# RDGFormer: Rough Domain Grassmann Transformer for Intention Prediction

Haibo Li, Qin Lai, Qicong Wang, Hongying Meng, *Senior Member, IEEE*

***Abstract*—The modeling of highly nonlinear dynamic relationships among humans, objects, and the environment from videos is a complex task. Moreover, extrapolating human intentions from these relationships presents an even greater challenge. To effectively parse such intentions, we propose the concept of intention prediction, which means to anticipate future interaction trends. In this paper, a novel Transformer-based intention prediction framework on Grassmann manifold has been proposed. Specifically, to decompose ambiguous video spatio-temporal dynamic information, a separation method based on rough sets has been proposed to decouple critical cues in distinct geometric spaces. To effectively capture topological-geometric information across diverse spaces, we introduce Grassmann manifold projection with learnable orthogonal bases for multi-scale topology learning. Finally, a homeomorphic evolution module with adaptive residual connections is proposed to refine the parametric embedding space. The experimental results on three datasets demonstrate that our proposed method not only achieves state-of-the-art (SOTA) performance in prediction accuracy, but also significantly outperforms other comparative methods in intention prediction.**

***Index Terms*—Rough Set, Grassmann Manifold, Homeomorphism, Intention Prediction.**

## I. Introduction

HUMAN intentions [1], [2] are abstract representations of goals or motivations shaped by individuals, environments, and objects. In complex interactions, intentions are expressed more ambiguously due to the intricate dynamics among multiple factors. While current research has explored intention recognition in images [3]–[6], video-based intention prediction remains understudied. We argue for explicitly encoding human intentions in video understanding, particularly through their manifestations in actions and object interactions across multiple temporal scales. Hence, we introduce the concept of intention prediction, which involves modeling the semantic relationship between humans and interacting objects in a given environment. To our knowledge, we are the first to propose an intention prediction task for videos.

In action recognition, anticipation, and video understanding, some methods focus solely on non-human contextual information [7]–[9] or exclusively on human [10] when recognizing motion processes, overlooking their combined influence. Intention prediction aims to integrate the environment, humans, and interacting objects to holistically model human intentions. This requires the model to capture dynamic human-object relationships and reason about complex interaction scenarios, enabling more accurate prediction of intentional dynamics shaped by human-object-environment interactions.

We observe that existing paradigms exhibit limitations when analyzed from two distinct dimensions: time and space. The time dimension characterizes the temporal interval between the observation and the prediction, challenging the model to infer future states from historical context. The space dimension, on the other hand, focuses on the semantic depth of the current interaction, which specifically involves the precise identification of verbs and nouns. Existing tasks typically exhibit an imbalance by focusing predominantly on one dimension. For example, long-term action anticipation emphasizes the time dimension, which aims to predict distant future events, but often at the cost of granular semantic detail. Conversely, tasks such as situation recognition and standard action anticipation prioritize the space dimension, targeting the recognition of immediate interactive semantics. However, standard action anticipation typically adopts a fixed anticipation time, which restricts the evaluation of the ability of a model to recognize the evolutionary changes of interactions across various temporal stages. Consequently, these tasks fail to comprehensively evaluate the ability of a model to understand the video as a whole.

To bridge this gap, our proposed intention prediction task is designed to simultaneously address both dimensions. We formally define two complementary sub-tasks: short-term intention prediction and long-term intention prediction. Both tasks take video sequences as input to forecast future intentions. By setting varying prediction horizons and evaluating fine-grained semantic labels, this task explicitly examines how well a model understands motion semantics and captures the dynamic evolution of human-object interactions.

In intention prediction, some clues may hinder reasoning, while others are beneficial for understanding the current situation. For example, as shown in Figure 1, although the two individuals in the first row exhibit similar actions and interactions with the ball, their intentions differ—one plans to play basketball, while the other plans to play rugby. Therefore, the model should predict the same actions, such as ball handling, but the corresponding interactive objects should be distinct—one will handle a rugby ball, and the other will handle a basketball. In the second row, despite significant differences in the actions and expressions of the

This work was supported by the National Natural Science Foundation of China under Grant No. 62571464, the Shenzhen Science and Technology Projects under Grant No. JCYJ20200109143035495, and the Natural Science Foundation of Fujian Province under Grant 2023J01003.

H. Li, Q. Lai and Q. Wang are with the Department of Computer Science and Technology, School of Informatics, Xiamen University, Xiamen 361102, China, and also with the Shenzhen Research Institute, Xiamen University, Shenzhen 518000, China (e-mail:qcwang@xmu.edu.cn).

H. Meng is with the Department of Electronic and Electrical Engineering, Brunel University London, UB8 3PH Uxbridge, U.K. (e-mail: hongying.meng@brunel.ac.uk).


Fig. 1. The impact of ambiguous cues on intention prediction

two individuals, both intend to score a goal. In this process, the action corresponding to their intention is shooting, and the interactive object is the football. The situation in the third row is more complex. Both the environmental and human information are somewhat ambiguous, and the clues from the environment and the person both point to football rather than catching the frisbee. This highlights the challenge of distinguishing between typical and ambiguous clues in the model, and when certain target cues become ambiguous, their semantic suggestions cannot be directly used for the inference process. Instead, a combination of multiple clues is required for a comprehensive judgment.

The need to model complex relationships—such as environmental context, interacting objects, and spatio-temporal dependencies of task subjects—leads to highly intricate geometric topologies in feature space, corresponding to dynamically evolving manifolds. While existing approaches across domains have attempted to model these relationships within specific manifolds, they consistently struggle to select appropriate manifold spaces for samples with ambiguous geometric-topological structures. This challenge significantly hinders learning performance in non-Euclidean spaces. In intention prediction, the rich and diverse semantics extracted from videos further compound this complexity. The intricate geometric and topological properties of the data make it particularly difficult to identify and adapt suitable manifold spaces. Consequently, effectively perceiving and interpreting these multi-modal semantic relationships remains a critical unresolved challenge.

To address the aforementioned challenges in intention prediction, we propose a Rough Domain Grassmann Transformer (RDGFormer) for modeling complex manifold relationships. Our approach consists of three key components: first, we introduce a learnable semantic separation module based on rough sets to decompose video content by directing heterogeneous relationships into different trajectory spaces; second, we develop a Grassmannian projection mechanism that maps features to adaptive Grassmann points, enabling explicit learning of geometric-topological relationships across trajectory subspaces while performing joint reasoning on different cues; finally, we construct a homeomorphic evolution module with dynamic residual connections to refine the latent space representation for robust intention reasoning.

The main contributions of this paper are as follows:

- We propose RDGFormer for the novel task of video intention prediction, effectively unifying short-term semantic understanding and long-term forecasting to achieve state-of-the-art performance.
- We design a separation mechanism based on rough sets to disentangle ambiguous video cues into distinct geometric spaces, enabling the model to filter noise and focus on critical semantics.
- We introduce Grassmann manifold projection to explicitly model the intrinsic geometric topology of dynamic interactions, enhancing reasoning capabilities in non-Euclidean spaces.
- We develop a homeomorphic evolution strategy with dynamic residuals to adaptively optimize the latent space during training, balancing model complexity and generalization.

## II. Formal Definition

Intention prediction aims to predict interaction intentions, focusing on forecasting the corresponding verbs, the objects of interaction (values), and the associated actions (verbs and values).

To clarify the distinction between intention prediction and traditional action anticipation related tasks, we emphasize two key differences. First, in traditional action anticipation, the anticipation time $t_a$ is typically fixed and dataset-dependent. We argue that this fixed setting limits the assessment of the ability of a model to perceive intentions across varying temporal spans. Second, long-term action anticipation tasks generally treat an action as a monolithic entity, often neglecting the granular understanding of verbs and nouns of a model. To address these limitations, we introduce intention prediction as a comprehensive task that draws upon action anticipation but extends it to evaluate the perception of human intention from multiple dimensions.

In the dynamic interaction process, the temporal dependencies between intentions are dynamic and uncertain. To this end, we further define the concept to be addressed as follows:

1) Short-term Intention Prediction: By observing past and current video segments, it predicts the verbs, values, and actions with objects starting at time $t_s$. The observation segment is the interaction process that occurs $t_o$ seconds before $t_a$, from time $t_s - (t_a + t_o)$ to $t_s - t_a$. Here, $t_a$ is the intention anticipation time, indicating how many seconds in advance the action is predicted.
2) Long-term Intention Prediction: The goal is to predict the verbs, values, and actions in the future $\mu T$ frames after observing $\nu T$ frames of the video. $\nu$ and $\mu$ represent the observation and prediction ratios of the entire video sequence $T$, respectively.

## III. Related work

### A. Action Anticipation Centered on Human Subjects

Action anticipation focuses on predicting future actions by analyzing human dynamics in videos. Early studies [11]–[13] have demonstrated initial success in forecasting actions by modeling temporal evolution patterns in video sequences. Nevertheless, the inherent challenges of high-dimensional sparsity in pixel space significantly complicate this task.

Some researchers have enhanced model performance by combining loss functions with temporal dynamics, for example, using self-distillation techniques and cycle consistency constraints to ensure consistency between forward and backward predictions [14], employing attention mechanisms to align and address semantic shifts to enhance predictions [15], and adopting mean squared error for predicting future features to encourage learning future representations [16], etc.

Analyzing long video sequences successfully preserves the temporal relationships between distant actions. Early works employed recurrent neural networks and their variants, such as long short-term memory networks and gated recurrent units, for sequential action processing [17]–[19]. However, due to the limitations of recurrent neural networks in handling very long sequences, subsequent work has shifted towards using Transformer architectures [20]–[23]. Recently, Yuan et al. [24] proposed a throughout procedural transformer to model the full temporal spectrum for anticipation. While effective in temporal modeling, this approach tends to overemphasize the temporal dimension at the expense of spatial semantic analysis, potentially overlooking the rich spatial cues crucial for fine-grained intention understanding.

Action anticipation requires predicting future actions and their categories, which introduces uncertainty. Ng et al. [25] proposed predicting entire action sequences instead of individual frames to avoid inconsistencies, and designed a new loss function to measure reliability. Piergiovanni et al. [26] developed an adversarial generative grammar model that produces diverse candidate sequences via stochastic rules for multi-hypothesis reasoning. However, these methods only treat uncertainty as an accompanying factor of the prediction results, without establishing an explicit mathematical modeling mechanism. To this end, we propose to use rough sets to effectively quantify and represent this kind of uncertain knowledge.

### B. Perception of Human Dynamic Processes in Non-Euclidean Spaces

Non-Euclidean spaces prove to be remarkably effective for representing the semantics of human actions in deep learning [27], [28]. This property indicates that some manifolds may align well with the geometric properties of the human body, whereas others may match the temporal characteristics of action sequences.

The geometric characteristics of human body movements, which are rich and complex, can be abstracted into a graph structure. Thus, researchers have explored processing and mapping based on the structural characteristics of data representations. Chen et al. [29] proposed a hyperbolic manifold-aware network that embeds dynamic sequences into hyperbolic space using the Poincaré model to capture hierarchical features and extract spatio-temporal information. Lin et al. [30] developed a manifold-guided graph update mechanism based on graph neural networks to describe spatio-temporal dependencies in frame sequences. Chen et al. [31] utilized hyperbolic space to simulate data structures and combined it with homotopy methods for efficient manifold learning.

