2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Mining Intention with Heterogeneous Attention and Distillation for Interaction Anticipation

Yueke Zheng[1], Danhuai Zhao[2], Yicheng Liu[1], Kanghua Pan[1], Yuping He[1], Guo Chen[1], Guangpin Tao[1], Kang Zheng[2], Wei Zhu[2], Tong Lu[1]

[1] Nanjing University
[2] China Mobile Zijin Innovation Institute

Abstract. Short-term Object Interaction Anticipation (STA) aims to enhance intelligent systems with predictive capabilities and decision-making support by forecasting future interactions between objects. Existing methods often rely on visual changes in short video inputs, which limits the depth and accuracy of motivation prediction. In this study, we propose a novel approach, termed IntenFormer, inspired by how humans make decisions by understanding intentions. Specifically, IntenFormer employs a heterogeneous attention mechanism to simultaneously mine long- and short-term information frameworks, while incorporating knowledge distillation by utilizing a pretrained global intention model as a teacher, enabling the model to learn intention patterns. Extensive experiments on the Ego4D-STA dataset demonstrate that IntenFormer achieves highly competitive results, underscoring the efficacy of a unified approach to intention prediction and knowledge distillation.

Keywords: Object Interaction Anticipation, Knowledge Distillation.

1 Introduction

Short-term Object Interaction Anticipation (STA) is an important research task aimed at predicting future interactions between objects and agents, thereby providing a basis for behavior prediction and decision-making support [1]. The STA task involves taking a video and a specific timestamp as input and predicting the timing, location, and type of object interaction that may occur at that specific moment. Compared to traditional tasks such as action recognition and hand-grasp analysis, there is growing interest in predicting future human-object interactions based on egocentric (first-person perspective) videos. Such predictions can be embedded into AI systems to support users by enabling more accurate behavior predictions. For example, STA can be used to predict pedestrian trajectories in autonomous driving or to anticipate user actions in smart kitchens to provide safety alerts. [2][3] In scenarios requiring proactive judgment and rapid response, STA is critical.

Humans instinctively rely on understanding others' intentions to make predictions and decisions. In communication, we observe facial expressions and body language to infer emotions and intentions, thereby predicting subsequent behaviors and interactions. Studies have shown that incorporating the concept of intention into the processing

of egocentric videos can significantly enhance model performance [4], leading to more accurate behavior predictions [5]. However, existing methods [1] [6] [7] primarily rely on short-term observed physical changes, such as hand motion trends and gaze fixation rates, to predict intentions. These approaches are constrained by the length of input videos and the richness of observations, leading to overly localized and simplistic predictions of intention, which in turn limits the efficiency and accuracy of the task.

