, it is often misinterpreted as multiple separate motions. Exploring temporal modeling offers a solution to this challenge. The proposed method first utilizes an LLM to obtain a TVS matrix, which explicitly encodes temporal relationships between frames. Then, the TVS matrix-based regularization is introduced to enforce temporal continuity in both the embedding space and the segmentation result, thereby reducing ambiguity in temporal motion transitions.

Figure 3 further visualizes the two-dimensional t-SNE embeddings of the extracted features u_i from six motions in the Weiz dataset. The proposed method produces more compact clusters than the strong baseline CDMS, highlighting

TABLE II

CLUSTERING PERFORMANCE OF COMPARED METHODS IN TERMS OF PR AND ARI ON FOUR HUMAN MOTION VIDEOSETS

(a) Results on Keck dataset

Method	Pr ↑	ARI ↑
SSC [74]	0.3511	0.2446
OSC [6]	0.3767	0.2743
TSC(M) [7]	0.4214	0.3457
LTS [8]	0.4457	0.3052
DSAE [34]	0.4195	0.3418
VSDA [10]	0.4311	0.3529
CDMS(M) [9]	0.5828	0.5174
ARLRR [19]	0.5772	0.5046
TVSH	**0.7559**	**0.7214**

(b) Results on MAD dataset

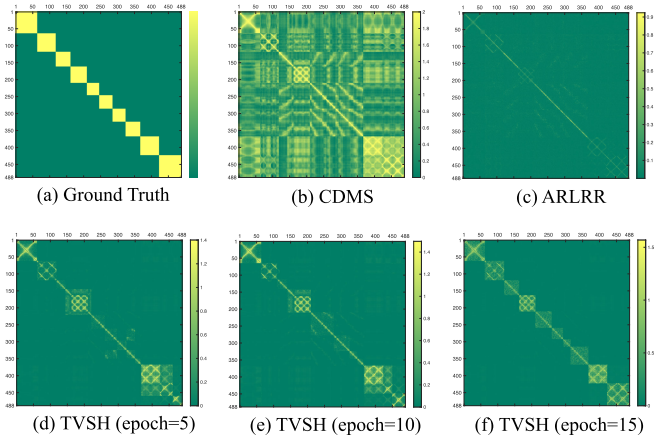
Method	Pr ↑	ARI ↑
SSC [74]	0.3151	0.1994
OSC [6]	0.4024	0.2403
TSC(K) [7]	0.5116	0.3724
LTS [8]	0.4673	0.3426
DSAE [34]	0.5492	0.3891
VSDA [10]	0.5667	0.3780
CDMS(K) [9]	0.5761	0.4128
ARLRR [19]	0.5426	0.4072
TVSH	**0.7043**	**0.6973**

(c) Results on UT dataset

Method	Pr ↑	ARI ↑
SSC [74]	0.5426	0.3772
OSC [6]	0.5426	0.3966
TSC(K) [7]	0.5864	0.4217
LTS [8]	0.5774	0.4457
DSAE [34]	0.6189	0.4895
VSDA [10]	0.6334	0.5202
CDMS(K) [9]	0.6466	0.5539
ARLRR [19]	0.6054	0.5213
TVSH	**0.7477**	**0.7153**

(d) Results on Weiz dataset

Method	Pr ↑	ARI ↑
SSC [74]	0.4469	0.3620
OSC [6]	0.5126	0.4422
TSC(K) [7]	0.5667	0.5324
LTS [8]	0.5991	0.5724
DSAE [34]	0.6233	0.5406
VSDA [10]	0.6180	0.5378
CDMS(K) [9]	0.6316	0.5561
ARLRR [19]	0.6211	0.5146
TVSH	**0.9012**	**0.8867**



(a) Ground Truth (b) CDMS (c) ARLRR

(d) TVSH (epoch=5) (e) TVSH (epoch=10) (f) TVSH (epoch=15)

Fig. 4. Visualizations of the similarity matrix. (a) Ground-truth similarity matrix of the Ido human motion sequence in the Weiz dataset. Yellow regions indicate high similarity between frames of the same motion, while green regions denote zero similarity between frames of the same motion. (b)–(c) show the similarity matrices provided by the baselines CDMS and ARLRR. (d)–(f) show the similarity matrices generated by the proposed TVSH at different iterations.

the effectiveness of temporal regularization in the embedding component of the proposed TVSH. Based on the embedded features shown in Figure 4(b), achieving better motion segmentation performance becomes easier. This also explains why the proposed method can achieve more temporally semantically accurate motion segments.

b) Superiority 2: joint optimization: Traditional clustering-based human motion segmentation methods typically perform feature extraction and clustering of embedded features in separate stages, resulting in an embedding process that does not receive feedback from the clustering outcome. However, a large clustering loss indicates that the embedded features are difficult to cluster, and it is meaningful to adjust the embedding to generate new features that minimize the clustering loss as much as possible. Therefore, our method adopts a feedback-enabled joint optimization framework, where the segmentation results iteratively refine the learned embeddings to achieve the smallest possible clustering loss.

Figures 4(b-c) presents the similarity matrices for different methods, while Figures 4(d)–(f) show the similarity matrices across different iterations from the joint optimization of the proposed TVSH. The similarity matrix of the proposed TVSH framework progressively exhibits a clearer block-diagonal structure, demonstrating strong alignment with the ground truth shown in Figure 4(a). In contrast, the similarity matrices obtained by baseline methods, such as CDMS and ARLRR (Figure 4(b) and (c)), are less structured and more diffuse. Obviously, a similarity matrix that is more consistent with the ground truth in Figure 4(a) facilitates more accurate motion segmentation.

TABLE III

ABLATION STUDY OF THE EFFECTS OF TEMPORAL MODELING AND JOINT OPTIMIZATION

	Keck		MAD		UT	
	Acc	NMI	Acc	NMI	Acc	NMI
TVSH (w/o joint optimization)	0.7423	0.7829	0.7648	0.8322	0.7854	0.8371
TVSH (w/o temporal model)	0.7152	0.7428	0.7211	0.7546	0.6714	0.6211
TVSH	0.8048	0.8090	0.7829	0.8438	0.8123	0.8488

c) Ablation study: To assess the contribution of temporal modeling and joint optimization in TVSH, we conduct ablation experiments by selectively removing the temporal model and the joint optimization module. The variant TVSH (w/o temporal model) in Table III removes the temporal prior by setting $G = I$, thereby disabling temporal regularization.

When the LLM-guided TVS is incorporated as the temporal model, performance consistently improves across all datasets. For instance, accuracy increases from 0.7152 to 0.8048 on the Keck dataset, from 0.7211 to 0.7829 on MAD, and from 0.6714 to 0.8123 on UT. These gains confirm that the TVS-based temporal modeling enhances human motion segmentation performance.

The variant TVSH (w/o joint optimization) in Table III runs the proposed TVSH for only one iteration. This version produces weaker segmentation quality, as reflected by a 6–8% drop in accuracy across the datasets. In contrast, the full TVSH model benefits from iterative feedback between clustering and embedding, progressively refining segment boundaries and aligning the learned representation with semantic motion transitions. These results validate that joint optimization is essential for achieving temporally semantically consistent motion segmentation.