Temporal information can typically be represented as covariance matrices or linear subspaces and embedded into symmetric positive definite (SPD) matrix manifolds or Grassmann manifolds. Zeng et al. [32] proposed a mixed modeling approach using SPD and Grassmann manifolds to extract deep representations. Ozdemir et al. [33] represented video data as tensors on the Grassmann manifold and introduced a new kernel to address the challenge of processing tensor data using matrix algebra. Wang et al. [34] constructed a stacked Riemannian autoencoder and incorporated residual blocks to improve the training accuracy of deep Riemannian networks. Kobler et al. [35] normalized data by computing the Fréchet mean and variance on the SPD manifold, outperforming tangent space approximation methods. These sequence-mapping methods primarily focus on discovering the relationships between elements within the sequence, enabling the capture of the topological structure of long sequences and the utilization of key frames relevant to dynamic information recognition. However, manifolds have fixed curvature and lack variable shapes, which has led to an intractable bottleneck in the non-Euclidean structured representation of action semantics. Recently, Wang et al. [36] introduced a Laplacian low-rank representation on product Grassmann manifolds for activity clustering, though it relies on static, non-learnable manifold points. Similarly, Wang et al. [37] proposed a transparent embedding space using Grassmann manifolds for interpretable image recognition, where class subspaces are treated as static points. These methods map data onto fixed, predefined manifold structures, limiting the ability of a model to adaptively learn the intrinsic geometric variations of dynamic video data.

### C. Human-Object-Environment Collaborative Perception and Interaction Recognition

Intention modeling fundamentally requires the synergistic integration of human, object, and environmental perception, as intentions emerge from dynamic multi-modal interactions rather than existing in isolation.

Human-Object Interaction (HOI) detection is a crucial task in computer vision. It involves localizing pairs of humans and objects in images and parsing their interaction relationships, holding significant value in the field of action understanding. Current research explores various dimensions of the interaction process [38]–[41]. However, the accuracy of prediction hinges entirely on the accuracy of localization, which in turn constrains the effectiveness of prediction.

Situation recognition [42] extends the scope of activity recognition and human-object-environment interaction by assigning semantic roles to define how actors, objects, and

environments participate in specific activities. This approach provides a more comprehensive understanding of images. Current works [3]–[5] have been able to effectively infer based on specific different roles and proposed methods adapted to this based on the knowledge graph [6]. However, the aforementioned research only considers single-frame images. Currently, related studies are shifting from single-frame analysis to long-term sequence modeling paradigms, adopting a direct action perception approach to understand ongoing actions [43]–[45].

Recent advancements have also explored the integration of multimodal information for intention understanding. Zhang et al. [46] introduced a multimodal intent recognition dataset (MIntRec) that combines text, video, and audio modalities, demonstrating that non-verbal cues significantly enhance recognition performance. Similarly, Chen et al. [47] proposed a domain-aware multimodal dialog system that leverages context-knowledge embeddings and user characteristics to improve intention modeling. These multimodal approaches offer promising directions. However, acquiring aligned and high-quality multimodal data, such as synchronized audio or textual descriptions, remains challenging in real-world scenarios. Consequently, our work focuses on video intention prediction.

Current approaches remain limited in their ability to jointly model the spatio-temporal dynamics of human-object-environment interactions. This critical gap hinders holistic understanding of interactive processes across varying time scales. Our proposed intention prediction framework addresses this limitation by explicitly integrating collaborative perception and multi-scale temporal reasoning.

## IV. Methodology

In this section, we present our proposed Rough Domain Grassmann Transformer (RDGFormer) for intention prediction. Firstly, we provide an overview of our proposed network and the designed loss function. Secondly, we introduce the fuzzy semantic separation method based on rough sets and correlation inference on the Grassmann manifold. Finally, a dynamic residual module is introduced that effectively optimizes the latent space of the model. Table I shows the notations and their corresponding definitions.

TABLE I
Notations and definitions

| Notations | Definitions |
|---|---|
| $PE$ | positional encoding |
| $S_0$ | video frame embeddings |
| $S'_m$ | GSROM module output in the m-th block |
| $S_m$ | output of the m-th RDGFormer block |
| $N_{query}$ | number of learnable queries |
| $\lambda$ | loss function weights |
| $L$ | loss function |
| $X$ | input samples |
| $IND$ | indiscernibility relation |
| $[x]$ | equivalence class of $x$ |
| $R_B$ | learnable classification relationship |
| $O_{max}$ | supremum of rough set feature extraction |
| $O_{min}$ | infimum of rough set feature extraction |
| $O_{mid}$ | center point of the equivalence class |
| $NR_B$ | negative region in rough set theory |
| $PR_B$ | positive region in rough set theory |
| $BND_B$ | boundary region in rough set theory |
| $\epsilon, \alpha, \beta, \gamma, \delta$ | learnable parameter |
| $\mathcal{G}(p, D)$ | Grassmann manifold space |
| $W \in R^{(d_{\text{in}} \times d_{\text{in}}) \times k}$ | weights of Grassmann projection heads |
| $Q \in R^{(d_{\text{in}} \times d_{\text{in}}) \times k}$ | k Grassmann learnable points |
| $P$ | Grassmann projection results |
| $O_G^X$ | Grassmann fused semantics representation |
| $O_{RS}^X$ | mixed semantic representation |
| $O^X$ | final output of the GSROM module |
| $t$ | learnable homeomorphism parameter |
| $\odot$ | hadamard product |
| $f(x)$ | target function on manifold |
| $\phi_i$ | dictionary atoms |
| $a_i(x)$ | expansion coefficients |
| $k$ | sparsity level / number of Grassmann points |
| $\hat{f}_k$ | approximation function with k terms |
| $N$ | intrinsic complexity of the function space |
| $M(x, t)$ | homeomorphic mapping function |
| $F(x)$ | complex manifold module function |
| $G(x)$ | simple module function |
| $\nabla$ | gradient operator |

### A. Overview

Given a video sequence $K \in \mathbb{R}^{C \times T \times H \times W}$, where $C$ is the number of channels in the video frames, $T$ is the temporal dimension, and $H$ and $W$ represent the height and width of the video frames, respectively. To obtain the video frame embeddings for convenient input into our model, we perform a 3D convolution operation on the input video frames and then transpose them for positional encoding to get the video frame embeddings $S \in \mathbb{R}^{d_s \times d_{in}}$, where $d_s = T' \times H' \times W'$ and $d_{in}$ is the dimension of the video frame embedding representation.

$$S_0 = Transpose(Conv3d(K)) \in \mathbb{R}^{d_s \times d_{in}} \tag{1}$$

Our proposed framework RDGFormer is shown in Figure 2. We follow the transformer architecture, where the video frame embedding $S_0$ is further processed with positional encoding before being fed into RDGFormer. The positional encoding is defined as follows:

$$\begin{aligned} PE(pos, 2i) &= \sin\left(\frac{pos}{10000^{2i/d}}\right) \\ PE(pos, 2i+1) &= \cos\left(\frac{pos}{10000^{2i/d}}\right) \end{aligned} \tag{2}$$

where $pos$ denotes the position in the sequence, and $i$ represents the dimension position in the positional encoding vector. The sequence $S_0$ is passed through RDGFormer blocks. Each RDGFormer block consists of a GSROM module we proposed, a residual connection with layer normalization, and a feedforward neural network. The GSROM is introduced into the learning process of fuzzy information, allowing it to continuously optimize during training. The propagation process of each module is as follows:

$$\begin{aligned} \boldsymbol{S}'_m &= Dropout\left(GSROM\left(LN\left(S_{m-1}\right)\right)\right) + S_{m-1} \\ \boldsymbol{S}_m &= Dropout\left(MLPs\left(LN\left(S_{m-1}\right)\right)\right) + S'_m \end{aligned} \tag{3}$$

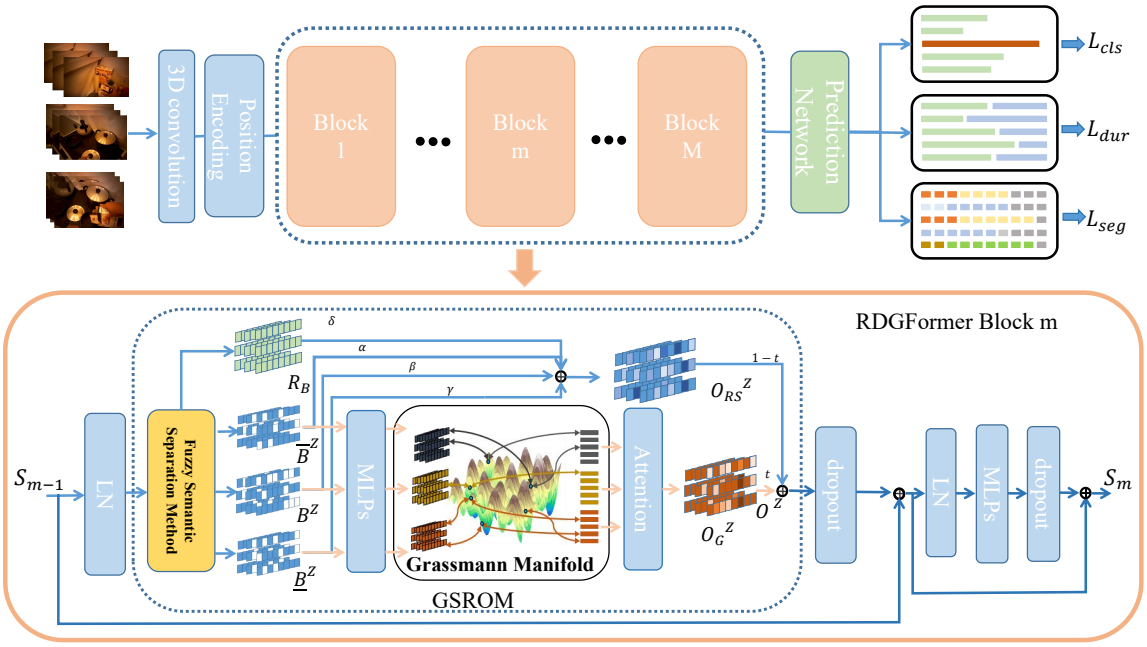



Fig. 2. The framework of RDGFormer is composed of multiple RDGFormer blocks. For the $m^{th}$ block, $S^{m-1}$ is the input to the model, and $S^m$ is the output of the model. We employ the fuzzy semantic separation method to extract various category information from video representations. This information is then projected onto an updatable Grassmann manifold points for learning the geometric and topological information of the corresponding manifold subspace. Finally, we utilize a dynamic residual module to optimize the latent space. The resulting representation is fed into a prediction network for loss function calculation, thereby achieving efficient intention prediction.

where $LN$ denotes layer normalization, $MLPs$ represents the multilayer perceptrons and $Dropout$ represents the dropout layer.

After being processed by multiple RDGFormer blocks, the video vector will be transformed through a prediction network to match the dimensions required for various loss function calculations. For short-term intention prediction, the prediction network consists of multiple parallel linear layers for dimensionality scaling. For long-term intention prediction, the prediction network is an RDGFormer block with cross-attention, which uses a learnable, specified number of queries to extract representations of the corresponding time steps. Subsequently, the vectors processed by the cross-attention RDGFormer block are scaled through different linear layers for the loss calculation of long-term intention prediction.