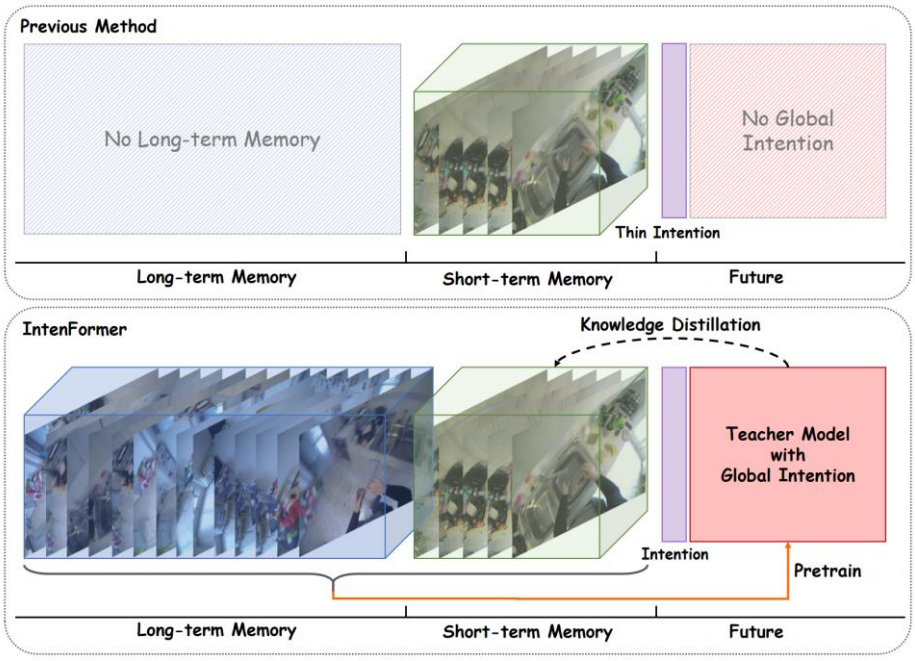

Fig. 1. Top: Previous methods could only use short-term memory and objective physical changes to predict implicit intentions. These methods often failed to uncover deeper subjective motivations. **Bottom**: Our **IntenFormer** additionally incorporates long-term memory and knowledge distillation, making it more capable of guiding action prediction from a macro perspective.

The effectiveness of intention prediction is often closely related to the richness of observational information. Based on this natural insight, we propose the core idea: combining long-term and short-term memory with knowledge-distilled global features to predict intentions. (see in **Fig. 1**). This approach offers the following 2 key merits:

- **Task-oriented prediction**: The model can accurately predict object interactions that are directly relevant to the current task. For example, an STA model that understands a user's historical behavior and current task objectives can accurately predict the next ingredient the user is likely to use while preparing a meal.
- **Personalized prediction**: The model can identify individual behavioral patterns. For instance, if the model learns from historical data that a specific user always wipes

their hands after using the kitchen sink, it can predict that the user is likely to wipe their hands again after finishing sink usage.

To this end, we propose an Intention-guided Transformer, termed **IntenFormer** for mining considered intentions to guide anticipation. Specifically, IntenFormer formulates the STA task as a set-prediction problem and defines a set of "Next-Active Objects" (NAO) queries to aggregate all intentions from memory and generate final predictions through a prediction head. To achieve effective intention modeling, IntenFormer leverages temporal attention and introduces a heterogeneous attention mechanism. It fully exploits local spatiotemporal features and dynamically models spatial information in short-term memory, while also extending historical information and broadening the model's perceptual scope in long-term memory. To further enrich the enhanced intentions, IntenFormer defines an additional set of queries that fully interact with all accumulated memories. These queries are supervised by global features obtained from a pretrained teacher model, enabling knowledge distillation for the NAO queries. Finally, the updated NAO queries are processed into final outputs through bipartite graph matching in the prediction head. We conducted extensive experiments on the Ego4D-STA dataset, including STAv1 and STAv2, to evaluate the performance of IntenFormer. Experimental results show that IntenFormer achieves highly competitive results compared to existing methods. Moreover, our ablation studies demonstrate the significant advantages of incorporating enhanced intentions into the STA task.

Our main contributions are as follows:

- We propose **IntenFormer**, an end-to-end model that jointly uses short-term memory, long-term memory, and the intention mined by them to complete STA task.
- A heterogeneous attention method is proposed, which not only fully exploits local spatiotemporal features but also extends historical information and broadens the model's perceptual scope.
- Furthermore, we integrate knowledge distillation to extract global features, enabling task-oriented prediction and personalized prediction, thereby enhancing the model's predictive capability.
- Our model has produced highly competitive results on Ego4D STA dataset and has proven the superiority of our approach through extensive ablation testing.

2 Related Work

2.1 Action Anticipation with Intention

Different prior works have explored incorporating Intention (termed as Goal in some papers) into the task of action prediction.

ANTGPT [8] proposed utilizing large language models to predict Goals in natural language form based on video action sequences extracted by the backbone. It then transforms these Goals into features using CLIP [9], attaches them to the input features, and feeds them into the prediction network for completion.

Method [10] proposed an action prediction model based on "Abstract Goal", which defined a new concept called “Abstract Goal” from observed visual feature sequences for action anticipation. Similar to Method [5], this Abstract Goal is designed as a distribution, whose parameters are estimated through a variational recurrent network. The model then samples multiple possible next actions and introduces a measure called goal consistency to determine which candidate action aligns best with the Abstract Goal.

Method [11] proposed an action prediction method based on "latent goal learning." It generates latent goals from observed video features using a stacked recurrent neural network and uses the learned latent goals to guide the generation of possible future actions. The loss is then calculated based on the quality of the predicted actions and the distance difference between the continuous predicted latent goals.