2) Performance Bottleneck: Although the proposed method achieves consistent improvements across all benchmarks, two

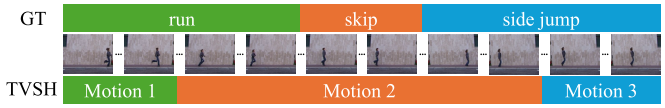


Fig. 5. Illustration of a failure case on the Weiz dataset (subject "ido") showing gradual transition boundaries between actions.

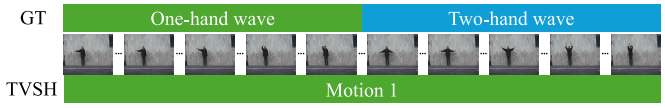
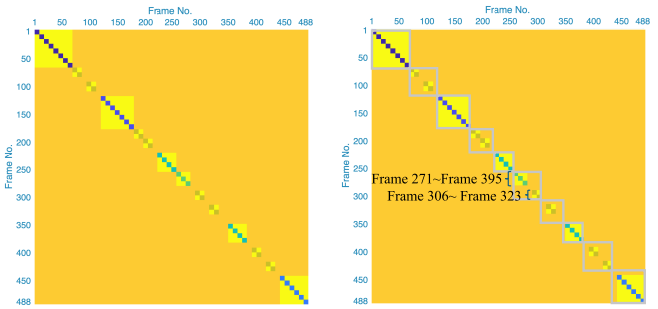


Fig. 6. Illustration of a failure case on the Weiz dataset (subject "ido") showing visually similar or look-alike actions.



(a) TVS matrix **G** (b) TVS matrix **G** with ground truth box

Fig. 7. (a) The TVS matrix **G** on Weiz dataset (person 'ido') generated by GPT-o1. (b) The TVS matrix **G** with ground truth label. The gray boxes indicate groups of frames that describe the same motion.

primary bottlenecks preventing TVSH from reaching perfect segmentation accuracy were identified and analyzed.

a) Bottleneck I: gradual transition boundaries: The first bottleneck arises in sequences where actions evolve smoothly without clear-cut temporal boundaries. As shown in Figure 5, when a motion transitions gradually from one motion to another (e.g., run → skip → side-jump), both visual and kinematic cues change continuously. These intermediate frames are semantically ambiguous, leading to minor drifts in segmentation boundaries or partial merging of adjacent segments.

b) Bottleneck Ii: look-alike motions: The second bottleneck arises when motions share highly similar morphological characteristics. As depicted in Figure 6, for visually related motions such as raising one hand and raising both hands, the motion segmentation algorithm may incorrectly classify these two motions as the same "raising hand" action. This occurs because the visual cues and kinematic features between these motions overlap significantly, making it difficult for the model to distinguish between them.

3) Limitations and Future Directions: Despite the effectiveness of the proposed method, several limitations remain. The first limitation arises in sequences where actions evolve smoothly without clear-cut temporal boundaries, causing minor drifts or partial merging of adjacent segments. The second limitation occurs when motions share highly similar morphological characteristics. This is due to the significant overlap in visual and kinematic features, making it difficult for the model to distinguish between them.

To address the limitations outlined above, future work will focus on enhancing the model's ability to handle gradual motion transitions and look-alike motions. For gradual transition boundaries, we plan to incorporate uncertainty-aware temporal modeling, which can adaptively capture smooth variations and probabilistic transition boundaries. This will help the model distinguish between genuine motion transitions and intra-action fluctuations. Additionally, we aim to integrate multimodal features, such as skeletal joint trajectories, optical flow, and motion energy maps, to provide richer dynamic and geometric context. These modalities will enable the model to better differentiate between visually similar but semantically distinct actions, improving segmentation accuracy and temporal coherence in motion sequences.

TABLE IV

PARAMETER SCALE, RUNTIME, AND ACCURACY OF DIFFERENT LLMS FOR TVS LEARNING ON THE WEIZ DATASET (SUBJECT "IDO")

Method	Parameters (B)	Runtime/question (s)	Acc (%)
GPT-o1	175	1.84	88.27
DeepSeek-v3-2-exp	67	0.95	85.71
Claude-Sonnet-4-5-20250929	70	0.72	86.88
Gemini-2.0-Flash-exp	120	0.36	87.47
Grok-4	314	3.69	84.17
Qwen3-235B-a22b	235	2.20	87.52

C. Effective Analysis of the LLM-Based TVS

This section analyzes the effectiveness, interpretability, and generalization of the proposed LLM-based TVS framework. Section IV-C1 visualizes the generated TVS matrices to assess their ability to capture temporal adjacency and motion coherence. Section IV-C2 compares different LLMs to evaluate how model architectures affect segmentation accuracy. Finally, Section IV-C3 summarizes the limitations of LLM-based TVS inference.

1) Visualization of TVS From LLM: We employ *GPT-o1* to generate the TVS matrix **G**. Figure 7(a) presents an example TVS matrix from the *Weiz* dataset for the person "ido". The TVS generated by the LLM exhibits minor inconsistencies. For instance, in the sixth motion, frames *271–323* correspond to the same "side walk" motion in the ground truth, yet the LLM identifies frames *271–295* as depicting the same motion and *306–323* as depicting the same motion separately. As shown in Figure 7(b), such partial inconsistencies occur in five out of ten motion segments. Nevertheless, the LLM does not introduce false distinctions, as it never explicitly labels segments *271–395* and *306–323* as different motions. Consequently, these minor inconsistencies do not mislead the subsequent TVSH algorithm. For the motions 1, 3, 5, 8, and 10, the LLM produces nearly perfectly consistent segmentations, demonstrating its effectiveness in capturing semantic motion coherence and temporal adjacency.

2) Comparison With Different LLM Models: We further evaluate six widely used multimodal LLMs for TVS learning, including *GPT-o1*, *DeepSeek-v3-2-exp*, *Claude-Sonnet-4-5-20250929*, *Gemini-2.0-Flash-exp*, *Grok-4*, and *Qwen3-235B-a22b*. Table IV summarizes the parameter scales, runtime, and

segmentation accuracy of the proposed TVSH, evaluated on the Weiz "ido" video (488 frames) using these LLM models under identical prompt and API configurations. All experiments were conducted on a MacBook Air equipped with an M4 chip and 32 GB of memory. Overall, *Gemini-2.0-Flash-exp* achieves the fastest runtime, followed by *Claude-Sonnet-4-5-20250929* and *DeepSeek-v3-2-exp*, whereas *Grok-4* is the slowest due to its extremely large parameter size. *GPT-o1* and *Qwen3-235B-a22b* fall in the mid range, balancing accuracy and computational cost.

In terms of segmentation accuracy, *GPT-o1* attains the best overall performance (88.27%), closely followed by *Qwen3-235B-a22b* and *Gemini-2.0-Flash-exp*, which achieve comparable results. *Claude-Sonnet-4-5-20250929* and *DeepSeek-v3-2-exp* exhibit slightly lower accuracy, while *Grok-4*, despite its largest parameter scale, yields the lowest performance. These results indicate that a larger model size does not necessarily guarantee better temporal reasoning or motion understanding; rather, architectural design and multimodal alignment play a more decisive role. Furthermore, model runtime generally increases with parameter size, reflecting the trade-off between computational complexity and inference precision. Nevertheless, all evaluated LLMs enhance TVS quality over non-LLM baselines (76.51%), confirming that integrating multimodal reasoning effectively strengthens temporal semantics and motion segmentation performance.

