

Vision and Intention Boost Large Language Model in Long-Term Action Anticipation

Congqi Cao¹, Lanshu Hu¹, Yating Yu¹,
Yanning Zhang¹

¹Northwestern Polytechnical University
{congqi.cao, ynzhang}@nwpu.edu.cn,
{hlshu, yatingyu}@mail.nwpu.edu.cn

Abstract

Long-term action anticipation (LTA) aims to predict future actions over an extended period. Previous approaches primarily focus on learning exclusively from video data but lack prior knowledge. Recent researches leverage large language models (LLMs) by utilizing text-based inputs which suffer severe information loss. To tackle these limitations single-modality methods face, we propose a novel Intention-Conditioned Vision-Language (ICVL) model in this study that fully leverages the rich semantic information of visual data and the powerful reasoning capabilities of LLMs. Considering intention as a high-level concept guiding the evolution of actions, we first propose to employ a vision-language model (VLM) to infer behavioral intentions as comprehensive textual features directly from video inputs. The inferred intentions are then fused with visual features through a multi-modality fusion strategy, resulting in intention-enhanced visual representations. These enhanced visual representations, along with textual prompts, are fed into LLM for future action anticipation. Furthermore, we propose an effective example selection strategy jointly considers visual and textual similarities, providing more relevant and informative examples for in-context learning. Extensive experiments with state-of-the-art performance on Ego4D, EPIC-Kitchens-55, and EGTEA GAZE+ datasets fully demonstrate the effectiveness and superiority of the proposed method.

Introduction

Predicting future actions is a crucial task in fields such as robotic collaboration (Koppula and Saxena 2015). This predictive capability enables systems to provide assistance (Rodin et al. 2021; Yao et al. 2019) at the appropriate moments, thereby enhancing both the naturalness and effectiveness of the interaction. For instance, in autonomous driving (Zhang et al. 2020), accurately anticipating the intentions behind the movements of other vehicles enables the autonomous system to make proactive preparations, thereby reducing potential hazards. Unlike other video understanding tasks, action anticipation requires not only understanding the observed context but also predicting future actions based on the observation. This task is inherently challenging, as it requires both strong logical reasoning capabilities and the ability to manage the uncertainty of future actions.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

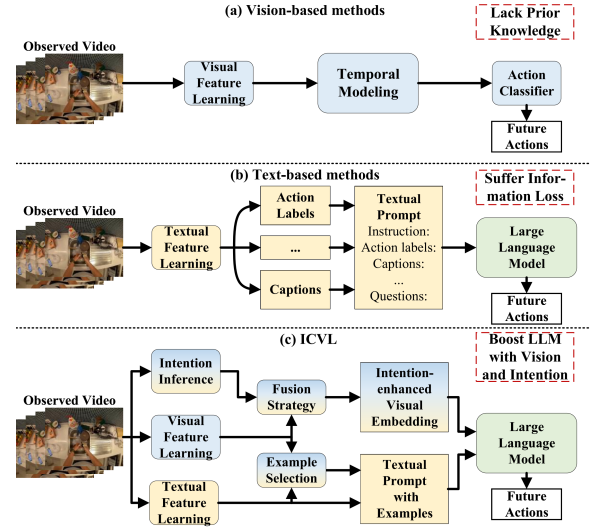


Figure 1: Illustration of different action anticipation methods. (a) Vision-based methods. (b) Text-based methods. (c) Our proposed Intention-Conditioned Vision-Language (ICVL) model.

To address the task of action anticipation, some approaches start by leveraging video data to learn visual features and model the temporal relationships between the features via neural networks, as shown in Figure 1 (a). LSTM/RNN-based (Furnari and Farinella 2020; Sener, Singhania, and Yao 2020; Sadegh Aliakbarian et al. 2017) and Transformer-based (Gong et al. 2022; Yuan et al. 2025; Wang et al. 2023) models are employed to model the dependencies between actions as well as the object-action inter-action relationships (Pasca et al. 2024). Furthermore, (Cao et al. 2024) proposes a hybrid Transformer-GRU architecture to make predictions. However, visual data is often redundant and low in information density. Methods relying solely on visual data lack prior knowledge, making it challenging to model the intrinsic evolution of actions and rendering them overly sensitive to visual variations.