However, these methods are too coarse-grained and lack the ability to capture micro-information in interactions. In contrast, our proposed method not only captures macro trends but also extracts subtle details such as minute hand movements and gaze shifts from short-term memory, enabling more comprehensive predictions.

2.2 Short-term Object Interaction Anticipation

Various approaches have explored Short-Term Object Interaction Anticipation task.

Method [12] proposed analyzing the trajectories of objects to enhance anticipation performance. Method [13] explored predicting visual attention probability maps from images, taking into account features of both hands and objects. With the advent of Vision Transformers [14], the application of transformers has begun to be explored by researchers. Method [15] introduced causal modeling of video features, employing sequence modeling of frame features to interpret interactions in consecutive future frames. Method [16] developed a method for a long-term understanding of videos, using Multiscale Transformers to hierarchically focus on stored memories.

We propose a novel approach that leverages logically derived explicit intentions from long-term memory to guide predictions, while also incorporate global intentions to capture more nuanced details.

3 Methodology

3.1 Problem Definition

We define the STA task as follows. Given an egocentric video V and a specific timestamp t, the input to the model is the video segment $V_{1:t}$ up to the current time. The goal of the model is to predict future interactions, which are represented by a set: $\Psi = \{\phi_i = (b, n, v, \delta, c)\}_{i=1}^{N_p}$, where b denotes the object's location (bounding box), n is the object's category, v is the action performed with the object, δ is the time until interaction, and c is the confidence score of the prediction.

The video segments $\{V_{1:t_i}\}_{i=1}^{N_v}$ are defined by annotation or sampling points $\{t_i\}_{i=1}^{N_v}$, where the annotation points are consecutive points from the same scene.

3.2 Model Architecture

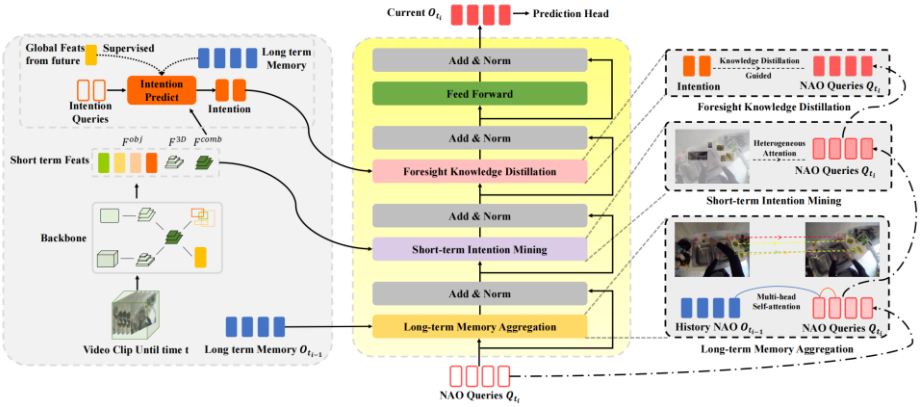

Fig. 2. A The overview of our IntenFormer. The light yellow section in the middle represents the main control flow, the light gray column on the left indicates the feature extraction and preparation of the data, and the boxes on the right visualize the specific modules.

As illustrated in **Fig. 2**, the input to **IntenFormer** consists of a video clip and the output from the previous time step $O_{t_{i-1}}$. The backbone processes the video clip and extracts 3 types of features: F^{obj} (object features), F^{3D} (3D motion features), and F^{comb} (combined features), which are used for the intention prediction task.

Next, we create a set of Next-Active Object (NAO) queries. We define these learnable parameters $Q_{t_i} \in R^{N_q \times C}$ to query for NAOs, where N_q is the number of NAOs we want to query, C is the dimensionality of the model, and t_i represents the current time. Following [17], we add learnable positional encodings to Q_{t_i} to enhance the predictive capability in the time sequence. These queries utilize a heterogeneous attention mechanism to interact with the 3 types of features and the previous output $O_{t_{i-1}}$, facilitating the mining of intentions from both long- and short-term memory. On the other hand, we introduce another set of intention queries Q_I, which query the long-term memory $O_{t_{i-1}}$ and the short-term memory F^{comb} to train the teacher model, and is supervised by global features from future. The knowledge obtained from this training is distilled into our model, enhancing the overall preformance.