When constructing the supervised task for intention prediction, the design of the loss function is crucial for the learning and prediction accuracy. In this paper, two loss functions are proposed, one for predicting short-term intention and the other for predicting long-term intention. This ensures that the model can effectively capture the temporal characteristics of events and predict future event sequences.

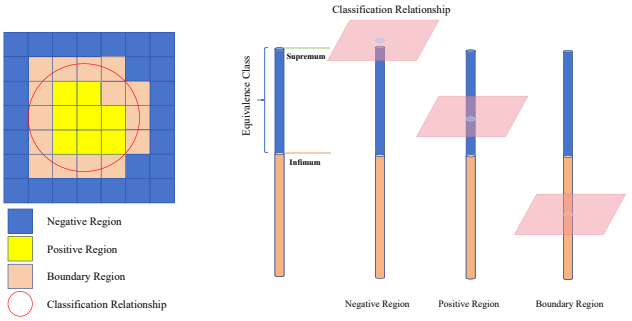
1) Loss function for short-term intention prediction

$$L_{\text{st}} = L_{\text{a}}^{st} + \lambda_1 L_{\text{v}}^{st} + \lambda_2 L_{\text{n}}^{st} \tag{4}$$

$L_{\text{a}}^{st}$ represents the cross-entropy function for action classification, where $a$, $v$, and $n$ represent action, verb, and noun, respectively.

2) Loss function for long-term intention prediction

$$L_{\text{lt}} = L_{\text{a}}^{lt} + \lambda_1 L_{\text{v}}^{lt} + \lambda_2 L_{\text{n}}^{lt} + \lambda_3 L_{\text{seg}}^{lt} + \lambda_4 L_{\text{dur}}^{lt} \tag{5}$$




(a) Rough Set Theory (b) Fuzzy Semantic Separation Method

Fig. 3. Visualization of Rough Sets:(a) The traditional rough set theory divides the universal set into upper approximation, lower approximation, and boundary region through the classification relation. Among these, the equivalence classes in the boundary region cannot be completely distinguished by the indiscernibility relation. (b) The proposed representation of equivalence classes under the rough domain employs a learnable classification relation $R_B$, which distinguishes the upper approximation, lower approximation, and boundary region.

We follow the settings of [48] and adopt action segmentation loss, duration loss, and label classification loss at a specified time step as the loss functions. Among them, $L_{\text{a}}^{lt}$, $L_{\text{v}}^{lt}$ and $L_{\text{n}}^{lt}$ have the same structure and are defined

as follows.

$$L_{\text{a}}^{lt} = -\sum_{i=1}^{N_{query}} \sum_{j=1}^{K+1} \mathbf{1}_{i \leq \Delta} A_{i,j} \log(\hat{A}_{i,j}) \tag{6}$$

In the context, $N_{\text{query}}$ represents the number of queries specified in the network architecture, which is also equal to the number of predicted time steps. $A_{i,j}$ represents the label of ground truth. $\hat{A}_{i,j}$ represents the predicted class probability by the model. $K+1$ represents $K$ action categories plus the "None" category. $\mathbf{1}_{i \leq \Delta}$ is an indicator function where $\Delta$ denotes the index of the first prediction result where the predicted category is "None".

$$\begin{aligned} L_{\text{seg}} &= -\sum_{i=1}^{T_O} \sum_{j=1}^{K} A_{i,j}^{\text{past}} \log \hat{A}_{i,j}^{\text{past}} \\ L_{\text{duration}} &= \sum_{i=1}^{N_{query}} 1_{\text{argmax}(\hat{A}_i) \neq \text{None}} (d_i - \hat{d}_i)^2 \end{aligned} \tag{7}$$

The event segmentation loss $L_{seg}$ employs the cross-entropy loss, where $A_{i,j}^{\text{past}}$ represents the ground-truth label of the action segmentation event in the observation, $\hat{A}_{i,j}^{\text{past}}$ is the predicted probability distribution by the model, and $T_O$ denotes the number of observed frames. The duration regression loss uses the L2 loss, where $d_i$ is the ground-truth duration, $\hat{d}_i$ is the predicted duration by the model.

### *B. Fuzzy Semantic Separation Method Based on Rough Sets*

If ambiguous information is incorporated alongside useful clues during training, the model may fail to identify the critical information required for accurate judgments. This idea comes from two main reasons:

1) Different types of information tend to interfere with each other when they interact. Ambiguous information often contains a large number of interfering clues. If the model cannot effectively distinguish and eliminate these interfering pieces of information that are not conducive to learning, it will have a negative impact on the overall understanding of the video.
2) Different types of mixed information typically exist in an extremely complex manifold space with highly intricate geometric topology, making it challenging for models to learn and understand effectively.

To overcome this challenge, we introduce a novel approach based on rough sets that segregates heterogeneous information into distinct geometrically-structured subspaces. This separation creates an optimized learning environment that enhances the ability of a model to identify and focus on the most discriminative features for accurate decision-making.

Rough set is a mathematical tool for dealing with incomplete and uncertain data. It defines the fuzzy parts in the data by describing the boundaries of knowledge through upper and lower approximations. For an information system, the indiscernibility relation can exist between any pair of sample points, $\forall x, y \in U$, The indiscernibility relation $IND(B)$ on the attribute set $B$ is defined as

$$IND(B) = \{(x, y) \in U \times U | f(x, a) = f(y, a), \forall a \in B\}. \tag{8}$$

The indiscernibility relation $IND(B)$ implies that objects under the indiscernibility relation cannot be distinguished. Therefore, the information can be partitioned through the equivalence relation $IND(B)$, resulting in multiple equivalence classes $U/IND(B) = \{X_1, X_2, \cdots, X_k\}$. Furthermore, the equivalence class containing object $x$ can be denoted as $[x]_B = \{y \in U | (x, y) \in IND(B)\}$. In rough set theory, two types of approximation operators are defined to describe fuzzy concept sets. The upper approximation $\overline{R}$ and the lower approximation $\underline{R}$ of samples set $X_B$ can be further defined:

$$\begin{aligned} PR_B(X_B) = \underline{R}(X_B) &= \cup\{[x]_B \in U/B | [x]_B \subseteq X\} \\ \overline{R}(X_B) &= \cup\{[x]_B \in U/B | [x]_B \cap X \neq \varnothing\} \end{aligned} \tag{9}$$

The negative region is the complement of the upper approximation, represents the set of elements that completely do not belong to the classification relationship. The boundary region is defined as the set of elements that cannot be distinguished. The negative region and the boundary region is defined as:

$$\begin{aligned} NR_B(X_B) &= U - \overline{R}(X_B) \\ BND_B(X_B) &= \overline{R}(X_B) - \underline{R}(X_B) \end{aligned} \tag{10}$$

We employ visualization to facilitate the understanding of rough set concepts in Figure 3(a). We can consider the part within the red circle as the range of set $X_B$, and this range corresponds to a classification relationship that divides the $U$ into three disjoint parts. For classification relationships, if an equivalence class can be completely distinguished, then it has a clear meaning and belongs to either the positive region or the negative region. Otherwise, it is considered to have fuzzy information and cannot be completely distinguished, thus belonging to the boundary region.

However, our analysis reveals that conventional rough set theory cannot be directly implemented for matrix operations in deep learning frameworks. To bridge this gap, we introduce the novel concept of a rough domain. Within this framework, we establish that elements in the rough domain can maintain independent semantic representations. Specifically, we employ the minimal matrix elements to approximate the equivalence class concept from rough set theory.

We assume that the input data is $X$ and define the supremum and infimum operators $Supremum_\theta$ and $Infimum_\xi$ as fine-grained rough set feature extraction modules. The structure consists of linear layers and batch normalization layers. The batch normalization module helps the sample representations in the rough domain space to undergo scaling and shifting, thereby achieving global alignment and enhancing the ability to measure uncertain knowledge. The space between supremum and infimum is defined as the range of the equivalence

class, and all elements within this range are considered to be part of the equivalence class of that information.

$$\begin{aligned} O_{max} &= \max\left(Supremum_{\theta}(X), Infimum_{\xi}(X)\right) \\ O_{min} &= \min\left(Supremum_{\theta}(X), Infimum_{\xi}(X)\right) \\ O_{mid} &= O_{min} + \frac{O_{max} - O_{min}}{2} \end{aligned} \tag{11}$$

To satisfy the strict size constraints of the supremum and infimum, we employ the min-max operation to re-integrate the information, ultimately obtaining the supremum $O_{max}$ and infimum $O_{min}$. $O_{mid}$ represents the center point of the equivalence class of the current position.

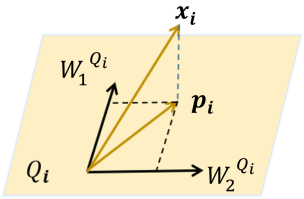



Fig. 4. Grassmann Manifold Projection: We project the original representations onto multiple Grassmannian subspaces $Q$ (each $Q_i$ defined by a set of orthogonal bases) and sum all the projected representations $p_i$ to obtain the final output.

Subsequently, we utilize the learnable classification relationship $R_B$ to partition the set composed of equivalence classes, as illustrated in Figure 3(b). We divide the corresponding positive and negative regions as well as the boundary class based on the degree of intersection between the supremum $O_{max}$ and infimum $O_{min}$ and $R_B$.

$$\begin{aligned} PR_B &= O_{mid} \odot PReLU(R_B - O_{max}) \\ NR_B &= O_{mid} \odot PReLU(O_{min} - R_B) \\ BND_B &= O_{mid} \odot (O_{max} - O_{min}) \end{aligned} \tag{12}$$

The magnitude of $R_B - O_{\text{max}}$ is designed to indicate the degree to which an equivalence class belongs to the positive region, and $O_{\text{min}} - R_B$ can indicate the degree to which an equivalence class belongs to the negative region. These two values serve as weights to be combined with the centers of the equivalence classes through a weighted sum to obtain $PR_B$ and $NR_B$. The boundary range can effectively capture the properties of the boundary class and is thus used for weighting.

### C. Correlation Inference on the Grassmann manifold

Diverse representations from rough set decomposition often correspond to manifolds with different geometric properties, which are abstract and variable. Thus, learning on a fixed manifold with constant curvature is not applicable. Our research reveals that the representations of the three concepts still have temporal alignment properties, which are well-suited for Grassmann space embedding. Therefore, we propose a learnable Grassmann space grouping embedding method. It groups and maps the representations to a space composed of multiple learnable Grassmann points for targeted learning. We also leverage the aligned properties of the learned representations for attention-based joint reasoning, effectively obtaining a fused representation with human-object-environment relationship.

A point $Y$ in the Grassmann manifold space can be defined as a matrix of size $D \times p$, whose columns consist of orthogonal bases. It parametrizes all possible subspaces of a given dimension in a vector space. Therefore, we can represent the Grassmann manifold as:

$$\mathcal{G}(p, D) = \left\{Y \in \mathbb{R}^{D \times p} : Y^T Y = I_p\right\} / \mathcal{O}(p) \tag{13}$$

where $\mathcal{O}(p)$ denotes the orthogonal group of order $p$. On the temporal dimension, data often exhibit different states or patterns, which can be regarded as subspaces at different time points. The Grassmann manifold, through its geometric structure, can naturally capture the changes and differences between these subspaces, thereby effectively representing the dynamic characteristics of time-series data. [49] has demonstrated that the curvature of the Grassmann manifold is not fixed. Therefore, we propose to employ learnable points in the Grassmann manifold space to facilitate the learning of geometric and topological properties in dynamic interactions.