3) Limitation: A notable limitation of the LLM-driven TVS is its tendency to misidentify a single motion as multiple distinct motions. This issue is evident in Figure 7(a), where the LLM splits a continuous motion into several segments. While using appropriate prompts can help mitigate this problem, it cannot be completely avoided. As a result, LLM-driven TVS cannot be directly applied to human motion segmentation in a straightforward manner. Additional steps, such as the proposed TVSH, are required. This issue becomes particularly pronounced when a motion involves multiple stages and occurs at high speed, but the camera's frame rate is low. Furthermore, different LLM models may lead to TVS matrices with varying accuracy, and calling the LLM API is typically time-consuming, with each query requiring between 0.36 and 3.69 seconds.

V. Conclusion

In this paper, we introduced a novel feedback-enabled subspace embedding approach for HMS, leveraging TVS embedded in human motion videos. We formulated the subspace embedding problem by integrating a temporal regularizer to capture the underlying temporal structure. Furthermore, we incorporated clustering with a temporal constraint to ensure that the clustering assignments reflect temporal characteristics. Finally, we developed a feedback-enabled framework to optimize the subspace embedding based on the segmentation results. Experimental results on benchmark datasets for HMS consistently demonstrated the superior performance of our approach compared to existing state-of-the-art techniques.

Appendix

A. Proof of Theorem 1

Let $\{\mathcal{N}_i\}_{i=1}^N$ be the temporal neighborhoods and define $A_{i\ell} = 1$ iff $\ell \in \mathcal{N}_i$, with $\deg(i) = |\mathcal{N}_i|$. Let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N] \in \mathbb{R}^{d\times N}$. The matrix $\mathbf{G} \in \mathbb{R}^{N\times N}$ is given by

$$G_{ii} = -\deg(i), \qquad G_{i\ell} = 1 \text{ if } \ell \in \mathcal{N}_i,$$
$$G_{i\ell} = 0 \text{ otherwise.} \tag{11}$$

Lemma 1: For every i,

$$(\mathbf{ZG})_{:i} = \sum_{\ell\in\mathcal{N}_i} (\mathbf{z}_i - \mathbf{z}_\ell). \tag{12}$$

Proof: By (11), the i-th column of $\mathbf{G}$ has $G_{ii} = -\deg(i)$ and $G_{\ell i} = 1$ for $\ell \in \mathcal{N}_i$. Hence

$$(\mathbf{ZG})_{:i} = \mathbf{z}_i(-\deg(i)) + \sum_{\ell\in\mathcal{N}_i} \mathbf{z}_\ell = \sum_{\ell\in\mathcal{N}_i} (\mathbf{z}_i - \mathbf{z}_\ell).$$

□

Define the isotropic graph total variation

$$\mathrm{GTV}_{\mathrm{iso}}(\mathbf{Z}; A) = \sum_{i=1}^N \left(\sum_{\ell\in\mathcal{N}_i} \|\mathbf{z}_i - \mathbf{z}_\ell\|_2^2 \right)^{1/2}. \tag{13}$$

By (12),

$$\|\mathbf{ZG}\|_{2,1} = \sum_{i=1}^N \Big\| \sum_{\ell\in\mathcal{N}_i} (\mathbf{z}_i - \mathbf{z}_\ell) \Big\|_2. \tag{14}$$

For each i, let $\mathbf{a}_\ell = \mathbf{z}_i - \mathbf{z}_\ell$. The triangle inequality and Cauchy–Schwarz give

$$\Big\| \sum_{\ell\in\mathcal{N}_i} \mathbf{a}_\ell \Big\|_2 \le \sum_{\ell\in\mathcal{N}_i} \|\mathbf{a}_\ell\|_2 \le \sqrt{\deg(i)} \left(\sum_{\ell\in\mathcal{N}_i} \|\mathbf{a}_\ell\|_2^2 \right)^{1/2}. \tag{15}$$

Summing (15) over i yields

$$\|\mathbf{ZG}\|_{2,1} \le \sqrt{\deg_{\max}}\, \mathrm{GTV}_{\mathrm{iso}}(\mathbf{Z}; A), \qquad \deg_{\max} = \max_i \deg(i). \tag{16}$$

Conversely, for each i,

$$\left(\sum_{\ell\in\mathcal{N}_i} \|\mathbf{z}_i - \mathbf{z}_\ell\|_2^2 \right)^{1/2} \le \sum_{\ell\in\mathcal{N}_i} \|\mathbf{z}_i - \mathbf{z}_\ell\|_2,$$

and, on degree-bounded graphs, the linear aggregation $\{\mathbf{a}_\ell\} \mapsto \sum_\ell \mathbf{a}_\ell$ is bounded. Therefore, there exist constants $c_1, c_2 > 0$, depending only on $\deg_{\max}$, such that

$$c_1\, \mathrm{GTV}_{\mathrm{iso}}(\mathbf{Z}; A) \le \|\mathbf{ZG}\|_{2,1} \le c_2\, \mathrm{GTV}_{\mathrm{iso}}(\mathbf{Z}; A). \tag{17}$$

Hence $\|\mathbf{ZG}\|_{2,1}$ is equivalent to the isotropic Graph-TV on the temporal graph. Moreover, $\|\mathbf{ZG}\|_{2,1} = 0$ if and only if $\mathbf{z}_i = \mathbf{z}_\ell$ for all $\ell \in \mathcal{N}_i$, implying that minimizers are piecewise constant on connected components induced by A, with discontinuities allowed only across missing edges.

B. Proof of Theorem 2

Let $\mathrm{eq}_k^\star \in \{0,1\}$ denote the true adjacencies on consecutive pairs, flipped independently with probability $p < \frac{1}{2}$ to yield eq_k. Assume the minimum true segment length is $L_{\min} \geq 1$. The TVS graph $\mathbf{G}$ is built from $\{\mathrm{eq}_k\}$, and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]$ minimizes $\|\mathbf{ZG}\|_{2,1}$ together with other convex terms.

Let

$$\mathcal{F} = \{k : \mathrm{eq}_k \neq \mathrm{eq}_k^\star\}, \qquad F = |\mathcal{F}|.$$

By independence, $\mathbb{E}[F] = p(N-1) = O(pN)$. Each flip creates at most one spurious boundary or removes one true boundary, so if B_{err} is the number of erroneous TVS boundaries, $\mathbb{E}[B_{\mathrm{err}}] \leq 2\,\mathbb{E}[F] = O(pN)$. Isolated flips only affect a local neighborhood: a flipped within-segment edge creates a short notch, while a flipped boundary merges two segments at one position. The TVS penalty admits the nodewise form $\|\mathbf{ZG}\|_{2,1} = \sum_{i=1}^N \left\| \sum_{\ell \in \mathcal{N}_i} (\mathbf{z}_i - \mathbf{z}_\ell) \right\|_2$, which is equivalent to an isotropic graph total variation penalizing differences along present edges. On any true segment S with $|S| \geq L_{\min}$ and no flips, the minimum TVS cost is achieved by constant $\mathbf{z}_i$ on S. A single spurious cut splits S into two parts but leaves $\Omega(|S|)$ edges between them; any nonconstant split incurs an extra TVS cost of order $\Omega(L_{\min})$, whereas keeping $\mathbf{z}_i$ constant yields zero increase. Hence isolated spurious cuts are suppressed.