After achieving significant success in natural language processing, large language models (LLMs) (Wang et al. 2024; Cui et al. 2024) have been adapted to the vision do-

main. This success has motivated researchers to leverage the strong prior knowledge and reasoning capabilities of LLMs to address the challenge of action anticipation. An intuitive solution is to generate appropriate textual substitutes of the original video content, enabling LLMs to predict future actions through a question-answering paradigm, as illustrated in Figure 1 (b). The simplest form of such substitutes is the observed action labels (Zhao et al. 2023), generated by off-the-shelf action recognition models. Nevertheless, due to the limited accuracy of existing recognition models, these action labels often contain substantial noise and errors. Another approach involves using a Vision-Language Model (VLM) to generate more detailed textual captions (Kim et al. 2024). However, fully understanding video content and providing accurate descriptions is inherently challenging. Methods relying solely on textual inputs suffer from significant information loss, limiting the ability of LLMs to make precise and contextually informed predictions.

To fully preserve the visual content and extract crucial clues for long-term action anticipation (LTA), we propose a novel Intention-Conditioned Vision-Language (ICVL) model that integrates complementary visual and textual information with the commonsense prior knowledge of LLMs. This approach addresses the limitations of single-modality methods by combining rich visual features with high-level behavioral intentions to boost the performance of LLMs in LTA. On the one hand, visual data, such as the presence of objects like “a bowl”, offers valuable insights for future action prediction, even when these objects are not explicitly mentioned in textual substitutes. On the other hand, behavioral intentions, such as “cleaning the kitchen”, represent high-level semantic concepts that guide the evolution of actions over time. By capturing these intentions, we can better understand the progression of actions and gain critical insights for predicting future events.

Specifically, our ICVL model employs a Vision-Language Model (VLM) to infer behavioral intentions directly from video data by analyzing the entire temporal dynamics of the observed video. This allows the model to generate textual features that capture the high-level intentions behind the actions. We then introduce a novel fusion mechanism, Intention-Context Attention Fusion (ICAF), which integrates visual features with the inferred behavioral intentions to produce intention-enhanced visual embeddings. These embeddings are more discriminative, with reduced redundancy and higher information density, as they focus on the most relevant aspects of the visual data guided by behavioral intentions. Combined with carefully designed textual prompts, these enriched embeddings are fed into the LLM, which has been fine-tuned in a parameter-efficient manner to adapt to the specific task of action anticipation.

Additionally, to further improve the reasoning capabilities of LLMs, we propose an effective example selection mechanism that leverages both visual and textual modalities to identify the most relevant examples for in-context learning. This ensures that the LLM is provided with the most pertinent data, enhancing its ability to make informed and accurate predictions. Extensive experiments across three datasets demonstrate the effectiveness of our approach, validating the

strength of combining vision, intention, and LLMs for long-term action anticipation.

Our key contributions can be summarized as follows:

- We propose a novel multimodal framework for long-term action anticipation that fully leverages both visual and textual information, integrating them with the prior knowledge and reasoning capabilities of LLMs.
- We introduce intention-enhanced visual features by fusing visual data with inferred behavioral intentions, addressing information loss and enriching the representations for more precise and reliable action predictions.
- We design an effective example selection mechanism that integrates both visual and textual modalities to identify the most relevant examples, improving long-term action anticipation via enhanced in-context learning.
- Extensive experiments demonstrate the effectiveness and superiority of our proposed method, achieving state-of-the-art performance on Ego4D, EPIC-Kitchens-55 and EGTEA GAZE+ datasets.

Related Works

Action Anticipation

Action anticipation aims at inferring future actions based on a period of observed video, and can be categorized into long-term and short-term anticipation tasks depending on the time to predict. Our work focus on the long-term action anticipation. Previous methods mainly make predictions by modeling the temporal dynamics solely from the visual features. (Furnari and Farinella 2020) uses a rolling LSTM to encode the input and an unrolling LSTM to make recurrent predictions for future actions. With the rise of Transformers, (Gong et al. 2022) adopts an end-to-end attention model, leveraging fine-grained visual features from previous frames for prediction. Furthermore, (Cao et al. 2024) proposes a hybrid architecture that utilizes a Transformer as the encoder for long-term sequence modeling, coupled with a GRU decoder for flexible recurrent predictions. Recently, approaches utilizing LLMs have become increasingly popular. (Zhao et al. 2023) firstly utilizes LLMs to solve the LTA task by simply substituting video content with observed action labels. (Kim et al. 2024) employs an image captioning model to generate descriptions of the video, thereby enriching the textual information. Additionally, (Pei et al. 2024) leverages a more advanced foundation model to extract richer visual features, generating more accurate observed action labels as input for the LLM. However, these methods depend excessively on a single modality. On the one hand, vision-based methods face information redundancy and lack prior knowledge, making it challenging to model long-term temporal relationships and make accurate predictions. On the other hand, text-based methods suffer from severe information loss and noise, struggling to generate accurate action recognition results or detailed video descriptions, thereby impairing predictive accuracy. In contrast to the aforementioned methods, we leverage the contextual information from visual data, the intentional information from textual descriptions, and the commonsense reasoning capabilities of LLMs to enhance long-term action anticipation.