To effectively learn the geometric and topological representations of rough set on the Grassmann manifold, we will construct learnable points in Grassmann manifold. These points are defined by the matrix $W \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}} \times k}$. We set the dimension of learnable points in the Grassmann manifold equal to the input dimension $d_{\text{in}}$ and $k$ is the number of the learnable points in the Grassmann manifold.

To compute the orthogonal matrix $Q_i$ for each $d_{\text{in}} \times d_{\text{in}}$ submatrix $W_i$ of the tensor $W$, we follow the Gram-Schmidt orthogonalization process. For each column vector $w_{ij}$ of $W_i$:

$$\begin{aligned} u_j^i &= w_{ij} - \sum_{m=1}^{j-1} \frac{w_{ij}^{\top} u_m}{u_m^{\top} u_m} u_m \\ Q_i &= [\frac{u_j^1}{\|u_j^1\|}, \frac{u_j^2}{\|u_j^2\|}, \ldots, \frac{u_j^{d_{in}}}{\|u_j^{d_{in}}\|}] \end{aligned} \tag{14}$$

Each learnable point in the Grassmann manifold has been transformed into a form that satisfies the geometric constraint conditions, which is represented by $Q \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}} \times k}$.

The information separated by the rough set process has complex topological structures. Therefore, we project the positive, negative, and boundary regions into the manifold for learning to adapt to the nonlinear geometric and topological properties. The projection formula is as follows:

$$P = Q(Q^T x) = \sum_{i=1}^{k} Q_i ({Q_i}^T x) = \sum_{i=1}^{k} \sum_{j=1}^{d_{\text{in}}} W_j^{Q_i} \left((W_j^{Q_i})^T x\right) \tag{15}$$

where $x_i$ represents the information of the i-th channel of the input vector. As shown in Figure 4, we use the orthogonal basis $W_j^{Q_i}$ corresponding to the variable-curvature point $Q_i$ in the linear subspace to compute the projection of the input vector. Specifically, we construct the projection matrix $Q_i Q_i^T$ which is symmetric and idempotent. It maps the input vector to the subspace, retaining its information within the subspace while removing components unrelated to the subspace, thus achieving topological feature extraction.

The projected result has a dimension of $d_s \times d_{\text{in}}$ and inherits the geometric properties of the manifold, which enhances the semantic relationships of the rough set in the nonlinear space and makes it more suitable for complex logical reasoning tasks.

The trajectory spaces separated by rough sets exhibit diverse geometric and topological properties. To address this, we map these spaces to distinct learnable points in the Grassmann space, thereby learning different geometric and topological information. We denote the representations of the positive region, negative region, and boundary class after MLPs and Grassmann manifold projection as $PR_B^{\mathcal{G}}$, $NR_B^{\mathcal{G}}$, and $BND_B^{\mathcal{G}}$, respectively.

To effectively jointly extract spatio-temporal representations in the Grassmann space, the negative region and boundary class are used as query and key, respectively, for attention computation. This further enables joint reasoning to obtain the auxiliary reasoning representation $O_{\mathcal{G}}^X$.

$$O_{\mathcal{G}}^X = \text{softmax}\left(\frac{{NR_B^{\mathcal{G}}}^T BND_B^{\mathcal{G}}}{\sqrt{d_{in}}}\right) PR_B^{\mathcal{G}} \tag{16}$$

Furthermore, we provide a theoretical basis for the number of Grassmann points $k$, grounded in Sparse Approximation theory [50]. Sparse approximation posits that a target function $f$ on a manifold can be approximated by a linear combination of atoms:

$$f(x) \approx \hat{f}_k(x) = \sum_{i=1}^{k} a_i(x)\phi_i(x) \tag{17}$$

where $\{\phi_i\}$ represents the dictionary of atoms, $a_i(x)$ denotes the expansion coefficients, and $\hat{f}_k(x)$ is the approximation with $k$ terms as the sparsity level. This formulation directly aligns with our Grassmann projection in Eq. (15): $P = \sum_{i=1}^{k} Q_i(Q_i^T x)$, where each Grassmann point $Q_i$ acts as an atom $\phi_i$ capturing specific interaction patterns, and the projection coefficient $Q_i^T x$ corresponds to $a_i(x)$.

The approximation error is theoretically bounded by the intrinsic complexity $N$ of the function space [51]:

$$\|f - \hat{f}_k\| \leq C_0 \cdot \min\left(1, \frac{N}{k}\right) \cdot \|f\| \tag{18}$$

where $\|\cdot\|$ denotes the function norm, and $C_0$ is a constant independent of $k$. This theorem implies that the error is proportional to $N/k$ when $k < N$. Thus, $k$ must be comparable to $N$ to minimize information loss.

In our framework, the function $f(x)$ models the mapping from video features to the intention manifold. Since complex intentions are composed of fundamental semantic units, we can assume that the intrinsic complexity $N$ corresponds to the number of distinct "action primitives" required to span this semantic space (i.e., the number of semantic labels). For example, EGTEA Gaze+ contains 15 verbs and 27 nouns, totaling 42 semantic labels; Breakfast contains 48 fine-grained action labels; 50 Salads contains 17 fine-grained action labels. Based on this, we can assume that the preliminary estimate of $N$ for each dataset is in the range of 17-48. Furthermore, increasing $k$ will introduce $d_{in}^2$ parameters for each point. Following the structural risk minimization principle, the optimal $k^*$ can balance the approximation error (decreasing with $k$) and model complexity (increasing with $k$). Therefore, it can be expected that $k^*$ will be slightly lower than the number of labels. From Figure 6(d), it can be seen that the experimental results are consistent with this expectation, and the observed optimal $k^*$ falls within the range of 15-20.

### *D. Dynamic Residual Module for Homeomorphic Evolution*

Given the complex parameter space inherent in Transformer-based models, learning various relationships in intention representation can be challenging, often leading to optimization difficulties during training due to the vast number of parameters. To effectively address this issue, we began our research by focusing on the training process. First, we examined the impact of different parameters on the training process. For the training process, the number of parameters and the learning rate are crucial to the learning progress.

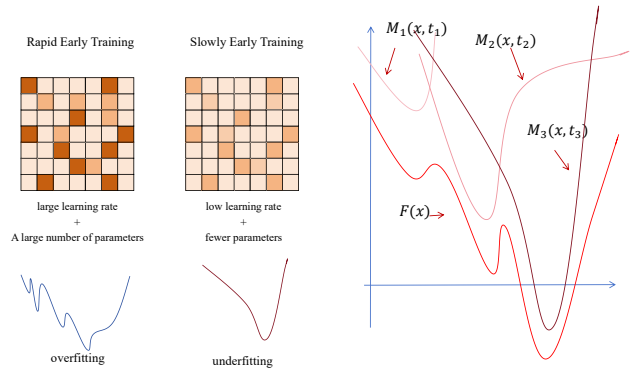
As shown in Figure 5(a), we use the shades of individual squares in a checkerboard pattern to represent the magnitude of the gradient. It is obvious that a large learning rate combined with an excessive number of parameters can lead to higher gradient values and interference among the parameters, resulting in unstable learning and even overfitting. On the other hand, a small number of parameters and a low learning rate often lead to underfitting, which in turn fails to capture the complex spatio-temporal relationships.

To address this, we came up with an idea: if we could dynamically determine the current learning rate and the parameters involved in decision-making, the model could adapt to its current state according to the training progress. This would help more complex models avoid overfitting. Some works [52] employ fixed weighted summation or residual connections for feature fusion, but these static strategies fail to adapt to training dynamics. To this end, we propose the dynamic residual module based on homeomorphic evolution, which, compared to the aforementioned methods, can automatically adjust the transition from simple to complex representations based on loss reduction efficiency.

The model is expected to converge to $F(x)$ as its learned function, define a simple module function as $G(x)$, and construct an explicit homeomorphic mapping $M_0(x,t)$.

$$M_0(x,t) = tF(x) + (1-t)G(x) \tag{19}$$

For $0 < t < 1$, $M_0(x,t)$ is a linear combination of $F(x)$ and $G(x)$. When $t$ is small, the influence of $G(x)$ is stronger, while that of $F(x)$ is relatively weaker. In our implementation, $t$ is initialized to $0.05$ and treated as a learnable parameter to be jointly updated with other network parameters through gradient descent. After each update, $t$ is truncated to the range of $[0,1]$ to satisfy theoretical assumptions about the coefficient range. The underlying dynamical principle is that for the function with weaker influence, most of its parameters do not actually participate in the direct computation of the final result. Thus, it can be considered that as $t$ varies from 0 to 1, the number of parameters directly acting on $M_0(x,t)$ gradually increases.




(a) different parameter initializations (b) Homeomorphic Evolution

Fig. 5. Schematic Diagram of Homeomorphic Evolution: We progressively introduce new prototype functions $M_i(x,t_i)$ based on existing $M_{i-1}(x,t_{i-1})$, enabling a gradual transition from simple to complex representations. This evolutionary process is designed with inspiration from homeomorphic transformations.

Note that $M_0(x,t)$ represents the initial state of a dynamic residual module. Assuming $M_{i-1}(x,t)$ has high solvability within its domain, in each learning iteration, the model can effectively learn a new function $M_i(x,t)$. Given that $M_{i-1}(x,t)$ has been learned, the condition $t<1$ indicates that some parameters in the encoder corresponding to the complex target function $F(x)$ have not been fully activated. By setting $t=t_{i-1}$ for the current step, we construct a new homeomorphic mapping function:

$$\begin{aligned} M_i(x,t) &= [(1-t)-(1-t_{i-1})]F(x)+(t-t_{i-1})G(x) \\ &+M_{i-1}(x,t_{i-1}) = (t-t_{i-1})\left[F(x)-G(x)\right]+M_{i-1}(x,t_{i-1}) \end{aligned} \tag{20}$$

The difference term $(t-t_{i-1})\left[F(x)-G(x)\right]$ plays a crucial role in the optimization process. It not only ensures the dynamic smoothness of the learning process but also achieves a gradual enhancement of the representational capabilities. As shown in Figure 5(b), our proposed method is capable of rapidly learning a homeomorphic function $M_0$. With the dynamic change of $t$, it gradually increases the number of parameters involved in decision-making and dynamically adjusts the learning progress. It then iteratively transitions to the next homeomorphic function and repeats this process until the desired outcome is achieved.

To strictly prove the effectiveness of the dynamic update of $t$, we analyze its gradient properties during backpropagation. Let $M(x,t)$ denote the general form of the homeomorphic mapping function (representing the current state $M_i(x,t)$). Let $L$ denote the global loss function. The gradient of $L$ with respect to the parameter $t$ can be derived via the chain rule:

$$\frac{\partial L}{\partial t} = \nabla_{M(x,t)}L \cdot \frac{\partial M(x,t)}{\partial t} = \nabla_{M(x,t)}L \cdot (F(x)-G(x)) \tag{21}$$

where $\nabla_{M(x,t)}L$ represents the gradient of the loss with respect to the module output $M(x,t)$. The parameter $t$ is updated via gradient descent: $t \leftarrow t-\eta\frac{\partial L}{\partial t}$. In the early stages of training, the simple module $G(x)$ converges faster, while the complex manifold module $F(x)$ is not yet fully optimized. As training progresses, $F(x)$ begins to capture complex topological structures that $G(x)$ cannot model, leading to a potentially lower loss state. Consequently, the direction from $G(x)$ to $F(x)$ aligns with the descent direction of the loss surface (i.e., $-\nabla_{M(x,t)}L$). This results in $\nabla_{M(x,t)}L \cdot (F(x)-G(x)) < 0$, making $\frac{\partial L}{\partial t}$ negative. Thus, the update term $-\eta\frac{\partial L}{\partial t}$ becomes positive, driving $t$ to increase. This internal dynamic mechanism ensures that the model automatically shifts its focus from the simple representation to the complex manifold reasoning as the latter becomes more effective. The term $F(x)-G(x)$ acts as a critical evolution indicator, guiding the gradient flow to prioritize the learning of complex structures that are not captured by the simple module, thereby facilitating a smooth transition from rough to fine-grained semantic reasoning.