For small p, flips are sparse and multiple adjacent flips within one segment are exponentially unlikely. Together with $\mathbb{E}[B_{\mathrm{err}}] = O(pN)$, this implies that TVS minimization smooths out isolated errors and yields embeddings that are piecewise constant on the true segments in expectation when p is small and $L_{\min}$ is large.

C. Proof of Proposition 1

We first consider the first two term of (1), which is denoted as

$$f(\mathbf{Z}) = \|\mathbf{XZ} - \mathbf{X}\|_{\mathrm{F}}^2 + \|\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\|_1 = \|\mathbf{XZ} - \mathbf{X}\|_{\mathrm{F}}^2 + \mathbf{e}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\mathbf{e}$$

The columns of $\mathbf{X}$ are in general position: $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K]$, where all the columns of submatrix $\mathbf{X}_\alpha$ lie in the same subspace $\mathcal{S}_\alpha$.

Assume $\mathbf{Z}^*$ minimizes the function $f(\mathbf{Z})$, and we decompose $\mathbf{Z}^*$ to be the sum of two matrices

$$\begin{aligned}\mathbf{Z}^* &= \mathbf{Z}^D + \mathbf{Z}^C \\ &= \begin{bmatrix} \mathbf{Z}_{11}^* & & & \mathbf{0} \\ & \mathbf{Z}_{22}^* & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{Z}_{KK}^* \end{bmatrix} \\ &\quad + \begin{bmatrix} \mathbf{0} & \mathbf{Z}_{12}^* & \cdots & \mathbf{Z}_{1K}^* \\ \mathbf{Z}_{21}^* & \mathbf{0} & \cdots & \mathbf{Z}_{2K}^* \\ \vdots & & \ddots & \vdots \\ \mathbf{Z}_{K1}^* & \mathbf{Z}_{K2}^* & \cdots & \mathbf{0} \end{bmatrix}\end{aligned}$$

where $\mathbf{Z}_{ij}^* \in \mathbb{R}^{\mathcal{N}_i \times \mathcal{N}_j}$. Note that both $\mathbf{Z}^D$ and $\mathbf{Z}^C$ are nonnegative.

According to the decomposition of $\mathbf{Z}^*$, any column of $\mathbf{Z}^*$ can be written as $\mathbf{z}_i^* = \mathbf{z}_i^D + \mathbf{z}_i^C$, with $\mathbf{z}_i^D$ and $\mathbf{z}_i^C$ supported on disjointed subset of indices. We can write $\|\mathbf{XZ}^* - \mathbf{X}\|_{\mathrm{F}}^2$ as

$$\begin{aligned}\|\mathbf{XZ}^* - \mathbf{X}\|_{\mathrm{F}}^2 &= \sum_{i=1}^N \|\mathbf{Xz}_i^* - \mathbf{x}_i\|_2^2 = \sum_{i=1}^N \|\mathbf{Xz}_i^D + \mathbf{Xz}_i^C - \mathbf{x}_i\|_2^2 \\ &= \sum_{i=1}^N \|\mathbf{Xz}_i^D - \mathbf{x}_i\|_2^2 + \sum_{i=1}^N \|\mathbf{Xz}_i^C\|_2^2 \\ &\quad + 2\sum_{i=1}^N \cos\theta_i \|\mathbf{Xz}_i^D - \mathbf{x}_i\|_2 \|\mathbf{Xz}_i^C\|_2\end{aligned}$$

where θ_i is the angle between vector $\mathbf{Xz}_i^D - \mathbf{x}_i$ and $\mathbf{Xz}_i^C$.

Since the matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K]$ is well arranged, any column $\mathbf{x}_i \in \mathbf{X}_\alpha$ and $\mathbf{x}_j \in \mathbf{X}_\beta$ lie in different subspaces if $\alpha \neq \beta$. Let $\mathbf{x}_i \in \mathcal{S}_\alpha$, according to the definition of $\mathbf{z}_i^D$ and $\mathbf{z}_i^C$, we have $\mathbf{Xz}_i^D \in \mathcal{S}_\alpha$ and $\mathbf{Xz}_i^C \notin \mathcal{S}_\alpha$. Based on the orthogonal subspace assumption, we have $(\mathbf{Xz}_i^D - \mathbf{x}_i) \perp \mathbf{Xz}_i^C$ and $\theta_i = \pi/2$, thus

$$\begin{aligned}\|\mathbf{XZ}^* - \mathbf{X}\|_{\mathrm{F}}^2 &= \|\mathbf{XZ}^D - \mathbf{X}\|_{\mathrm{F}}^2 + \|_2\|\mathbf{XZ}^C\|_{\mathrm{F}}^2 \\ &\geq \|\mathbf{XZ}^D - \mathbf{X}\|_{\mathrm{F}}^2 \end{aligned} \tag{18}$$

Based on the nonnegativity of $\mathbf{Z}^*$, $\mathbf{Z}^C$, and $\mathbf{Z}^D$, we have

$$\begin{aligned}&\|(\mathbf{Z}^*)^{\mathrm{T}}\mathbf{Z}^*\|_1 \\ &= \sum_{i,j} |(\mathbf{z}_i^*)^{\mathrm{T}}\mathbf{z}_j^*| = \sum_{i,j} (\mathbf{z}_i^*)^{\mathrm{T}}\mathbf{z}_j^* = \sum_{i,j} (\mathbf{z}_i^C + \mathbf{z}_i^D)^{\mathrm{T}}(\mathbf{z}_j^C + \mathbf{z}_j^D) \\ &\geq \sum_{i,j} (\mathbf{z}_i^D)^{\mathrm{T}}\mathbf{z}_j^D + \sum_{i,j} (\mathbf{z}_i^C)^{\mathrm{T}}\mathbf{z}_j^C = \|(\mathbf{z}^D)^{\mathrm{T}}\mathbf{z}^D\|_1 + \|(\mathbf{z}^C)^{\mathrm{T}}\mathbf{z}^C\|_1 \\ &\geq \|(\mathbf{z}^D)^{\mathrm{T}}\mathbf{z}^D\|_1 \end{aligned} \tag{19}$$

From inequalities (18) and (19) we have $f(\mathbf{Z}^*) \geq f(\mathbf{Z}^D)$. Because $\mathbf{Z}_{ij}^* \in \mathbb{R}^{\mathcal{N}_i \times \mathcal{N}_j}$, we have $f(\mathbf{Z}^*) = f(\mathbf{Z}^D)$ and $\mathbf{Z}^C = \mathbf{0}$, thus $\mathbf{Z}^* = \mathbf{Z}^D$.