Finally, a feedforward layer completes one iteration. Each iteration follows a residual connection and a LayerNorm layer [18]. After 3 iterations, the output O_{t_i} is determined. This output serves two purposes: it acts as the long-term memory for the next time step t_{i+1} and also enters the prediction head to complete the predictions.

3.3 Feature Extraction

We designed a dual-branch feature extraction method based on [6]. Given an input video $V_{1:t_i}$, this module takes as input the high-resolution frame V_{t_i} extracted from the

last frame and a low-resolution video $\widetilde{V_{(t_\iota-l):t_\iota}}$, obtained by spatially downsampling the input video of length l. These are processed separately using image backbone networks to generate two sets of features: $\mathcal{B}_{2\mathcal{D}}(V_{t_i})$ and $\mathcal{B}_{3\mathcal{D}}(\widetilde{V_{(t_\iota-l):t_\iota}})$.

The 2D features $\mathcal{B}_{2\mathcal{D}}(V_{t_i})$ are passed through a standard feature pyramid layer to obtain $\widetilde{\mathcal{B}_{2D}}(V_{t_i})$. $\widetilde{\mathcal{B}_{2D}}(V_{t_i})$ is used by a Region Proposal Network (RPN) [20] to predict region proposals. From these region proposals, we extract 2D object features (processed through an average pooling layer) and object position features using an RoIAlign layer and linear layers. These two features are summed to form the final object features $F_{t_i}^{obj}$.

Also, we mix the 2D features $\mathcal{B}_{2\mathcal{D}}(V_{t_i})$ and 3D features $\mathcal{B}_{3\mathcal{D}}(\widetilde{V_{(t_\iota-l):t_\iota}})$ to create a combined feature. We upsample 3D feature map to match 2D spatial resolution, average along its temporal dimension, and apply a 2D 3×3 convolution to match the dimensions of $\mathcal{B}_{2\mathcal{D}}(V_{t_i})$, resulting in $\mathcal{B}_{3\mathcal{D}}^{2\mathcal{D}}(\widetilde{V_{(t_\iota-l):t_\iota}})$. These features are added to $\mathcal{B}_{2\mathcal{D}}(V_{t_i})$, and each layer undergoes a 1×1 convolution, producing combined feature $F_{t_i}^{comb}$.

Finally, $F_{t_i}^{3D}$ (i.e., $\mathcal{B}_{3\mathcal{D}}(\widetilde{V_{(t_\iota-l):t_\iota}})$), $F_{t_i}^{comb}$, and $F_{t_i}^{obj}$ serve as outputs of the feature extraction module for subsequent networks. In our experiments, we use ResNet-50 as the 2D CNN and X3D-M as the 3D CNN.

3.4 Long- and Short-term Heterogeneous Attention

We introduce the heterogeneous attention mechanisms employed by our model to obtain information from both long- and short-term memory, combining multi-head self-attention (MHSA) [18], multi-head cross-attention (MHCA) [19], and multi-scale multi-head deformable attention (MSMHDA) [20].

For Long-term Memory Aggregation, we adopt an iterative approach. We adopt an iterative approach. (Although more distant memories may become diluted, aligning with the principles of memory in the real world), we define $O_{t_{i-1}}$ as the long-term memory at time t_i. For any given moment t_i in the video sequence $\{V_{1:t_i}\}_{i=1}^{N_v}$, the model learns information from t_1 to t_{i-1} through $O_{t_{i-1}}$. To achieve this, we concatenate $O_{t_{i-1}}$ and Q_{t_i}, then apply MHSA [18]. The part of the concatenated vector corresponding to the original Q_{t_i} is selected as the new Q_{t_i}.

For short-term intention mining, as illustrated in **Fig. 3** , we obtain information from 3 different perspectives to enhance intention prediction.