Finally we propose the dynamic residual module for homeomorphic evolution. We define $F(x)$ as $O_{\mathcal{G}}^X$, while $G(x)$ is defined as the direct linear combination of multiple concepts. To enable the subsequent modules to effectively distinguish these different rough set concepts, we employ updatable tagging factors for alignment and weighting, thereby obtaining a final representation:

$$O_{RS}^X = \alpha \odot NR_B + \beta \odot PR_B + \gamma \odot BND_B + \delta \odot R_B \tag{22}$$

where $\alpha$, $\beta$, $\gamma$, and $\delta$ are learnable tagging factors used to control the scaling ratios of different parts. To ensure that the overall result represents the offset of various category information relative to the resolvable relationships, we incorporate $\delta \odot R_B$ to correct the overall distribution of the final representation, which fully utilizes the characteristics of positive and negative regions as well as fuzzy information. We have integrated the technologies we mentioned to build a Grassmann manifold-based Semantic Reasoning Optimization Method (GSROM). The final result we obtain is:

$$O^X = (1-t)O_{RS}^X + tO_{\mathcal{G}}^X = GSROM(X) \tag{23}$$

## V. Experiments and Analysis

### A. Datasets

We evaluate the proposed method on three datasets: EGTEA Gaze+ [53], Breakfast [54] and 50 Salads [55]. Here, EGTEA Gaze+ is utilized for short-term intention prediction, while Breakfast and 50 Salads are employed for long-term intention prediction. We divide action labels into verb labels and noun labels, which correspond to the actions of interaction and the objects being interacted with, respectively. Finally, the action labels, verb labels, and noun labels serve as the criteria for classification.

**EGTEA Gaze+:** The EGTEA Gaze+ dataset is a rich dataset for activity understanding and gaze prediction, containing 1,549 video sequences captured from 18 participants performing 104 unique cooking activities. Each video is annotated with detailed action labels and synchronized eye-tracking data, providing insights into where participants focus their attention. The dataset includes over 1.2 million frames with precise temporal action annotations and gaze points, making it a valuable resource for training and evaluating models in activity recognition and gaze prediction tasks.

**Breakfast:** The Breakfast dataset encompasses 1,712 video clips featuring 52 distinct individuals preparing breakfast across 18 unique kitchen settings. Each video is classified into one of the 10 activities associated with breakfast preparation. A total of 48 fine-grained action labels are utilized to delineate these activities. On average, each video spans approximately 2.3 minutes and encompasses around 6 actions. All videos have been resampled to a resolution of 240×320 pixels, with a frame rate of 15 fps. The dataset is divided into 4 separate training and test splits, and we report the average performance across all splits.

**50 Salads:** The 50 Salads dataset consists of 50 video recordings of 25 individuals making salads. This dataset includes over 4 hours of RGB-D video data, annotated with 17 fine-grained action labels and 3 high-level activities. Given that the videos in 50 Salads are typically longer than those in the Breakfast dataset, each video averages 20 actions. Every video in this dataset maintains a resolution of 480×640 pixels and a frame rate of 30 fps. The dataset is split into 5 training and test sets, and we report the average results across all these splits.

### *B. Experimental Setup*

We conducted multi-scale experiments on the intention prediction task. For Short-term Intention Prediction, the time horizon $t_a$ is set to 0.5s, 1s, and 2s. For Long-term Intention Prediction, the parameters $\nu$ are set to 0.2 and 0.3, while $\mu$ is set to 0.1, 0.2, 0.3, and 0.5.

**Architecture details:** For EGTEA Gaze+ datasets, we employed 12 RDGFormer blocks for training and testing, with a hidden layer size of 768. For the Breakfast and 50 Salads datasets, we used 3 and 4 RDGFormer blocks respectively, with hidden layer sizes of 128 and 512. For the multi-head attention mechanism, we employ a fixed number of heads set to 12. For long-term intention prediction, we employed multiple learnable query tokens to learn future labels. Specifically, we set 8 tokens for Breakfast and 20 tokens for 50 Salads.

**Training and testing:** For short-term intention forecasting, we utilize 16-frame segments at a resolution of 224×224. These segments are evenly extracted from a 64-frame observed video sequence, which lasts roughly 2 seconds. For long-term intention prediction, I3D features are used and the length of the input frames is not fixed. To facilitate a gradual model initialization, we set the initial value of $t$ in the Dynamic Residual Module to 0.05. For EGTEA Gaze+, we set the number of epochs to 100, while for Breakfast and 50 Salads, it is set to 160. The network utilizes a batch size of 16, and the Top-1 accuracy is evaluated for short-term intention prediction, while the mean over classes (MoC) accuracy is reported for long-term intention prediction. All training processes are carried out using 4 NVIDIA 4090 GPUs with 24 GB memory each and the learning rate is set to 0.0001 with Riemannian Adam optimizer.

### *C. Ablation Experiments*

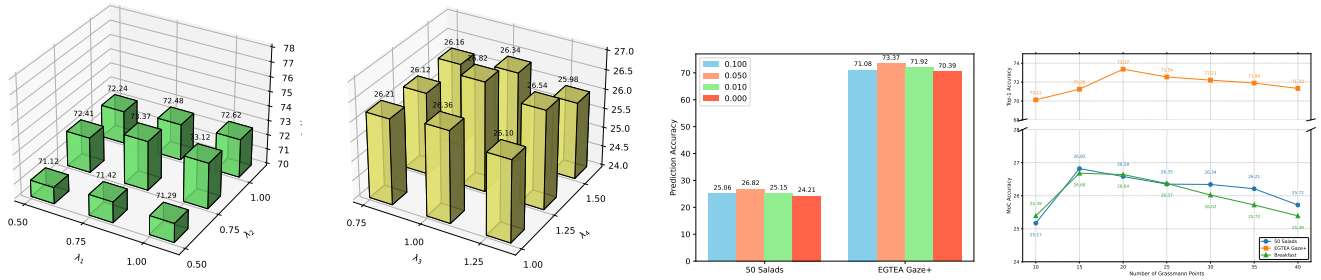
Ablation studies were executed on the EGTEA Gaze+ dataset and the 50 Salads dataset to assess the efficacy of each suggested component. In addition, we investigated the most effective arrangement of the principal parameters and furnished a thorough examination and dialogue on the outcomes of the experiments. We present the average Mean Over Classes (MoC) performance under different conditions specifically for the 50 Salads dataset.

TABLE II
ABLATION STUDY ON EGTEA GAZE+ DATASET AND 50 SALADS DATASET

| Methods | 50 Salads | | | EGTEA Gaze+ | | |
|---|---|---|---|---|---|---|
| | Action | Verb | Noun | Action | Verb | Noun |
| baseline | 22.95 | 74.06 | 70.25 | 68.52 | 79.01 | 78.54 |
| FSS | 23.78 | 75.43 | 71.07 | 69.76 | 80.49 | 79.35 |
| FSS+GM | 25.15 | 76.02 | 72.25 | 70.73 | 81.25 | 80.18 |
| FSS+GM+DRM | 26.82 | 76.61 | 72.73 | 73.37 | 83.03 | 81.25 |

**The effectiveness of the proposed modules:** As shown in Table II, we conducted ablation experiments on the three innovations we proposed. FSS represents the fuzzy semantic separation method based on rough sets, GM indicates correlation inference on the Grassmann manifold, and DRM stands for dynamic residual module for homeomorphic evolution. It is evident that our method achieves a more significant improvement in the action metric on both datasets. Through the series of innovations we employed, we realized a 4%–5% increase in terms of action metric, which effectively demonstrates the effectiveness of our proposed innovations. Specifically, FSS can decompose the complex information space, providing a foundation for learning geometric topological information using learnable Grassmann point projections into a simpler space. Therefore, the combination of FSS and GM can achieve a substantial performance improvement, which is effectively demonstrated in the 50 Salads dataset. However, we found that the performance improvement brought by the combination of the two is relatively limited in the EGTEA Gaze+ dataset. For instance, FSS + GM only brought a 1.64% increase in the Noun metric. We believe this is due to the optimization difficulties encountered by our method during the training process in larger datasets. Thus, our proposed DRM can effectively mitigate this issue and bring a more noticeable performance improvement.

**The impact of main parameters:** As illustrated in Figure 6(a), we investigated the loss weights of verb, noun and action on the EGTEA Gaze+ dataset. The experiment revealed that, although the semantics of action are complex, they significantly aid the learning of Verb and Noun. Consequently, the optimal weights for both the noun and the verb are less than 1, with the optimal value being 0.75 for each. However, the model is more sensitive to the weight of Noun during training, resulting in greater fluctuation of the result. This sensitivity arises because, in complex environments, the target cues of objects are more difficult to learn compared to human actions. As shown in Figure 6(b), we conducted a study on the weights of segmentation and duration loss in the 50 Salads dataset. The experiment revealed that the weights of these two losses have a relatively low impact on the final results. During the experiments, we set $\lambda_1$ and $\lambda_2$ to 0.75. The optimal weight for segmentation, $\lambda_3$, was found to be




(a) Verb loss weight $\lambda_1$ and Noun loss weight $\lambda_2$.

(b) Segmentation loss weight $\lambda_3$ and Duration loss weight $\lambda_4$

(c) Dynamic Residual Module $t$

(d) The number of Grassmann points $k$

Fig. 6. Ablation of different parameters on 50 Salads, Breakfast and EGTEA Gaze+.