We then consider the third term in (1), namely, $g(\mathbf{Z}) = \lambda_2\|\mathbf{ZG}\|_{2,1} = \sum_{i=1}^N \sum_{l \in \mathcal{N}_i} \|\mathbf{z}_i - \mathbf{z}_l\|_2$. Given the subspace assumption for the samples, the construction of the neighbor set $\mathcal{N}_i$ for the ith sample based on cosine measurements captures all the temporal neighbors of the ith sample. It is therefore straightforward to demonstrate that $g(\mathbf{Z}) \geq \sum_{i=1}^N \sum_{l \in \mathcal{N}_i} \|\mathbf{z}_i^* - \mathbf{z}_l^*\|_2$, confirming that $\mathbf{Z}^*$ also minimizes $g(\mathbf{Z})$.

Finally, we address the last term in (1). With $\mathbf{Z}$ being block-diagonal, the (i, j)th element of $\mathbf{Z}$ is nonzero if and only if the ith and jth samples are situated in the same subspace. Consequently, $\|\mathbf{Z}^*\|_{\mathbf{Q}^*} = 0$.

D. Proof of Theorem 3

Let the sequence be partitioned into true motion segments $\{\mathcal{S}_g\}_{g=1}^K$ with $|\mathcal{S}_g| \geq L_{\min}$. Data in $\mathcal{S}_g$ lie near a subspace $\mathcal{U}_g \subset \mathbb{R}^D$ with variance σ^2, and subspaces satisfy $\Delta_{\mathrm{sub}}^2 = \min_{g \neq h} \mathrm{dist}^2(\mathcal{U}_g, \mathcal{U}_h) > 0$. Let $(\mathbf{Z}, \mathbf{Q})$ solve (1) with TVS matrix $\mathbf{G}$, built from noisy adjacencies obtained by independently flipping true labels with probability $p < \frac{1}{2}$. Denote the normalized segmentation error by $\mathrm{Err}_{\mathrm{HMS}}$.

The objective combines data-fitting/clustering terms with the TVS regularizer $\lambda_G\|\mathbf{ZG}\|_{2,1}$, yielding $\mathbb{E}[\mathrm{Err}_{\mathrm{HMS}}] \leq \mathbb{E}[\mathrm{Err}_{\mathrm{TVS}}] + \mathbb{E}[\mathrm{Err}_{\mathrm{sub}}]$. Let F be the number of flipped adjacencies; then $\mathbb{E}[F] = p(N-1)$ and the number of erroneous

TVS boundaries is $O(pN)$ in expectation. Since $\|\mathbf{ZG}\|_{2,1}$ is an isotropic graph total variation, it favors constant embeddings on long runs. Within any true segment of length at least $L_{\min}$, an isolated flip removes only one local edge while $\Omega(L_{\min})$ links remain, so any spurious split incurs an extra TVS cost of order $\Omega(L_{\min})$. Thus isolated boundary errors are suppressed and $\mathbb{E}[\mathrm{Err}_{\mathrm{TVS}}] \le C_1 \frac{p}{L_{\min}}$.

Within segments, noise of variance σ^2 and subspace separation Δ^2_{sub} imply a misassignment probability bounded by $C_2\,\sigma^2/\Delta^2_{\mathrm{sub}}$ under standard subspace perturbation arguments. The data-fitting and clustering terms promote block-diagonal embeddings, giving $\mathbb{E}[\mathrm{Err}_{\mathrm{sub}}] \le C_2 \frac{\sigma^2}{\Delta^2_{\mathrm{sub}}}$. Therefore, $\mathbb{E}[\mathrm{Err}_{\mathrm{HMS}}] \le C_1 \frac{p}{L_{\min}} + C_2 \frac{\sigma^2}{\Delta^2_{\mathrm{sub}}}$.

If the TVS neighborhoods match the true segments, the TVS graph has no spurious cross-segment edges; minimizing $\|\mathbf{ZG}\|_{2,1}$ enforces constant $\mathbf{z}_i$ within each segment, and $\Delta^2_{\mathrm{sub}} > 0$ prevents inter-segment coupling, yielding exact segmentation up to label permutation.

E. Proof of Proposition 2

Let $\mathbf{P} = \mathbf{ZG} - \gamma^{-1}\mathbf{F}$. Completing the square,

$$\frac{\gamma}{2}\|\mathbf{H} - \mathbf{ZG}\|_F^2 + \langle \mathbf{F}, \mathbf{H} - \mathbf{ZG}\rangle = \frac{\gamma}{2}\|\mathbf{H} - \mathbf{P}\|_F^2 + \text{const},$$

so (22) is equivalent to

$$\min_{\mathbf{H}} \ \|\mathbf{H}\|_{2,1} + \frac{\gamma}{2}\|\mathbf{H} - \mathbf{P}\|_F^2. \tag{20}$$

Since $\|\mathbf{H}\|_{2,1} = \sum_i \|\mathbf{h}_i\|_2$ and the quadratic term is separable, the problem decouples column-wise:

$$\mathbf{h}_i^\star = \arg\min_{\mathbf{h}} \|\mathbf{h}\|_2 + \frac{\gamma}{2}\|\mathbf{h} - \mathbf{p}_i\|_2^2.$$

The optimality condition yields the group soft-thresholding rule

$$\mathbf{h}_i^\star = \max\left(1 - \frac{1}{\gamma\|\mathbf{p}_i\|_2}, 0\right)\mathbf{p}_i, \qquad i = 1, \ldots, N. \tag{21}$$

Because (20) is strongly convex, (21) is the unique global minimizer. Hence the $\mathbf{H}$-update attains the global optimum of (22) at every iteration.

F. Proof of Proposition 4

The cost associated with a sample located in the k-th cluster, whose center is denoted by $\boldsymbol{\mu}_k$, can be expressed as the sum of the squared Euclidean distances between the sample and the cluster center, augmented by a term accounting for the influence of neighboring samples. Specifically, the cost is given by $\sum_{i\in\mathcal{C}_k}\left(\|\mathbf{u}_i - \boldsymbol{\mu}_k\|_2^2 + \eta\sum_{j\in\mathcal{N}_i}\|\mathbf{u}_j - \boldsymbol{\mu}_k\|_2^2\right)$, which can be expanded as $\sum_{i=1}^N \mathbb{1}(i \in \mathcal{C}_k)\|\mathbf{u}_i - \boldsymbol{\mu}_k\|_2^2 + \sum_{i=1}^N \mathbb{1}(i \in \mathcal{C}_k)\sum_{j=1}^N \mathbb{1}(j \in \mathcal{N}_i)\|\mathbf{u}_j - \boldsymbol{\mu}_k\|_2^2$. To simplify, we define $n_k(i)$, the number of times the i-th frame is considered a neighbor of samples in the k-th cluster, as $n_k(i) = \sum_{j\in\mathcal{C}_k}\mathbb{1}(i \in \mathcal{N}_j)$. Substituting this into the previous expression, the cost becomes $\sum_{i=1}^N \mathbb{1}(i \in \mathcal{C}_k)\|\mathbf{u}_i - \boldsymbol{\mu}_k\|_2^2 + \sum_{j=1}^N \eta n_k(j)\|\mathbf{u}_j - \boldsymbol{\mu}_k\|_2^2$. This formulation can be further compacted into the following expression $\sum_{i=1}^N \|\mathbf{u}_i - \boldsymbol{\mu}_k\|_2^2\,(\mathbb{1}(i \in \mathcal{C}_k) + \eta n_k(i))$. This equation reveals the total cost, which is the sum of the direct distance between each sample and the cluster center, and the weighted influence of its neighbors, with the weight determined by η. The term $n_k(i)$ quantifies how many times sample i is considered a neighbor within the k-th cluster.