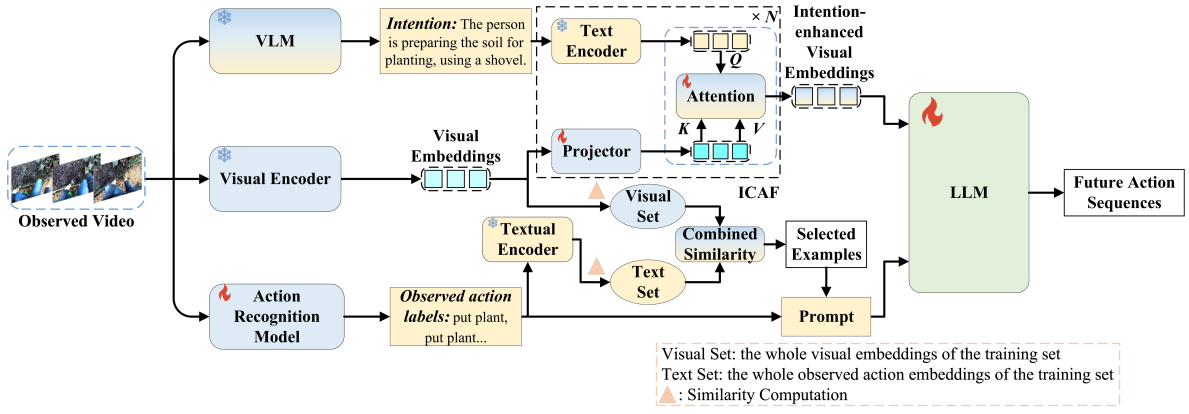


Figure 2: Illustration of Intention-Conditioned Vision-Language (ICVL) model. Given a video, we use a VLM, a visual encoder, and an action recognition model to extract behavioral intention, original visual embeddings, and observed action labels respectively. The behavioral intention and visual embeddings are then integrated into the intention-enhanced visual embeddings through our proposed Intention-Context Attention Fusion (ICAF) module, in which visual features serve as the keys (K) and values (V), while textual intention features act as the queries (Q). Then we consider both visual similarity and textual similarity based on observed action labels to select examples from the training set for in-context learning. Finally, the textual prompt—composed of instructions, observed action labels, and selected examples—along with the intention-enhanced visual embeddings, are fed into the LLM to generate predictions for future action sequences.

Large Language Model

LLMs based on the Transformer architecture, typically make predictions in an autoregressive manner. These models, often containing billions of parameters, are trained on vast amounts of data and have demonstrated remarkable performance in natural language processing. Notable examples include GPT-4 (Achiam et al. 2023) and LLaMA (Touvron et al. 2023). To adapt these models more effectively to downstream tasks, some approaches (Hu et al. 2021) propose to fine-tune part of the model parameters on specific datasets, while others (Brown et al. 2020) attempt to leverage the LLMs’ in-context learning ability by providing high quality examples. Both strategies can further enhance the performance of LLMs, yielding higher-quality responses. Moreover, LLMs exhibit a profound understanding of the textual structure and semantics, as well as the ability to comprehend rich information from other modalities after alignment, such as visual and audio data. As a result, LLMs have been successfully applied to the visual domain (Tang et al. 2025), demonstrating significant performance. For example, (Li, Wang, and Jia 2024) uses context tokens based on multimodal fusion to represent an entire image and content tokens to encapsulate visual cues for video or image question-answering tasks. Inspired by this approach and the unique nature of LTA task to predict future actions, we creatively leverage high-level behavioral intentions to bridge past and future actions. By combining intentions with visual features, we generate intention-enhanced visual embeddings, which improve prediction by making visual features more discriminative and providing cues related to action evolution.