TABLE III
LONG-TERM INTENTION PREDICTION: PREDICTING VERBS, VALUES, AND ACTIONS IN FUTURE $\mu T$ FRAMES AFTER OBSERVING $\nu T$ FRAMES ON THE 50 SALADS DATASET AND BREAKFAST DATASET

| | Datasets | 50 Salads | | | | | | | | Breakfast | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\nu$ | 0.2 | | | | 0.3 | | | | 0.2 | | | | 0.3 | | | |
| | $\mu$ | 0.1 | 0.2 | 0.3 | 0.5 | 0.1 | 0.2 | 0.3 | 0.5 | 0.1 | 0.2 | 0.3 | 0.5 | 0.1 | 0.2 | 0.3 | 0.5 |
| Action | FUTR [48] | 30.80 | 25.63 | 20.60 | 16.84 | 28.26 | 23.05 | 19.83 | 15.61 | 23.36 | 22.01 | 20.75 | 20.60 | 28.50 | 26.01 | 24.19 | 22.89 |
| | DiffAct [56] | 20.75 | 18.66 | 18.73 | 16.13 | 21.94 | 19.75 | 18.32 | 16.75 | 21.62 | 20.03 | 20.92 | 17.30 | 25.96 | 25.57 | 23.04 | 22.51 |
| | GTDA [57] | 18.82 | 18.61 | 18.81 | 16.57 | 18.62 | 19.49 | 21.07 | 17.27 | 18.18 | 17.24 | 16.76 | 16.17 | 20.37 | 20.56 | 20.11 | 18.32 |
| | RDGFormer | **31.42** | **29.15** | **24.08** | **19.35** | **38.87** | **27.71** | **23.37** | **20.56** | **26.74** | **24.59** | **23.88** | **21.92** | **33.18** | **29.41** | **27.79** | **25.94** |
| Verb | FUTR [48] | 85.83 | 79.05 | 71.49 | 58.96 | 89.81 | 81.84 | 73.66 | 61.83 | 79.57 | 67.18 | 61.05 | 54.46 | 84.77 | 76.18 | 71.99 | 63.61 |
| | DiffAct [56] | 78.73 | 66.23 | 59.07 | 49.29 | 85.97 | 75.30 | 68.15 | 58.05 | 78.92 | 66.17 | 60.00 | 52.57 | 85.82 | 75.65 | 68.87 | 59.49 |
| | GTDA [57] | 83.01 | 74.57 | 70.02 | **63.64** | 84.80 | 73.98 | 66.71 | 57.95 | 77.35 | 63.40 | 54.54 | 44.43 | 83.55 | 72.27 | 64.72 | 55.26 |
| | RDGFormer | **86.53** | **79.55** | **72.74** | 61.54 | **91.54** | **83.08** | **73.87** | **64.02** | **81.41** | **69.43** | **63.31** | **57.57** | **87.30** | **78.81** | **73.58** | **66.32** |
| Noun | FUTR [48] | **86.20** | 72.80 | 65.69 | 55.09 | 87.03 | 78.29 | **71.62** | 58.21 | 81.77 | 72.50 | 65.92 | 62.08 | 87.82 | 80.68 | 75.54 | 71.69 |
| | DiffAct [56] | 77.97 | 65.48 | 56.51 | 45.98 | 83.65 | 70.80 | 64.00 | 54.58 | 82.74 | 69.95 | 63.13 | 56.95 | 88.34 | 78.98 | 72.79 | 63.94 |
| | GTDA [57] | 81.28 | 67.31 | 57.52 | 47.04 | 83.50 | 69.66 | 60.01 | 50.14 | 79.92 | 67.58 | 59.66 | 50.63 | 85.78 | 75.60 | 68.55 | 60.17 |
| | RDGFormer | 83.73 | **75.06** | **67.02** | **56.32** | **88.03** | **79.83** | 71.54 | **60.32** | **84.01** | **74.71** | **69.26** | **64.62** | **90.86** | **84.03** | **79.87** | **75.15** |

TABLE IV
SHORT-TERM INTENTION PREDICTION PERFORMANCE COMPARISON ON THE EGTEA GAZE+ DATASET: ACTION, VERB, AND NOUN PREDICTION AT $t_a = 0.5$, $t_a = 1.0$, AND $t_a = 2.0$ SECONDS

| | Action | | | Verb | | | Noun | | |
|---|---|---|---|---|---|---|---|---|---|
| $t_a$ | 0.5 | 1.0 | 2.0 | 0.5 | 1.0 | 2.0 | 0.5 | 1.0 | 2.0 |
| AVT [16] | 43.02 | 41.66 | 40.04 | 54.92 | 53.15 | 52.63 | 52.57 | 51.83 | 50.23 |
| AFFT [58] | 42.54 | 41.07 | 39.83 | 53.46 | 51.84 | 50.44 | 50.46 | 39.84 | 38.15 |
| InAViT [59] | 67.89 | 66.96 | 64.17 | 79.33 | 77.91 | 76.53 | 76.35 | 75.21 | 73.52 |
| RDGFormer | **73.37** | **72.92** | **70.84** | **83.03** | **81.21** | **80.96** | **81.25** | **80.62** | **78.71** |

1.00, while the optimal weight for duration, $\lambda_4$, was 1.25. We believe this is because segmentation helps the model better understand different action categories, but it lacks information on future prediction. In contrast, duration focuses more on learning the length of actions, which enables the model to better capture temporal dependencies. Figure 6(c) shows that the optimal initial value of t is 0.05. When the value is too small, the training accuracy will be significantly reduced. We believe this is due to the small gradient, which prevents the dynamic residual module from learning effectively in the early stage. As shown in Figure 6(d), we performed ablation experiments on the number of Grassmann points $k$ across three datasets. The results indicate that the optimal value of $k$ falls within the range of [15, 20], which is consistent with our theoretical prediction based on the intrinsic complexity $N$ of the action space. Specifically, for the 50 Salads and Breakfast datasets, the performance peaks at $k = 15$. While 50 Salads has a naturally smaller label space (17 classes), the result for Breakfast (48 classes) is particularly interesting. We attribute this to its third-person viewpoint, which relies more on gross body poses rather than the complex hand-object topology found in first-person views. This implies a lower effective intrinsic complexity $N$, allowing a smaller $k$ to suffice. In contrast, for the EGTEA Gaze+ dataset, which combines a first-person view with a rich label set, the optimal $k$ increases to 20 to capture the intricate interaction details. These findings confirm that $k$ should be comparable to the effective intrinsic complexity to balance approximation error

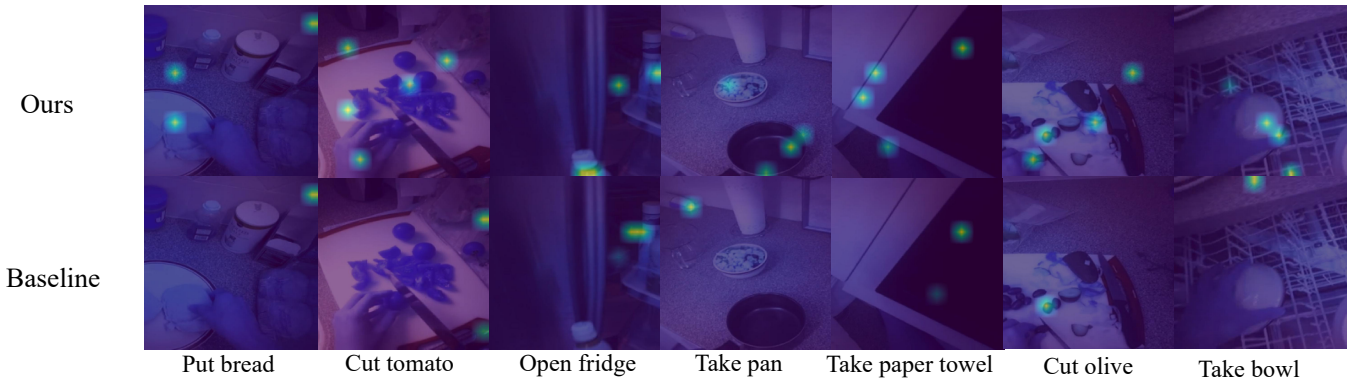



Fig. 7. Attention Map Visualization: Using the attention map visualization method to show the reasoning clues for a single frame in the video, the brighter areas have stronger weights.

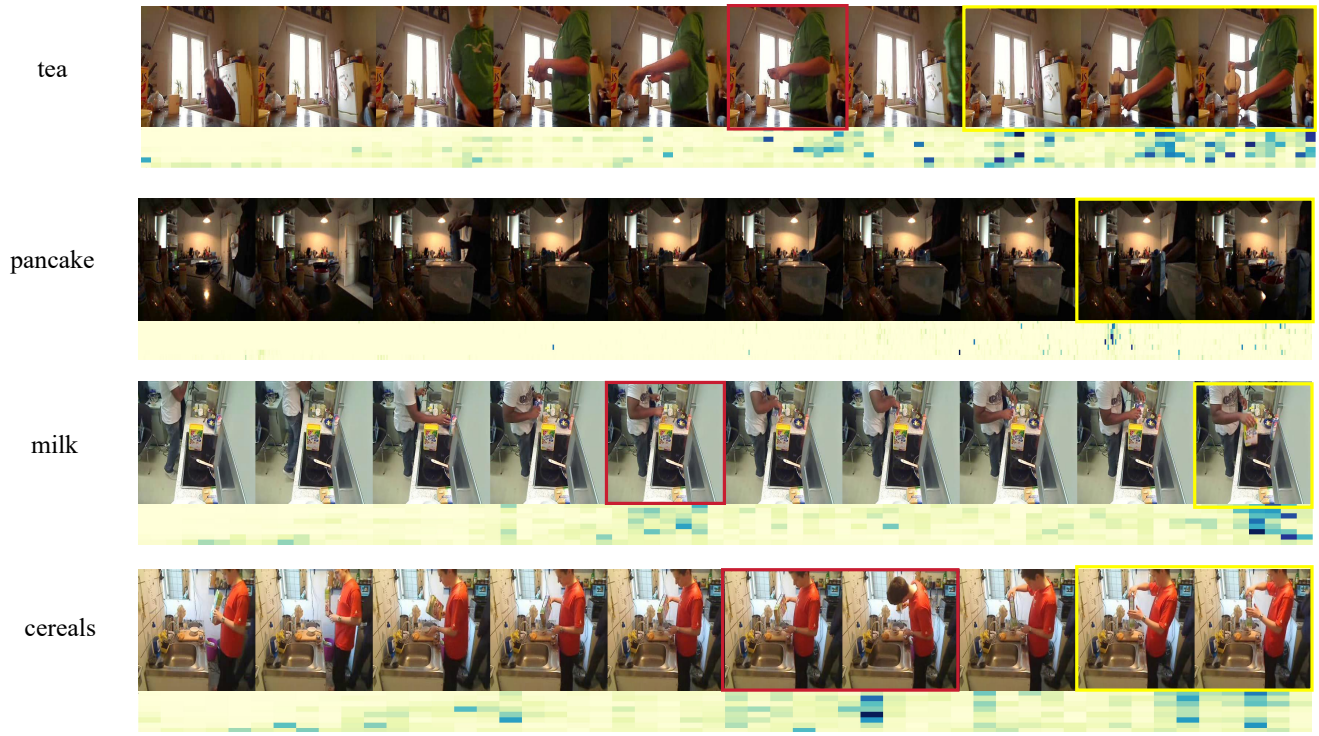



Fig. 8. Cross-Attention Map Visualization: Using the cross-attention map visualization method to display the weights of different video frames, the areas with deeper colors have stronger weights.

and model complexity.

### *D. Comparison with Existing Methods*

Given the novelty of intention prediction, we have adopted models that have demonstrated excellent performance in related fields for comparison. Specifically, for long-term intention prediction, we have modified the FUTR [48] and DiffAct [56], which have shown strong performance in long-term action anticipation and action segmentation, and GTDA [57], which excels in long-term dense anticipation. For short-term intention prediction, we have employed the AVT [16], AFFT [58], and InAViT [59], all of which have proven effective in action anticipation. To ensure a fair comparison, we appended an additional prediction network to the original architectures of these networks. This prediction network shares the same structure as ours, employing cross-attention mechanisms with multiple learnable query tokens designed to learn action duration and corresponding action labels. Furthermore, we integrated the same loss terms used in our approach into their original objective functions. Specifically, we set 8 tokens for Breakfast and 20 tokens for 50 Salads. These modifications align the supervision signals and training protocols, ensuring comparability across all methods. We provide a more comprehensive set of comparative experiments. Additionally, the performance of other methods is listed for reference. Table III shows the comparative experiments of our proposed method with other

methods on the long-term intention prediction task using the 50 Salads dataset and Breakfast dataset. Meanwhile, Table IV presents the comparative experiments of our proposed method with other methods on the short-term intention prediction task using the EGTEA Gaze+ dataset.