G. Proof of Proposition 3

To prove that the given problem is equivalent to the spectral clustering problem, that is, solving $\min_{\mathbf{Q}}\ \mathrm{Tr}(\mathbf{Q}^\top(\mathbf{D} - \mathbf{A})\mathbf{Q})$, where $\mathbf{D}$ is a diagonal matrix with elements $\mathbf{D}_{j,j} = \sum_i A_{i,j}$, we begin by expanding the objective function. Recall that our problem is $\min_{\mathbf{Q}}\ \frac{1}{2}\sum_{i,j} A_{i,j}\|\mathbf{q}_i - \mathbf{q}_j\|_2^2$, which can be expanded as $\frac{1}{2}\sum_{i,j} A_{i,j}\|\mathbf{q}_i - \mathbf{q}_j\|_2^2 = \frac{1}{2}\sum_{i,j} A_{i,j}(\mathbf{q}_i^\top\mathbf{q}_i - 2\mathbf{q}_i^\top\mathbf{q}_j + \mathbf{q}_j^\top\mathbf{q}_j)$. This expression can be broken into three parts: $\frac{1}{2}\sum_{i,j} A_{i,j}\mathbf{q}_i^\top\mathbf{q}_i - \sum_{i,j} A_{i,j}\mathbf{q}_i^\top\mathbf{q}_j + \frac{1}{2}\sum_{i,j} A_{i,j}\mathbf{q}_j^\top\mathbf{q}_j$. Now, observing each part, we can express it in matrix form. First, $\sum_{i,j} A_{i,j}\mathbf{q}_i^\top\mathbf{q}_i = \sum_i \mathbf{q}_i^\top\mathbf{q}_i\sum_j A_{i,j} = \sum_i \mathbf{q}_i^\top\mathbf{q}_i\mathbf{D}_{i,i} = \mathrm{Tr}(\mathbf{Q}^\top\mathbf{DQ})$, and similarly, $\sum_{i,j} A_{i,j}\mathbf{q}_j^\top\mathbf{q}_j = \sum_j \mathbf{q}_j^\top\mathbf{q}_j\sum_i A_{i,j} = \sum_j \mathbf{q}_j^\top\mathbf{q}_j D_{j,j} = \mathrm{Tr}(\mathbf{Q}^\top\mathbf{DQ})$. Finally, we have $-\sum_{i,j} A_{i,j}\mathbf{q}_i^\top\mathbf{q}_j = -\mathrm{Tr}(\mathbf{Q}^\top\mathbf{AQ})$. Thus, combining these parts, we arrive at the following: $\frac{1}{2}\sum_{i,j} A_{i,j}\|\mathbf{q}_i - \mathbf{q}_j\|_2^2 = \frac{1}{2}\mathrm{Tr}(\mathbf{Q}^\top\mathbf{DQ}) + \frac{1}{2}\mathrm{Tr}(\mathbf{Q}^\top\mathbf{DQ}) - \mathrm{Tr}(\mathbf{Q}^\top\mathbf{AQ}) = \mathrm{Tr}(\mathbf{Q}^\top\mathbf{DQ}) - \mathrm{Tr}(\mathbf{Q}^\top\mathbf{AQ}) = \mathrm{Tr}(\mathbf{Q}^\top(\mathbf{D} - \mathbf{A})\mathbf{Q})$. This establishes the equivalence between the given problem and the spectral clustering problem.

For fixed $(\mathbf{Z}, \mathbf{Q}, \mathbf{F}, \gamma)$, the $\mathbf{H}$-subproblem

$$\min_{\mathbf{H}} \ \|\mathbf{H}\|_{2,1} + \langle \mathbf{F}, \mathbf{H} - \mathbf{ZG}\rangle + \frac{\gamma}{2}\|\mathbf{H} - \mathbf{ZG}\|_F^2 \tag{22}$$

is the proximal operator of the $\ell_{2,1}$ norm and admits a closed-form solution.

H. Proof of Theorem 4

Recall the augmented problem and its scaled augmented Lagrangian $\mathcal{L}_\gamma$. At iteration t, the algorithm (exactly or non-increasingly) updates $\mathbf{Z}, \mathbf{H}, \mathbf{Q}$ blockwise and then performs

$$\mathbf{F}^{t+1} = \mathbf{F}^t + \gamma_t(\mathbf{H}^{t+1} - \mathbf{Z}^{t+1}\mathbf{G}), \qquad \gamma_t \uparrow \gamma_\infty \in (0, \infty).$$

Blockwise minimization gives

$$\mathcal{L}_{\gamma_t}(\mathbf{Z}^{t+1}, \mathbf{H}^{t+1}, \mathbf{Q}^{t+1}; \mathbf{F}^t) \le \mathcal{L}_{\gamma_t}(\mathbf{Z}^t, \mathbf{H}^t, \mathbf{Q}^t; \mathbf{F}^t),$$

and the standard ADMM identity (using the dual update and nondecreasing γ_t) yields some $c > 0$ such that

$$\mathcal{L}_{\gamma_{t+1}}(\mathbf{Z}^{t+1}, \mathbf{H}^{t+1}, \mathbf{Q}^{t+1}; \mathbf{F}^{t+1}) \tag{23}$$

$$\le \mathcal{L}_{\gamma_t}(\mathbf{Z}^t, \mathbf{H}^t, \mathbf{Q}^t; \mathbf{F}^t) - c\|\mathbf{H}^{t+1} - \mathbf{Z}^{t+1}\mathbf{G}\|_F^2. \tag{24}$$

Since $\mathcal{L}_\gamma$ is bounded below, it converges and hence

$$\|\mathbf{H}^{t+1} - \mathbf{Z}^{t+1}\mathbf{G}\|_F \to 0, \mathbf{F}^{t+1} - \mathbf{F}^t = \gamma_t(\mathbf{H}^{t+1} - \mathbf{Z}^{t+1}\mathbf{G}) \to \mathbf{0}. \tag{25}$$

The decrease of $\mathcal{L}_{\gamma_t}$ together with feasibility implies $\{(\mathbf{Z}^t, \mathbf{H}^t, \mathbf{Q}^t, \mathbf{F}^t)\}$ is bounded, so along a subsequence

$$(\mathbf{Z}^t, \mathbf{H}^t, \mathbf{Q}^t, \mathbf{F}^t) \to (\mathbf{Z}^\star, \mathbf{H}^\star, \mathbf{Q}^\star, \mathbf{F}^\star), \qquad \mathbf{H}^\star = \mathbf{Z}^\star\mathbf{G}.$$

Exact block optimality gives (at each t) the inclusions passing to the limit using $\gamma_t \to \gamma_\infty$ and (25) yields the primal–dual stationarity conditions at $(\mathbf{Z}^\star, \mathbf{H}^\star, \mathbf{Q}^\star; \mathbf{F}^\star)$. Therefore every accumulation point is a first-order stationary point; if the $\mathbf{Q}$-subproblem is solved globally, each limit point is a KKT point of the relaxed problem.