The Proposed Method: ICVL

We introduce our proposed Intention-Conditioned Vision-Language (ICVL) model in this section, which combines

LLM with intention-enhanced visual embeddings and carefully designed textual prompts to predict future actions, as shown in Figure 2.

Action Recognition and Intention Inference

Actions labels. Long-term action anticipation requires predicting future actions over an extended period, where upcoming actions are inferred from an observed video. The observed video can be divided into several segments $\{S^i\}_{i=1}^{N_{seg}}$, with each segment S^i corresponding to an action label A^i . Consequently, an observed video can be represented as $\{A^1, A^2, \dots, A^{N_{seg}}\}$. These action labels are represented as verb-noun pairs, where each action is composed of a verb and a noun $\{v^i, n^i\}$, such as *put plant*. To make a fair comparison, we follow (Zhao et al. 2023) and use the CLIP visual encoder to extract video features and get N_{seg} visual embeddings represented as $\{E^1, E^2, \dots, E^{N_{seg}}\}$. Then we use a Transformer-based architecture as the action recognition model, which consists of a Transformer encoder to model the visual embeddings and two MLP heads to decode the verb and noun. For each video segment S^i , we can obtain the action label based on the identified verb-noun pair. The action recognition model is trained using the cross-entropy loss between predictions and ground-truth action labels.

Intention Inference. Human actions are inherently driven by high-level intentions, which guide the evolution of actions over time. Therefore, understanding an individual’s intention is crucial for accurately predicting the future actions. While (Zhao et al. 2023) uses an LLM to infer goals (i.e., intentions) from observed action labels, these labels often contain substantial noise and errors, making it difficult for the LLM to infer correct intentions. Instead, we leverage observed visual cues through a VLM to obtain more

accurate intentions. Specifically, we first uniformly sample N_{frm} frames $\{f_1, f_2, \dots, f_{N_{frm}}\}$ from an observed video. These frames can be regarded as a condensed representation of the video’s content, sufficiently indicating the developmental trends of future actions. We then employ a pre-trained VLM \mathcal{E} to sequentially infer behavioral intentions $\{I_1, I_2, \dots, I_{N_{frm}}\}$ from each frame in chronological order, using the prompt P^I “What does the person want to do?”. The intentions inferred from all the preceding frames are also used as input to provide contextual information, supporting the VLM’s interpretation of the current image’s intention. This can be formulated as:

$$I_t = \mathcal{E}(P^I, f_t, \sum_{i=1}^{t-1} I_i), \quad (1)$$

where t is the index of the current image. The final behavioral intention is derived from the text generated by the VLM based on the last frame and its corresponding context.

Intention-Context Attention Fusion

Multi-modality fusion has been proven effective in short-term action anticipation tasks (Furnari and Farinella 2020; Cao et al. 2024). However, in the field of long-term action anticipation, this approach remains underexplored, particularly for LLM-based methods. In this section, we introduce our proposed Intention-Context Attention Fusion strategy.

Visual and Intention embeddings. For each video segment S^i , we use a pretrained vision encoder to extract the original visual embeddings through k uniformly sampled video frames, resulting in $E_v^i \in \mathbf{R}^{k \times d_v}$. The N_{seg} video segments’ visual embeddings can be concatenated to represent the whole video’s embeddings $E_v' \in \mathbf{R}^{T \times d_v}$ where $T = N \times k$. These visual embeddings serve as visual prompts, which are integrated with textual intention prompts as input for the LLM. To enhance the model’s understanding of sequential information, we add 2D fixed positional encoding (Vaswani 2017) to the visual embeddings. Then we can get the modified visual embeddings E_v'' after adding the positional embeddings. Furthermore, to align the dimension of visual embeddings with the embedding space d_l of the LLM, a linear project layer is used to get the final visual embeddings $E_v \in \mathbf{R}^{T \times d_l}$.

For the intention embeddings, a pretrained text encoder can be directly used to encode the behavioral intentions, denoted as $E_i \in \mathbf{R}^{seq \times d_l}$, where seq and d_l represent the sequence length of the intention embeddings and the embedding dimension of the LLM, respectively.