Firstly, $F_{t_i}^{obj}$ captures object-related information. We apply MHCA [18] to learn the features and spatial distribution of potential NAOs, as follows:

$$I_{t_i}^{obj} = \mathrm{MHCA}(Q_{t_i}, F_{t_i}^{obj}) \tag{1}$$

Secondly, $F_{t_i}^{3D}$ contains multi-scale and multi-dimensional short-term memory features. To process these, we designed MSMHDA_{3D} [20]. 3D coordinates of Q_{t_i} vectors are computed through a learnable projection. The $F_{t_i}^{3D}$ layers are dimensionally adjusted using 1×1 convolutions. To distinguish the feature layers, we add learnable scale embeddings $\{e_l^{3D}\}_{l=1}^{L}$ and calculate the final 3D feature $I_{t_i}^{3D}$:

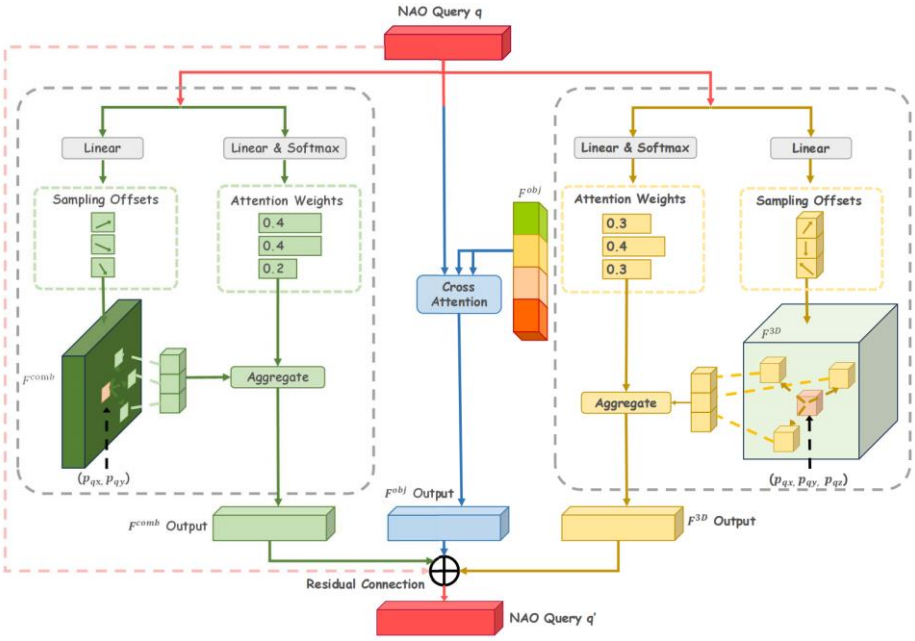

Fig. 3. The architecture of Short-term Intention Mining. From left to right, we utilize 2D deformable attention to learn F^{comb}, cross-attention to learn F^{obj}, and 3D deformable attention to learn F^{3D}, respectively.

$$I_{t_i}^{3D} = \text{MSMHDA}_{3D}\left(Q_{t_i}, \widehat{p_{Q_{t_i}}}, \text{Conv1x1}\left(F_{t_i}^{3D}\right) + \{e_l^{3D}\}_{l=1}^{L}\right) \tag{2}$$

Lastly, $F_{t_i}^{comb}$ merges high-resolution and low-resolution video features to form a multi-scale 2D feature map. We apply MSMHDA [20] to enhance efficiency by reducing redundancy. After computing 2D coordinates, adjusting dimensions, and adding scale embeddings $\{e_l^{comb}\}_{l=1}^{L}$, Q_{t_i}, $\widehat{p_{Q_{t_i}}}$, and $F_{t_i}^{comb}$ are processed as follows:

$$I_{t_i}^{comb} = \text{MSMHDA}\left(Q_{t_i}, \widehat{p_{Q_{t_i}}}, \text{Conv1x1}\left(F_{t_i}^{comb}\right) + \{e_l^{comb}\}_{l=1}^{L}\right) \tag{3}$$

Finally, the aggregated features are summed and residual connections are added:

$$Q_{t_i} = \left(I_{t_i}^{obj} + I_{t_i}^{3D} + I_{t_i}^{comb}\right) + Q_{t_i} \tag{4}$$

3.5 Foresight Knowledge Distillation

To effectively predict future interactions, we initialize a set of learnable Intention Queries $Q_{t_i}^{I} \in R^{N_I \times C}$. These queries are processed through multiple stages of intention prediction, forming the basis of our teacher model. Then the model is used to supervise and distill knowledge to the model to enhance its ability to predict.