REFERENCES

[1] F. Zhou, F. De la Torre, and J. K. Hodgins, "Hierarchical aligned cluster analysis for temporal clustering of human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 582–596, Mar. 2013.

[2] M. Keuper, S. Tang, B. Andres, T. Brox, and B. Schiele, "Motion segmentation & multiple object tracking by correlation co-clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 140–153, Jan. 2020.

[3] R. Poppe, "Vision-based human motion analysis: An overview," *Comput. Vis. Image Understand.*, vol. 108, nos. 1–2, pp. 4–18, Oct. 2007.

[4] X. Zheng and W. Zhao, "Unsupervised action segmentation via fast learning of semantically consistent actoms," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 6270–6278.

[5] J. F.-S. Lin, M. Karg, and D. Kulić, "Movement primitive segmentation for human motion modeling: A framework for analysis," *IEEE Trans. Hum.-Mach. Syst.*, vol. 46, no. 3, pp. 325–339, Jun. 2016.

[6] S. Tierney, J. Gao, and Y. Guo, "Subspace clustering for sequential data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1019–1026.

[7] S. Li, K. Li, and Y. Fu, "Temporal subspace clustering for human motion segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4453–4461.

[8] L. Wang, Z. Ding, and Y. Fu, "Low-rank transfer human motion segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 1023–1034, Feb. 2019.

[9] T. Zhou et al., "Consistency and diversity induced human motion segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 197–210, Jan. 2023.

[10] Y. Bai, L. Wang, Y. Liu, Y. Yin, H. Di, and Y. Fu, "Human motion segmentation via velocity-sensitive dual-side auto-encoder," *IEEE Trans. Image Process.*, vol. 32, pp. 524–536, 2023.

[11] X. Cao, C. Zhang, C. Zhou, H. Fu, and H. Foroosh, "Constrained multi-view video face clustering," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4381–4393, Nov. 2015.

[12] Z. Xing and W. Zhao, "Block-diagonal structure learning for subspace clustering," *Expert Syst. Appl.*, vol. 285, Aug. 2025, Art. no. 127767.

[13] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14, 2001, pp. 849–856.

[14] M. Rahmani and G. Atia, "Innovation pursuit: A new approach to the subspace clustering problem," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2874–2882.

[15] Z. Xing and J. Chen, "Constructing indoor region-based radio map without location labels," *IEEE Trans. Signal Process.*, vol. 72, pp. 2512–2526, 2024.

[16] S. Wang, C. Li, Y. Li, Y. Yuan, and G. Wang, "Self-supervised information bottleneck for deep multi-view subspace clustering," *IEEE Trans. Image Process.*, vol. 32, pp. 1555–1567, 2023.

[17] S. Wang, Z. Lin, Q. Cao, Y. Cen, and Y. Chen, "Bi-nuclear tensor Schatten-p norm minimization for multi-view subspace clustering," *IEEE Trans. Image Process.*, vol. 32, pp. 4059–4072, 2023.

[18] Z. Chen, X.-J. Wu, T. Xu, and J. Kittler, "Fast self-guided multi-view subspace clustering," *IEEE Trans. Image Process.*, vol. 32, pp. 6514–6525, 2023.

[19] Y. Tang, Y. Xie, and W. Zhang, "Affine subspace robust low-rank self-representation: From matrix to tensor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9357–9373, Aug. 2023.

[20] Y. Chen, S. Wang, Y.-P. Zhao, and C. L. P. Chen, "Double discrete cosine transform-oriented multi-view subspace clustering," *IEEE Trans. Image Process.*, vol. 33, pp. 2491–2501, 2024.

[21] Z. Xing and W. Zhao, "Segmentation and completion of human motion sequence via temporal learning of subspace variety model," *IEEE Trans. Image Process.*, vol. 33, pp. 5783–5797, 2024.

[22] X. Wang, D. Guo, and P. Cheng, "Support structure representation learning for sequential data clustering," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108326.

[23] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2004, pp. 819–826.

[24] A. Fod, M. J. Matarić, and O. C. Jenkins, "Automated derivation of primitives for movement classification," *Auto. Robots*, vol. 12, no. 1, pp. 39–54, Jan. 2002.

[25] J. Barbič, A. Safonova, J.-Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard, "Segmenting motion capture data into distinct behaviors," in *Proc. Graph. Interface*, 2004, pp. 185–194.

[26] P. Beaudoin, S. Coros, M. V. D. Panne, and P. Poulin, "Motion-motif graphs," in *Proc. ACM SIGGRAPH/Eurographics Symp. Comput. Animation*, 2008, pp. 117–126.

[27] C. Lea, A. Reiter, R. Vidal, and G. D. Hager, "Segmental spatiotemporal CNNs for fine-grained action segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 36–52.

[28] F. De la Torre, J. Campoy, Z. Ambadar, and J. F. Cohn, "Temporal segmentation of facial behavior," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Jun. 2007, pp. 1–8.

[29] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[30] Y. Qin, X. Zhang, L. Shen, and G. Feng, "Maximum block energy guided robust subspace clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2652–2659, Feb. 2023.

[31] L. Wang, Z. Ding, and Y. Fu, "Learning transferable subspace for human motion segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 7422–7429.

[32] J. Cui, Y. Li, H. Huang, and J. Wen, "Dual contrast-driven deep multi-view clustering," *IEEE Trans. Image Process.*, vol. 33, pp. 4753–4764, 2024.

[33] Z. Xing and W. Zhao, "Block-diagonal guided DBSCAN clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 11, pp. 5709–5722, Nov. 2024.

[34] Y. Bai, L. Wang, Y. Liu, Y. Yin, and Y. Fu, "Dual-side auto-encoder for high-dimensional time series segmentation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2020, pp. 918–923.

[35] J. Guo et al., "From images to textual prompts: Zero-shot visual question answering with frozen large language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10867–10877.

[36] Y. Ge, X. Zeng, J. S. Huffman, T.-Y. Lin, M.-Y. Liu, and Y. Cui, "Visual fact checker: Enabling high-fidelity detailed caption generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 14033–14042.

[37] H. Zhi et al., "LSceneLLM: Enhancing large 3D scene understanding using adaptive visual preferences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 3761–3771.

[38] D. Zheng, S. Huang, and L. Wang, "Video-3D LLM: Learning position-aware video representation for 3D scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 8995–9006.

[39] T. Kojima, S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 22199–22213.

[40] F. Li, D. Wang, J. Wu, and Y. Zuo, "LLMs as zero-shot graph learners: Alignment of GNN representations with LLM token embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 5950–5973.

[41] B. Feng et al., "Breaking down video LLM benchmarks: Knowledge, spatial perception, or true temporal understanding?," 2025, *arXiv:2505.14321*.

[42] R. Liu, C. Li, H. Tang, Y. Ge, Y. Shan, and G. Li, "ST-LLM: Large language models are effective temporal learners," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 1–18.

[43] Y. Yuan et al., "VideoRefer suite: Advancing spatial–temporal object understanding with video LLM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 18970–18980.

[44] M. Nie et al., "Slowfocus: Enhancing fine-grained temporal understanding in video LLM," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 81808–81835.