Fusion strategy. Our fusion strategy, based on cross-attention, integrates both visual and intention embeddings to obtain enhanced intention-enhanced visual embeddings $E_{ic} \in \mathbf{R}^{seq \times d_l}$. This process can be formulated as:

$$E_{ic} = \text{Attention}(E_i, E_v, E_v), \quad (2)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_l}}\right)V, \quad (3)$$

where visual embeddings serve as keys (K) and values (V), intention embeddings act as queries (Q). And scaling factor $\sqrt{d_l}$ is introduced to avoid gradient vanishing. As intention is embedded in visual features and guides the evolution of actions, it can enhance the visual embeddings to be more discriminative. Besides, this fusion also enhances the LLM’s ability to comprehend visual information and improves its interpretability, resulting in more discriminative and intention-consistent cross-modal representations.

Example Selection

Existing research (Brown et al. 2020) has shown that augmenting LLMs with relative demonstration examples can significantly enhance their generative capabilities. However, selecting appropriate examples for action anticipation tasks remains challenging due to the diversity of scenarios and the variability of actions even within the same scenario. To address this, we propose an example selection mechanism that jointly considers both visual and textual modalities as shown in Figure 2. And this mechanism can provide more relevant and appropriate examples for in-context learning, thereby improving generalizability. We first introduce single-modality selection and then explain how to extend it to a comprehensive multi-modality approach.

Single-Modality Selection. Taking the visual modality selection as an example, after obtaining the whole original visual embeddings E_v' , we apply average pooling to derive the averaged visual embeddings \bar{E}_v as a global representation of visual features. We then utilize L2 distance to obtain the similarity scores s_v between the query video and all the training videos, due to its clear geometric interpretation and effectiveness in capturing variations in both vector magnitude and feature dimensions, where a smaller similarity score s_{vi} between two features indicates greater similarity. Finally, we select the top- k examples based on the similarity scores. This process can be formulated as below:

$$U = \arg \min_{U \subset \Omega, |U|=k} \sum_{\bar{E}_v^i \in \Omega} \|E_q - \bar{E}_v^i\|_2, \quad (4)$$

where Ω represents the complete set of \bar{E}_v in the training set, E_q represents the embeddings of the query and U represents the set of the top- k most similar selected embeddings. Based on U , corresponding examples are then extracted from the training set, comprising observed action labels and future action sequences.

The example selection mechanism for the textual modality adheres to the same principles as that of the visual modality, where observed action labels are encoded to obtain textual embeddings.

Multi-Modality Selection. After obtaining the similarity results of the visual and textual modalities, we adopt a weighted summation approach to comprehensively consider the similarities of both modalities. First, the similarity scores for the visual and textual modalities are normalized as shown in the following formula:

$$s_{ti}^n = \frac{s_{ti} - \min(s_t)}{\max(s_t) - \min(s_t)}, \quad (5)$$

<u>Prompting using “In-Context Learning”</u>	
Instruction:	Given a video and 8 observed actions, you are supposed to predict the next most possible 20 actions in the format of verb noun pair in sequence that are consistent with logic and common sense.
Examples:	remove leaf, put plant, take grass, put grass, take trowel, put plant, dig plant, take plant -> dig plant, remove plant, put plant, put plant, pull plant, take plant, put plant, take garbage, put garbage, dig plant, put plant, put plant, take plant, put plant, take garbage, take plant, put garbage, put trowel, take mask, move rope ...more examples...
Observed actions:	put plant, put plant, put plant, put plant, put plant, put plant, put plant, cut plant
Intention-enhanced visual embeddings:	...
Predictions:	

Figure 3: Illustration of prompt for LLMs using in-context learning. The prompt is composed of an instruction, selected examples based on multi-modality similarity, observed actions and intention-enhanced visual embeddings.

$$s_{vi}^n = \frac{s_{vi} - \min(s_v)}{\max(s_v) - \min(s_v)}, \quad (6)$$

where s_{ti}^n represents the normalized similarity score for the textual modality of the i -th representation, and s_{vi}^n represents the normalized similarity score for the visual modality of the i -th representation.

The comprehensive similarity score is then calculated using a weighted summation based on s_{ti}^n and s_{vi}^n :

$$s_i = \alpha \times s_{ti}^n + (1 - \alpha) \times s_{vi}^n, \quad (7)$$

where α is a weighting factor that reflects the balance between the two modalities. When α is set to either 1 or 0, the selection mechanism becomes solely dependent on a single modality. Based on the comprehensive similarity scores, the top-k examples are selected. The final prompt is illustrated in the Figure 3.