The intention prediction process involves both Long-term memory $O_{t_{i-1}}$, Short-term memory $F_{t_i}^{comb}$, and the global features extracted from future frames. To handle Short-term memory, we apply average pooling to each layer of $F_{t_i}^{comb}$. Each vector is then processed through a 1×1 convolution to standardize the dimensions to C:

$$F_{t_i}^{comb.avg} = \text{Flatten}\left(\text{AvgPool}\left(F_{t_i}^{comb}\right)\right) \tag{5}$$

$$F_{t_i}^{comb.m} = \text{Concat}\left(\text{Conv1x1}\left(F_{t_i}^{comb.avg}\right)\right) \tag{6}$$

where k denotes the unified height of feature maps after AvgPool, resulting in l distinct one-dimensional vectors of varying dimensions.

Then Intention Queries $Q_{t_i}^{I}$ are passed through a Multi-Head Self-Attention (MHSA) [18] layer to learn internal representations. Then, we concatenate $O_{t_{i-1}}$ and $F_{t_i}^{comb.m}$ to process them using Multi-Head Cross-Attention (MHCA) [19]. Scale-level embeddings $\{e_l^{comb.m}\}_{l=1}^{L}$ are added to enhance $F_{t_i}^{comb.m}$ across different feature levels.

During training, we supervise I_{t_i} using global features from future video frames. Specifically, we extract the global features of the next n frames using CLIP [9], compute their average and compare it with the predicted intention I_{t_i}. The dimensions of $I_{t_i}^{avg}$ are aligned with the CLIP features using a 1×1 convolution. After L_I iterations, the final intention prediction $I_{t_i} \in R^{N_I \times C}$ is obtained. This predicted intention is used to guide the NAO Queries Q_{t_i} via MHCA [19]. Finally, we can compute the explicit intention loss: $L_{intention} = \left|\left|I_{t_i}^{gt} - I_{t_i}^{avg}\right|\right|_2$.

3.6 Prediction Head

After L_Q iterations, we obtain the NAO prediction set features O_{t_i}. Each prediction in O_{t_i} passes through two separate 3-layer perceptions to get bounding box (bbox) and time to contact (ttc), and through two linear layers to obtain noun and verb categories. Following [6], background classes are added for nouns and verbs, indicating "no object" and "no action." During training, bipartite graph matching is used to find the optimal pairing based on ground truth and calculate the total loss.

Specifically, in matching, for any sample i, we calculate the noun cross-entropy loss, verb cross-entropy loss, and the L1 and IoU losses for the bounding box. We determine the optimal matching loss L_{match} by selecting the i that minimizes the weighted sum of costs, referred to as i^*.Unmatched samples are set to the background classes, giving the unmatched loss:

$$L_{\text{matched}} = min\left(\alpha L_{\text{noun}}(i) + \beta L_{\text{verb}}(i) + \gamma L_{\text{bbox-L1}}(i) + \delta L_{\text{bbox-IoU}}(i)\right) \tag{7}$$

$$L_{\text{unmatched}} = \sum_{j \neq i^*}\left(\rho L_{\text{noun}}(j) + \sigma L_{\text{verb}}(j)\right) \tag{8}$$

Combined with $L_{intention}$ and the L1 loss for time to contact, the total loss of our model is expressed as:

$$L_{total} = L_{matched} + L_{unmatched} + \epsilon\left|ttc_{i^*} - ttc_{\text{gt}}\right| + \phi L_{intention} \quad (9)$$

During inference, we take the result with the highest sum of noun and verb confidence scores as the final result.

4 Experiments

4.1 Dataset

Our model is trained using the Ego4D-STA dataset, which is widely adopted in the STA field. Ego4D-STA v1 [1] contains 120 hours of videos, 27,801 training samples, 17,217 validation samples, and 19,780 test samples, classified into 87 noun and 74 verb categories. Ego4D-STA v2 [1] expands to 243 hours, 98,276 training samples, 47,395 validation samples, and the same 19,780 test samples, using a richer taxonomy with 128 noun and 81 verb categories.