As shown in Table III, our model exhibits remarkable performance on the long-term intention prediction task. Typically, an increase in the observation rate from 0.2 to 0.3 allows models to access more video segments, which generally enhances prediction accuracy. However, on the 50 Salads dataset, we observed a counterintuitive phenomenon: the action accuracy sometimes decreased when the observation rate rose. This is attributed to the relatively small training set for each split in the 50 Salads dataset, which limits the ability of a model to generalize effectively when more segments are observed. Our model, however, successfully overcomes this challenge. For instance, when the prediction proportion is 0.1, the action prediction accuracy of our model increases by 7 percentage points as the observation rate rises from 0.2 to 0.3. In contrast, other models either maintain their performance or experience a decline. Moreover, our model consistently outperforms other models across most metrics, thereby validating the effectiveness of our proposed framework in the manifold space. During this process, we found that FUTR performed exceptionally well in verb and noun metrics, even surpassing our model in some Noun metrics. However, it lagged significantly behind our model in the action metric, a trend observed in other models as well. On the 50 Salads dataset, with an observation rate of 0.3 and a prediction proportion of 0.1, our method outperformed other methods by more than 10 percentage points. We speculate that most models have difficulty integrating actions, verbs, and nouns during the learning process. Although they can understand the meanings of verbs and nouns, they often fail to accurately predict the corresponding actions. Our model addresses this by separately modeling different information and conducting joint reasoning, which better captures the distinct information categories and achieves a unified understanding of action, verb, and noun.

Table IV presents the comparative experiments of our method with other models. We selected prediction times $t_a$of 0.5, 1.0, and 2.0 seconds to test different models. The overall results show that as the prediction time increases, the prediction accuracy under different metrics will decrease slightly. However, since the short-term prediction trends within 2 seconds usually do not change significantly, the accuracy drop is not substantial compared to long-term intention prediction. It is worth noting that all the comparison methods selected for this task utilize the Transformer architecture, which underscores the advanced nature of the Transformer in the fields of video understanding and prediction. In terms of performance, our method significantly outperforms the others. This advantage stems from our adoption of a semantic disentanglement approach based on rough sets which can quickly identify the most salient clues for inferring the current action from short video segments and perform collaborative reasoning with other target cues in the environment. As a result, our method boasts superior reasoning and real-time judgment capabilities. It is found that predicting action labels in long-term intention prediction is significantly more challenging than in short-term intention prediction. The reason is that in long-term prediction, the observed segments may include multiple different actions performed on the same object, or the same action performed on different objects. In contrast, behaviors in short-term intention prediction are generally more stable, which accounts for this difference.

### *E. Fine-grained Analysis*

In this section, we conduct a more detailed analysis to validate the effectiveness of the specific designs within our proposed modules. We compare our designs with standard alternatives or simplified versions to demonstrate the necessity of each component. The results on the Breakfast dataset are summarized in Table V.

TABLE V
FINE-GRAINED ABLATION STUDY ON BREAKFAST DATASET

| Settings | Metrics | | |
|---|---|---|---|
| | Action | Verb | Noun |
| **Semantic Separation** | | | |
| w/o Max-Min Constraint | 24.53 | 70.72 | 76.64 |
| Standard Self-Attention | 25.27 | 71.29 | 76.21 |
| **Manifold Reasoning** | | | |
| w/o Orthogonalization (Linear Average) | 25.52 | 71.63 | 76.96 |
| w/o Grassmann Projection (Euclidean) | 24.95 | 71.21 | 76.49 |
| **Homeomorphic Evolution Strategy** | | | |
| Fixed $t = 1$ (Only $F(x)$) | 25.18 | 71.82 | 76.92 |
| Fixed Ratio ($F(x) + x$) | 25.97 | 72.15 | 77.59 |
| **Ours (Full Model)** | **26.68** | **72.22** | **77.81** |

**Analysis of Rough Set Constraints:** We investigated the impact of the maximum and minimum constraints in our fuzzy semantic separation method. By removing the max-min comparison constraint (denoted as "w/o Max-Min Constraint"), the model loses the ability to strictly define the boundaries of equivalence classes. As shown in Table V, this leads to a performance drop, resulting in accuracy even lower than the standard self-attention mechanism ("Standard Self-Attention"). This indicates that the strict definition of upper and lower approximations is crucial for effectively distinguishing fuzzy semantics from clear cues.

**Analysis of Grassmann Orthogonalization:** To validate the necessity of learning on the Grassmann manifold, we disabled the orthogonalization process of the learnable Grassmann points. This effectively degrades the projection into a simple weighted summation of multiple linear layers (denoted as "w/o Orthogonalization"). Although this simplified version performs slightly better than completely removing the Grassmann component ("w/o Grassmann Projection"), it still falls short of our full model. The reason is that standard linear layers operate in Euclidean space, which treats all feature dimensions uniformly and often struggles to model the intrinsic nonlinear geometric structures of complex actions. In contrast, the orthogonal bases in the Grassmann manifold explicitly enforce a subspace structure, allowing the model to capture

the rotational invariant and topological relationships essential for distinguishing subtle intention differences.

**Analysis of Homeomorphic Evolution Strategy:** We further evaluated the impact of different evolution strategies for the dynamic residual module. We compared our dynamic $t$ strategy with two static alternatives: using only the transformed feature $F(x)$ (equivalent to fixing $t = 1$) and using a fixed ratio combination. The results show that both static methods perform much worse than our proposed method. If we fix $t = 1$, the model is forced to use the complex manifold mapping from the very beginning. This makes optimization difficult and causes early overfitting. On the other hand, using a fixed ratio cannot adapt to the learning progress. Dynamic adjustment of $t$ acts like a smooth transition, allowing the model to start with simple representations and gradually learn complex topological structures. This step-by-step learning strategy effectively prevents underfitting in the early stages and overfitting in the later stages, ensuring stable training.

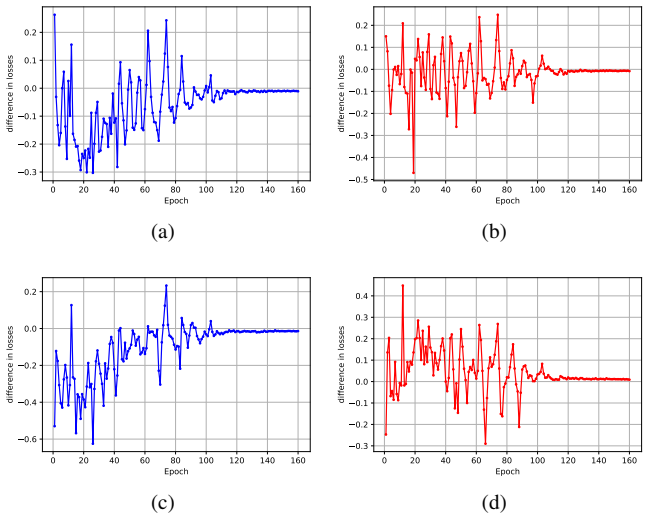



Fig. 9. (a) $L_{0.05} - L_{0.01}$, (b) $L_{0.05} - L_{0.10}$, (c) $L_{0.05} - L_{0.00}$, (d) $L_{0.05} - L_{0.50}$

### F. Visualization

To further investigate the effectiveness of our proposed method, we visualized the attention maps for a subset of videos processed by our model and annotated them with the corresponding action labels in Figure 7. Compared to the baseline, our method can assign higher weights to more clues that are helpful for prediction. For example, for the "Put bread" action, the focus of our proposed model is on the bread clues and the nearby condiments that may interact later. For the "Cut tomato" action, the attention is more concentrated on the already cut tomatoes, hand movements, and the cutting board and other interactive objects. For the "Take paper towel" action, the visualization results show that our model pays more attention to the edges of the cabinet, which can help the model understand the subsequent grasping actions and the geometric characteristics of the cabinet. In summary, our proposed method can better help the model focus on multiple clues that are useful for reasoning, thereby predicting future actions. The above visualization results show that our method has achieved target discovery and integrated reasoning from multiple clues.

To assess the temporal learning effectiveness, we believe that longer time sequences better capture this aspect. Therefore, we adopt long-term intention prediction and visualize the cross-attention maps in the prediction network of our model. The results are shown in Figure 8. We used four sets of video sequences with different themes, and selected 10 frames at equal intervals from each set as visualization samples. The width of the attention map is equal to the length of the input frames, while the height is equal to the number of queries. In the attention map, the darker the color, the more significant the information is in reasoning. To facilitate the description, we used different colors to highlight the corresponding target frames. For the video with the "tea" theme, it is found that the model focuses on the representative actions of putting tea leaves into the cup (highlighted in red) and turning to brew tea with hot water (highlighted in yellow). The action of brewing tea with hot water has a higher weight, indicating a strong relevance to the tea theme. For the "pancake" theme video, we observed that the attention weights are concentrated in the latter part. Upon inspection, we found that due to filming issues, the flour container completely blocked the person's hand movements and the interacting objects. When the flour container was moved at the end, revealing the person's hands, the attention weights increased significantly. This demonstrates that our model can effectively recognize ambiguous information and focus more on clear cues. For the "milk" theme video, the attention weights are concentrated on the actions of pouring milk (highlighted in red) and pouring a beverage (highlighted in yellow). We believe these two actions have strong event semantics. Similarly, for the "cereal" theme video, the attention is focused on the actions of pouring cereal and gently tapping the cereal box with the hand. Tapping the cereal box indicates that it is empty, which implies the subsequent action of throwing the cereal box into the trash.

By combining the results of the two types of attention visualization, our model demonstrates strong intention learning capabilities in both the spatial and temporal dimensions of the video, making it more suitable for intention prediction tasks.

### G. Analysis of the Dynamic Residual Module for Homeomorphic Evolution

To effectively demonstrate the impact of our proposed dynamic residual module for homeomorphic evolution on the training process, we adjusted the initial values of the dynamic parameters during training and analyzed their effects on the training loss. To clearly illustrate these differences, we visualized the loss differences under various configurations. Figure 9 shows the loss difference between the cases when we use different value of $t$.

We found that when the initial value of $t$ is set to 0.01 or even 0.00, the model exhibits a slower loss reduction in the early stages of training. During the early training phase (around epochs 5 to 40), the loss experiences some fluctuations and becomes slightly higher than when $t$ is initialized to

0.05. This suggests that with $t = 0.05$, more parameters are activated in the early stages compared to $t = 0.01$ and $t = 0.00$, which enhances the overall training performance initially.

Nevertheless, when $t$ equals 0.10, we found that the difference in loss compared to when $t$ equals 0.05 is not significant. Therefore, we adjusted $t$ to 0.50 for the subsequent experiments. We observed that when $t = 0.50$, the model achieves an even faster loss reduction in the early training phase compared to $t = 0.05$. However, during the mid-training period (epochs 40 to 100), the loss for $t = 0.50$ experiences some fluctuations. After a period of fluctuation, the model with $t = 0.50$ eventually converges to a smaller loss. Despite this lower loss, the performance on the test set is worse than when $t = 0.05$. We believe this is because, although a higher initial value of $t$ (such as 0.50) activates more parameters and leads to a rapid reduction in training loss early on, it fails to address the subsequent optimization issues in the latent space due to the simultaneous training of too many parameters. This ultimately results in optimization difficulties and overfitting.