Training

As shown in Figure 2, the visual and textual encoders in ICVL are frozen, while the ICAF module are fully trainable. Given the significant computational cost of fully training LLMs, we adopt the LoRA (Low-Rank Adaptation) (Hu et al. 2021) for fine-tuning the LLM. All trainable parameters are optimized based on the text generated by the LLM. As the model is tasked with predicting a future action sequence, We employ the next-token prediction loss with negative log-likelihood to optimize the predicted tokens:

$$\mathcal{L}_{CE}(\theta) = - \sum_{t=1}^M \log p_{\theta}(y_t | y_{<t}), \quad (8)$$

where θ represents the parameters of the model, y_t denotes the target token to be predicted at position t , M refers to the total number of tokens to be predicted, $y_{<t}$ represents the tokens predicted prior to position t , and p_{θ} indicates the probability of successfully predicting the token at position t based on $y_{<t}$. This loss measures the difference between

the action sequence output by LLM and the corresponding ground truth action sequence.

We adopt an end-to-end training process that both the ICAF module and the LoRA Adapter module are fine-tuned simultaneously.

Experiment

In this section, we first introduce the datasets and evaluation metrics, followed by providing implementation details. Subsequently, we compare ICVL with state-of-the-art methods under various popular benchmarks, and finally present ablation studies of the proposed strategies.

Datasets and Evaluation Metrics

Ego4D (Grauman et al. 2022). This dataset is a large-scale egocentric dataset encompassing hundreds of scenarios, such as home, outdoor, and workplace environments. We conduct experiments on its *Forecasting* subset, which includes a total of 243 hours of video, 3472 annotated clips. It has 117 verbs and 521 nouns for the LTA task. We adhere to the dataset’s standard splits for evaluation.

EPIC-KITCHENS-55 (EK-55) (Damen et al. 2020). This dataset contains 55 hours of egocentric videos centered around cooking scenarios, recorded by 32 participants in 32 different kitchens. It contains 125 verb categories and 352 noun categories. We follow the splits provided by (Nagarajan et al. 2020).

EGTEA Gaze+ (EGTEA) (Li, Liu, and Rehg 2018). This dataset is a first-person dataset containing 86 densely labeled cooking videos over 26 hours, with 19 verb categories and 51 noun categories. We also follow the splits provided by (Nagarajan et al. 2020).

Evaluation Metrics. For Ego4D, we employ the default edit distance (ED) metric using the Damerau-Levenshtein distance (Damerau 1964). ED is computed separately for verbs, nouns, and actions sequences. Given an observed video with $N_{seg} = 8$, we report the minimum edit distance between $K = 5$ predicted sequences, each of length $Z = 20$. For the EK-55 and EGTEA datasets, we follow the setting in (Nagarajan et al. 2020), using mean average precision (mAP) for multi-label classification as the evaluation metric. The task involves observing the first $P\%$ of each video and predicting the actions that will occur in the remaining $(100 - P)\%$ of the video. Here actions are defined to verbs only. We consider $P = [25, 50, 75]$ to represent different anticipation horizons and report performance on the validation set for all target actions (All), frequently appeared actions (Freq), and rarely appeared action (Rare) respectively.

Implementation Details

For action recognition, we utilize the frozen encoder CLIP ViT-L/14 to extract visual features and then employ a Transformer encoder with 8 attention heads. For ICAF module, we utilize BLIP2-OPT-2.7B (Li et al. 2023) as the frozen visual encoder, LLaMA 3.2-9B as the VLM to derive behavioral intentions, along with LLaMA 3-8B (Dubey et al. 2024) as the text encoder and the LLM for anticipation. The

Method	Venue	Visual Encoder	Noun ↓	Verb ↓	Action ↓
PaMsEgoAI (Ishibashi et al. 2023)	arXiv’23	-	0.6291	0.6702	0.8753
HAI-PUI (Zhong et al. 2024)	arXiv’24	-	0.6733	0.7721	0.9242
AntGPT (Zhao et al. 2023)	ICLR’23	CLIP	0.6755	0.6728	0.8931
PlausiVL* (Mittal et al. 2024)	CVPR’24	-	0.6466	0.6618	0.8771
EgoVideo (Pei et al. 2024)	arXiv’24	EgoVideo-V	<u>0.6264</u>	<u>0.6576</u>	<u>0.8619</u>
PALM (Kim et al. 2024)	ECCV’24	EgoVLP	0.6465	0.7111	0.8819
ICVL (Ours)	-	CLIP	0.6194	0.6516	0.8570

Table 1: Long-term action anticipation performance on Ego4D. The results with **bold** and underline indicate the highest and second-highest values, respectively. * denotes our reproduced results. Rows with gray shading represent LLM-based method. Visual Encoder refers to the visual encoder of the action recognition model.