4.2 Evaluation Metrics

Evaluation is performed using Top-K Average Precision (AP) and mean Average Precision (mAP), with no penalties for predicting unannotated objects. We specifically evaluate various Top-5 AP/mAP metrics: AP/mAP for nouns (b+n), nouns and verbs (b+n+v), nouns and time to contact (b+n+t), and the full combination (b+n+v+t). The overall metric (b+n+v+t) measures the model's ability to predict interactions of future active objects by considering nouns, verbs, and time to contact as in [1].

4.3 Experiments Settings and Computational Performance

For training, we use a batch size of 1 and 50 epochs, setting the optimizer as Adam. The base learning rate is 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 10^{-4}. The learning rate is decayed at the 40-th epoch by 0.1. Compared to our baseline, Stillfast, we compute the number of parameters and inference time for comparison, the results are (M/min/mAP), Stillfast 101/29.05/2.48, our Intenformer 142/38.67/4.01. We perform evaluations on an 8-card NVIDIA GeForce RTX 4090 setup.

4.4 Comparison with State-of-the-arts (SOTA)

Table 1. (AP) Results of our model and other baseline methods on Ego4D-STAv1.

model (AP)	b	b+n	b+n+t	b+n+v	b+t	b+v	b+v+t	b+n+v+t
Slowfast [1]	40.5	24.5	5.00	4.90	8.40	8.16	1.90	1.50
AVT [15]	40.5	24.5	4.39	4.52	7.12	8.45	1.15	1.71
ANACTO[21]	40.5	24.5	4.55	5.10	7.47	8.90	1.54	1.91
MeMVIT [16]	40.5	24.5	4.95	5.89	9.27	10.0	2.11	1.34
NAOGAT[22]	**45.3**	**27.0**	9.00	6.54	16.6	12.2	4.18	2.47
IntenFormer	42.5	25.8	**9.12**	**7.60**	**16.9**	**13.1**	**4.63**	**3.01**

As shown in **Table 1**, our model exhibits strong performance on the STAv1 dataset, particularly in predicting verbs and time to contact (ttc). Compared to previous models, we observe improvements of 0.29 in AP_{b+t}, 0.9 in AP_{b+v}, and 0.45 in AP_{b+v+t}. These gains are due to the model's ability to capture macro trends and dynamically predict object positioning and timing. However, for AP_b and AP_{b+n}, our model shows slightly lower performance compared to NAOGAT [22]. Their work focuses on Next-Active Object (NAO) prediction, emphasizing bounding boxes and nouns, which has a disadvantage of overfitting in comprehensive prediction.

Table 2. (mAP) Results of our model and other baseline methods on Ego4D-STAv1&2.

	v1	v2	v1	v2	v1	v2	v1	v2	v1	v2
	b+n		b+v		b+n+t		b+n+v		b+n+v+t	
1-stage StillFast [6]	16.2	20.3	-	-	4.94	7.16	7.32	10.37	2.48	3.96
2-stage Slowfast [1]	17.5	21.4	-	3.31	5.37	6.92	5.19	7.74	2.07	2.94
IntenFormer	**19.2**	**22.5**	**8.41**	**10.21**	**7.13**	**9.35**	**7.47**	**10.06**	**4.01**	**5.18**

In terms of mAP performance, as seen in **Table 2**, our model sets a new state-of-the-art (SOTA) benchmark on both STAv1 and STAv2 datasets, surpassing previous models. The significant improvements on STAv2 are mainly attributed to the larger dataset size and more detailed annotations, demonstrating our model's strong generalization capability across different datasets and evaluation conditions.

4.5 Ablation Study Based on the STAv1 Dataset

Feature Mining

Table 3. Ablation study considering different feature combinations in Feature Mining.