## VI. Conclusion

In this paper, we have proposed a novel intention prediction framework named Rough Domain Grassmann Transformer (RDGFormer) to address the challenging task of predicting human intentions in complex scenarios. By leveraging rough set theory and Grassmann manifold, our method effectively disentangles and learns the diverse semantics from videos, capturing the intricate geometric and topological relationships among humans, objects, and the environment. The dynamic residual module further enhances the learning efficiency and optimization process. Extensive experiments on multiple datasets demonstrate the superior performance of our RDGFormer compared to other methods in both short-term and long-term intention prediction tasks. Our work not only provides a new perspective for intention prediction but also highlights the potential of combining rough set theory and manifold learning in complex video understanding. Future work may focus on exploring more advanced manifold representations and incorporating additional contextual information to further improve the prediction accuracy and robustness.

## References


[1] J. Xu, Y. Rao, J. Zhou, and J. Lu, "Transferable unintentional action localization with language-guided intention translation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[2] R. Uziel and O. Bialer, "Optimizing vision-language model for road crossing intention estimation," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 1702–1712.

[3] A. Mallya and S. Lazebnik, "Recurrent models for situation recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 455–463.

[4] M. Suhail and L. Sigal, "Mixture-kernel graph attention network for situation recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 363–10 372.

[5] T. Cooray, N.-M. Cheung, and W. Lu, "Attention-based context aware reasoning for situation recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4736–4745.

[6] W. Yu, H. Wang, G. Li, N. Xiao, and B. Ghanem, "Knowledge-aware global reasoning for situation recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8621–8633, 2023.

[7] Y. Luo, Z. Huang, Z. Wang, Z. Zhang, and M. Baktashmotlagh, "Adversarial bipartite graph learning for video domain adaptation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 19–27.

[8] B. Pan, Z. Cao, E. Adeli, and J. C. Niebles, "Adversarial cross-domain action recognition with co-attention," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 815–11 822.

[9] Y. Xu, H. Cao, K. Mao, Z. Chen, L. Xie, and J. Yang, "Aligning correlation information for domain adaptation in action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[10] K.-Y. Lin, J. Zhou, and W.-S. Zheng, "Human-centric transformer for domain adaptive action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[11] K.-H. Zeng, W. B. Shen, D.-A. Huang, M. Sun, and J. Carlos Niebles, "Visual forecasting by imitating dynamics in natural sequences," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2999–3008.

[12] J. Gao, Z. Yang, and R. Nevatia, "Red: Reinforced encoder-decoder networks for action anticipation," *arXiv preprint arXiv:1707.04818*, 2017.

[13] Y. Zhong and W.-S. Zheng, "Unsupervised learning for forecasting action representations," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1073–1077.

[14] M. Moniruzzaman, Z. Yin, Z. He, M. C. Leu, and R. Qin, "Jointly-learnt networks for future action anticipation via self-knowledge distillation and cycle consistency," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 7, pp. 3243–3256, 2022.

[15] V. Tran, Y. Wang, Z. Zhang, and M. Hoai, "Knowledge distillation for human action anticipation," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2518–2522.

[16] R. Girdhar and K. Grauman, "Anticipative video transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 505–13 515.

[17] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Forecasting future action sequences with neural memory networks," *arXiv preprint arXiv:1909.09278*, 2019.

[18] M. Xu, M. Gao, Y.-T. Chen, L. S. Davis, and D. J. Crandall, "Temporal recurrent networks for online action detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5532–5541.

[19] F. Sener, D. Singhania, and A. Yao, "Temporal aggregate representations for long-range video understanding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 2020, pp. 154–171.

[20] W. Wang, X. Peng, Y. Su, Y. Qiao, and J. Cheng, "Ttpp: Temporal transformer with progressive prediction for efficient action anticipation," *Neurocomputing*, vol. 438, pp. 270–279, 2021.

[21] C.-Y. Wu, Y. Li, K. Mangalam, H. Fan, B. Xiong, J. Malik, and C. Feichtenhofer, "Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 587–13 597.

[22] C. Cao, Z. Sun, Q. Lv, L. Min, and Y. Zhang, "Vs-transgru: A novel transformer-gru-based framework enhanced by visual-semantic fusion for egocentric action anticipation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[23] H. Girase, N. Agarwal, C. Choi, and K. Mangalam, "Latency matters: Real-time action forecasting transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 759–18 769.

[24] H. Yuan, Y. Chen, Z. Ji, Z. Zheng, Y. Gu, and J. Zhou, "Throughout procedural transformer for online action detection and anticipation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[25] Y. B. Ng and B. Fernando, "Forecasting future action sequences with attention: a new approach to weakly supervised action forecasting," *IEEE Transactions on Image Processing*, vol. 29, pp. 8880–8891, 2020.

[26] A. Piergiovanni, A. Angelova, A. Toshev, and M. S. Ryoo, "Adversarial generative grammars for human activity prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 507–523.

[27] Y. Lee, "Mmp++: Motion manifold primitives with parametric curve models," *IEEE Transactions on Robotics*, 2024.

[28] M. Zong, Z. Ma, F. Zhu, Y. Ma, and R. Wang, "Laplacian eigenmaps based manifold regularized cnn for visual recognition," *Information Sciences*, vol. 689, p. 121503, 2025.
[29] J. Chen, C. Zhao, Q. Wang, and H. Meng, "Hmanet: Hyperbolic manifold aware network for skeleton-based action recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 2, pp. 602–614, 2022.
[30] X. Li, C. Li, X. Wei, and F. Yang, "Manifold guided graph neural networks for skeleton-based action recognition in human computer interaction videos," in *2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*. IEEE, 2021, pp. 239–244.
[31] J. Chen, Z. Jin, Q. Wang, and H. Meng, "Self-supervised 3d behavior representation learning based on homotopic hyperbolic embedding," *IEEE Transactions on Image Processing*, vol. 32, pp. 6061–6074, 2023.
[32] X. Zeng, J. Guo, Y. Wei, and Y. Zhuo, "Deep hybrid manifold for image set classification," *Image and Vision Computing*, vol. 143, p. 104935, 2024.
[33] C. Ozdemir, R. C. Hoover, K. Caudle, and K. Braman, "Tensor discriminant analysis on grassmann manifold with application to video based human action recognition," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 8, pp. 3353–3365, 2024.
[34] R. Wang, X.-J. Wu, Z. Chen, T. Xu, and J. Kittler, "Dreamnet: A deep riemannian manifold network for spd matrix learning," in *Proceedings of the Asian conference on computer vision*, 2022, pp. 3241–3257.
[35] R. J. Kobler, J.-i. Hirayama, and M. Kawanabe, "Controlling the fréchet variance improves batch normalization on the symmetric positive definite manifold," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3863–3867.
[36] B. Wang, Y. Hu, J. Gao, Y. Sun, and B. Yin, "Laplacian lrr on product grassmann manifolds for human activity clustering in multicamera video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 554–566, 2016.
[37] J. Wang, H. Liu, and L. Jing, "Transparent embedding space for interpretable image recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 3204–3219, 2024.
[38] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4116–4125.
[39] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "Ppdm: Parallel point detection and matching for real-time human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 482–490.
[40] O. Ulutan, A. Iftekhar, and B. S. Manjunath, "Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 617–13 626.
[41] R. Li and B. Bhanu, "Energy-motion features aggregation network for players' fine-grained action analysis in soccer videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 955–972, 2023.
[42] M. Yatskar, L. Zettlemoyer, and A. Farhadi, "Situation recognition: Visual semantic role labeling for image understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5534–5542.
[43] T. Han, Y. Xu, J. Yu, Z. Yu, and S. Zhao, "Action-driven semantic representation and aggregation for video captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
[44] T.-D. Truong and K. Luu, "Cross-view action recognition understanding from exocentric to egocentric perspective," *Neurocomputing*, vol. 614, p. 128731, 2025.
[45] N. Nigam, T. Dutta, and H. P. Gupta, "Factornet: Holistic actor, object, and scene factorization for action recognition in videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 976–991, 2021.
[46] H. Zhang, H. Xu, X. Wang, Q. Zhou, S. Zhao, and J. Teng, "Mintrec: A new dataset for multimodal intent recognition," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 1688–1697.
[47] X. Chen, X. Song, J. Zuo, Y. Wei, L. Nie, and T.-S. Chua, "Domain-aware multimodal dialog systems with distribution-based user characteristic modeling," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 2, pp. 1–22, 2024.
[48] D. Gong, J. Lee, M. Kim, S. J. Ha, and M. Cho, "Future transformer for long-term action anticipation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3052–3061.
[49] T. Bendokat, R. Zimmermann, and P.-A. Absil, "A grassmann manifold handbook: Basic geometry and computational aspects," *Advances in Computational Mathematics*, vol. 50, no. 1, p. 6, 2024.
[50] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.
[51] R. A. DeVore, "Nonlinear approximation," *Acta numerica*, vol. 7, pp. 51–150, 1998.
[52] C. Liu and Y. Mu, "Multi-granularity interaction for multi-person 3d motion prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1546–1558, 2023.
[53] K. Min and J. J. Corso, "Integrating human gaze into attention for egocentric activity recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1069–1078.
[54] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 780–787.
[55] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013, pp. 729–738.
[56] D. Liu, Q. Li, A.-D. Dinh, T. Jiang, M. Shah, and C. Xu, "Diffusion action segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 10 139–10 149.
[57] O. Zatsarynna, E. Bahrami, Y. A. Farha, G. Francesca, and J. Gall, "Gated temporal diffusion for stochastic long-term dense anticipation," in *European Conference on Computer Vision*. Springer, 2024, pp. 454–472.
[58] Z. Zhong, D. Schneider, M. Voit, R. Stiefelhagen, and J. Beyerer, "Anticipative feature fusion transformer for multi-modal action anticipation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6068–6077.
[59] D. Roy, R. Rajendiran, and B. Fernando, "Interaction region visual transformer for egocentric action anticipation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6740–6750.

**Haibo Li** is currently pursuing the graduation degree with the Department of Computer Science and Technology, Xiamen University, Fujian, China. His research interests include computer vision, machine learning and rough set.

**Qin Lai** is currently pursuing the graduation degree with the Department of Artificial Intelligence, Xiamen University, Fujian, China. Her research interests include computer vision and machine learning.

**Qicong Wang** received the Ph.D. degree in information and communication engineering from Zhejiang University, Hangzhou, China. He is currently an Associate Professor at the Department of Computer Science and Technology, Xiamen University, Xiamen, China. His research interests include computer vision, machine learning, big data analytic.

**Hongying Meng** (M'10-SM'17) received his Ph.D. degree in Communication and Electronic Systems from Xian Jiaotong University, Xian China. He was an associate editor for IEEE Transactions on Circuits and Systems for Videos Technology (TCSVT) and IEEE Transactions on Cognitive and Developmental Systems (TCDS). He has authored over 200 publications including IEEE TIP, TCYB, TFS, TAC, TCSVT, TBE, TCDS, ICASSP and CVPR. He is currently a Professor at the Department of Engineering, Brunel University of London, U.K. His research interests include digital signal processing, affective computing, machine learning, human computer interaction, and computer vision.