[45] G. Zhang et al., "A novel spatial–temporal learning method for enhancing generalization in adaptive video streaming," *IEEE Trans. Mobile Comput.*, vol. 24, no. 12, pp. 12852–12866, Dec. 2025.

[46] X. Ding and L. Wang, "Do language models understand time?," in *Companion Proc. ACM Web Conf.*, May 2025, pp. 1855–1868.

[47] A. Deng et al., "Seq2Time: Sequential knowledge transfer for video LLM temporal grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 13766–13775.

[48] L.-H. Chen et al., "MotionLLM: Understanding human behaviors from human motions and videos," 2024, *arXiv:2405.20340*.

[49] L. Li et al., "Human motion instruction tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 17582–17591.

[50] Z. Li, S. Deldari, L. Chen, H. Xue, and F. D. Salim, "SensorLLM: Aligning large language models with motion sensors for human activity recognition," 2024, *arXiv:2410.10624*.

[51] Y. Wang et al., "Scaling large motion models with million-level human motions," 2024, *arXiv:2410.03311*.

[52] B. Wang, Y. Tian, S. Wang, and L. Yang, "Multimodal large models are effective action anticipators," *IEEE Trans. Multimedia*, vol. 27, pp. 2949–2960, 2025.
[53] L. Lu, Y. Lu, R. Yu, H. Di, L. Zhang, and S. Wang, "GAIM: Graph attention interaction model for collective activity recognition," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 524–539, Feb. 2020.
[54] Z. Zhang, Q. Shen, Z. Hu, Q. Liu, and H. Shen, "Credit risk analysis for SMEs using graph neural networks in supply chain," in *Proc. Int. Conf. Big Data, Artif. Intell. Digit. Economy*, Jul. 2025, pp. 81–85.
[55] K. Li et al., "Research on reinforcement learning based warehouse robot navigation algorithm in complex warehouse layout," in *Proc. 6th Int. Conf. Artif. Intell. Comput. Appl. (ICAICA)*, Nov. 2024, pp. 296–301.
[56] B. Yan et al., "On protecting the data privacy of large language models (LLMs) and LLM agents: A literature review," *High-Confidence Comput.*, vol. 5, no. 2, Jun. 2025, Art. no. 100300.
[57] M. Sui, Y. Su, J. Shen, and W. Zhang, "Intelligent anti-money laundering on cryptocurrency: A CNN-GNN fusion approach," China Nanjing Univ. Posts Telecommun., Nanjing, China, Tech. Rep., 2026, doi: 10.22541/au.176824645.56752786/v3.
[58] Q. Wang, B. Huang, and Q. Liu, "Deep learning-based design framework for circular economy supply chain networks: A sustainability perspective," in *Proc. 2nd Int. Conf. Digit. Economy Comput. Sci.*, Oct. 2025, pp. 836–840.
[59] D. Yu et al., "Machine learning optimizes the efficiency of picking and packing in automated warehouse robot systems," in *Proc. Int. Conf. Comput. Eng., Netw. Digit. Commun.*, Mar. 2025, pp. 1325–1332.
[60] Z. Xing and W. Zhao, "Trajectory map-matching in urban road networks based on RSS measurements," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 4, pp. 4647–4660, Apr. 2025.
[61] Y. Yang, P. Hu, J. Shen, H. Cheng, Z. An, and X. Liu, "Privacy-preserving human activity sensing: A survey," *High-Confidence Comput.*, vol. 4, no. 1, Mar. 2024, Art. no. 100204.
[62] W. Zhao et al., "PointLIE: Locally invertible embedding for point cloud sampling and recovery," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1345–1351.
[63] J. Xi, W. Zhang, Z. Xu, S. Zhu, L. Tang, and L. Zhao, "Three-dimensional dynamic gesture recognition method based on convolutional neural network," *High-Confidence Comput.*, vol. 5, no. 1, Mar. 2025, Art. no. 100280.
[64] Z. Xing and W. Zhao, "Calibration-free indoor positioning via regional channel tracing," *IEEE Internet Things J.*, vol. 12, no. 5, pp. 5449–5461, Mar. 2025.
[65] W. Zhao, H. Zhang, C. Zheng, X. Yan, S. Cui, and Z. Li, "CPU: Codebook lookup transformer with knowledge distillation for point cloud upsampling," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 3917–3925.
[66] Z. Xing and J. Chen, "Blind construction of angular power maps in massive MIMO networks," *IEEE Trans. Signal Process.*, vol. 73, pp. 4539–4555, 2025.
[67] M. R. Kabir, F. S. Shishir, S. Shomaji, and S. Ray, "Digital twins in healthcare IoT: A systematic review," *High-Confidence Comput.*, vol. 5, no. 3, Sep. 2025, Art. no. 100340.
[68] Z. Xing and J. Chen, "Blind radio mapping via spatially regularized Bayesian trajectory inference," 2025, *arXiv:2512.13701*.
[69] J. Xu, K. Xu, K. Chen, and J. Ruan, "Reweighted sparse subspace clustering," *Comput. Vis. Image Understand.*, vol. 138, pp. 25–37, Sep. 2015.
[70] S. Boyd, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.
[71] R. H. Bartels and G. W. Stewart, "Solution of the matrix equation Ax+*$Xb*=$C [f4]," *Commun. ACM*, vol. 15, no. 9, pp. 820–826, 1972.
[72] K. Haynes, P. Fearnhead, and I. A. Eckley, "A computationally efficient nonparametric approach for changepoint detection," *Statist. Comput.*, vol. 27, no. 5, pp. 1293–1305, Sep. 2017.
[73] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[74] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.
[75] Z. Jiang, Z. Lin, and L. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 533–547, Mar. 2012.
[76] D. Huang, S. K. Yao, Y. Wang, and F. D. L. Torre, "Sequential max-margin event detectors," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 410–424.
[77] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1593–1600.
[78] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.
[79] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, 1955.
[80] K. Zhan, F. Nie, J. Wang, and Y. Yang, "Multiview consensus graph clustering," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1261–1270, Mar. 2019.

Zheng Xing (Member, IEEE) received the B.S. degree from the Ocean University of China, Qingdao, China, in 2017, the M.S. degree from Beihang University, Beijing, China, in 2020, and the Ph.D. degree from The Chinese University of Hong Kong, Shenzhen, China, in 2025. He is currently an Assistant Professor with the College of Computer Science and Software Engineering, Shenzhen University. He has authored several papers in prestigious journals and conferences, including IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE INTERNET OF THINGS JOURNAL, AAAI, ICC, GLOBECOM, WCNC, and ICASSP. His research interests encompass optimization, mobile computing, and machine learning.

Weibing Zhao (Member, IEEE) received the B.S. degree in computer science and technology from the School of Information Science and Technology, Beijing Normal University, Beijing, China, in 2018, and the Ph.D. degree from The Chinese University of Hong Kong, Shenzhen, China, in 2024. She is currently an Associate Professor with Shenzhen MSU-BIT University. She has contributed to multiple top conferences or journals, including ICCV, ECCV, AAAI, IJCAI, MM, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE INTERNET OF THINGS JOURNAL, and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. Her research interests include computer vision, point cloud analysis, medical image analysis, and image super-resolution.