Long	Short	Future	b+n	b+v	b+t	b+v+t	b+n+v+t
	✓		25.07	11.45	15.38	3.20	2.13
		✓	24.50	10.58	14.54	3.53	2.05
✓		✓	24.82	12.74	15.01	4.24	2.75
	✓	✓	25.28	12.58	16.74	4.48	2.78
✓	✓		25.35	12.11	16.43	4.15	2.60
✓	✓	✓	**25.83**	**13.10**	**16.89**	**4.63**	**3.01**

Table 3 shows the impact of different feature combinations for mining intentions. Given the initialization of F^{comb} using pre-trained Deformable DETR parameters, results vary significantly with its introduction. Thus, we only consider combinations including F^{comb}. Both F^{obj} and F^{3D} improve overall performance; F^{obj} increases AP_{b+n}, and F^{3D} enhances AP_{b+v}, AP_{b+t}, and AP_{b+v+t}. Combining all features achieves the best results.

Information Interaction

Table 4. Ablation study considering the different combinations of Information Interaction.

F^{obj}	F^{3D}	F^{comb}	b+n	b+v	b+t	b+v+t	b+n+v+t
		✓	25.17	12.40	15.97	4.22	2.23
✓		✓	25.70	12.76	16.03	4.29	2.75
	✓	✓	25.21	12.89	16.58	4.51	2.89
✓	✓	✓	**25.83**	**13.10**	**16.89**	**4.63**	**3.01**

Table 4 shows the impact of different information interaction combinations. Results with short-term memory intention are initially better, but global knowledge distillation improves with long-term memory aggregation. Global knowledge surpasses short-term memory intention when long-term memory aggregation is present, and combining all three yields the best performance.

5 Conclusion

This paper presents **IntenFormer**, a model combining long- and short-term memory to mine intentions for STA. With Heterogeneous Attention and Distillation mechanism, this dual approach enhances behavior prediction performance. Tests on the Ego4D STA dataset confirm its superiority.

References

1. Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, MiaoLiu, Xingyu Liu, et al., "Ego4d: Around the world in 3,000 hours of egocentric video," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18995–19012.
2. Lihuan Li, Maurice Pagnucco, and Yang Song, "Graph-based spatial transformer with memory replay for multi-future pedestrian trajectory prediction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2231–2241.
3. Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al., "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," International Journal of Computer Vision, pp. 1–23, 2022.
4. Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles, "Procedure planning in instructional videos," in European Conference on Computer Vision. Springer, 2020, pp. 334–350.
5. Esteve Valls Mascaro, Hyemin Ahn, and Dongheui Lee, "Intention- conditioned long-term human egocentric action forecasting," arXiv preprint arXiv:2207.12080, 2022.
6. Francesco Ragusa, Giovanni Maria Farinella, and Antonino Furnari, "Stillfast: An end-to-end approach for short-term object interaction anticipation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3635–3644.
7. Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and J¨urgen 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.
8. Qi Zhao, Ce Zhang, Shijie Wang, Changcheng Fu, Nakul Agarwal, Kwonjoon Lee, and Chen Sun, "Antgpt: Can large language mod- els help long-term action anticipation from videos?," arXiv preprint arXiv:2307.16368, 2023.
9. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning. PMLR, 2021, pp. 8748–8763.
10. Debaditya Roy and Basura Fernando, "Predicting the next action by modeling the abstract goal," arXiv preprint arXiv:2209.05044, 2022.
11. Debaditya Roy and Basura Fernando, "Action anticipation using latent goal learning," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2745–2753.
12. Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Gio- vanni Maria Farinella, "Next-active-object prediction from egocentric videos," Journal of Visual Communication and Image Representation, vol. 49, pp. 401–411, 2017.
13. Jingjing Jiang, Zhixiong Nan, Hui Chen, Shitao Chen, and Nanning Zheng, "Predicting short-term next-active-object through visual attention and hand position," Neurocomputing, vol. 433, pp. 212–222, 2021.
14. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weis- senborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
15. Rohit Girdhar and Kristen Grauman, "Anticipative video transformer," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 13505–13515.
16. Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph 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. 13587–13597.
17. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin, "Convolutional sequence to sequence learning," in Interna- tional conference on machine learning. PMLR, 2017, pp. 1243–1252.
18. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
19. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r- cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.
20. Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable detr: Deformable transformers for end-to-end object detection," arXiv preprint arXiv:2010.04159, 2020.
21. Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue, "Anticipating next active objects for egocentric videos," arXiv preprint arXiv:2302.06358, 2023.
22. Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue, "Leveraging next-active objects for context-aware anticipation in egocentric videos," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 8657– 8666.