Method	Venue	EK-55			EGTEA		
		ALL ↑	FREQ ↑	RARE ↑	ALL ↑	FREQ ↑	RARE ↑
Timeception (Hussein, Gavves, and Smeulders 2019a)	CVPR’19	35.6	55.9	26.1	74.1	79.7	59.7
VideoGraph (Hussein, Gavves, and Smeulders 2019b)	arXiv’19	22.5	49.4	14.0	67.7	77.1	47.2
EGO-TOPO (Nagarajan et al. 2020)	CVPR’20	38.0	56.9	29.2	73.5	80.7	54.7
Anticipatr (Nawhal, Jyothi, and Mori 2022)	ECCV’22	39.1	58.1	29.1	76.8	83.3	55.1
AntGPT (Zhao et al. 2023)	ICLR’23	40.1	58.8	<u>31.9</u>	80.2	84.8	72.9
PALM (Kim et al. 2024)	ECCV’24	<u>40.4</u>	<u>59.3</u>	30.3	<u>80.7</u>	<u>85.0</u>	<u>73.5</u>
ICVL (Ours)	-	43.3	61.6	33.8	81.0	85.2	73.7

Table 2: Long-term action anticipation performance on EK-55 and EGTEA datasets.

Adam optimizer is used for end-to-end training with a learning rate of 5×10^{-5} , over 8 epochs.

Results and Analysis

Comparison to state-of-the-art. We compare ICVL with the current state-of-the-art approaches. Table 1 shows the performance comparison on the Ego4D dataset where our method consistently outperforms the previous SOTA (Kim et al. 2024; Pei et al. 2024) in terms of edit distance for noun, verb, and action with an improvement of $\{2.71\%, 5.95\%, 2.49\% \}$ and $\{0.7\%, 0.6\%, 0.49\% \}$, respectively. Notably, most methods using LLMs need to obtain the observed action labels, and the accuracy of action recognition models varies with different visual encoders. Specifically, the action recognition accuracy is 7.97% for CLIP encoder (Radford et al. 2021), 20.63% for EgoVLP encoder (Lin et al. 2022), and 27.64% for EgoVideo encoder (Pei et al. 2024). A direct relationship between the recognition accuracy and the final anticipation performance can be clearly observed. Results show that ICVL achieves a significant performance improvement of $\{5.61\%, 2.12\%, 3.61\% \}$ over other approach using the same CLIP encoder (Zhao et al. 2023). Additionally, it still outperforms methods that employ stronger visual encoders, delivering the best anticipation performance overall. This indicates that our method is more robust and reliable, where the learned intention-enhanced visual embeddings and selected examples effectively mitigate the noise of observed action labels. Among the LLM-based methods, AntGPT, PlausiVL, PALM, EgoVideo only use textual inputs while PlausiVL focuses only on the original visual embeddings. Our method emphasizes the integration of infor-

mation from both modalities, demonstrating that LLMs can achieve accurate predictions by leveraging enhanced visual features and carefully designed textual prompts.

Table 2 presents a comparison between our method and previous state-of-the-art approaches on the EK-55 and EGTEA datasets. Our method achieves the best performance on both datasets, with particularly notable results on EK-55 dataset, showing an improvement of 2.9%, 2.3% and 1.9% on all actions, frequently happened actions and rarely happened actions respectively.

Ablation Studies

Effectiveness of the two proposed modules. The results on the Ego4D dataset of ICAF and Example Selection modules are provided in Table 3. It is evident that both modules contribute to a significant overall improvement in model performance, with ICAF having the greatest impact. This is primarily because intentions can enhance the extraction of discriminative information from visual features, providing critical visual cues for actions’ evolution and aiding LLMs in making predictions. Additionally, the carefully selected examples also enrich the inputs to the LLM, enhancing the in-context learning ability of the LLM.

Design of ICAF. We examine the design of the ICAF module from two aspects: the approach to generating intentions and the integration between intention and visual embeddings. Results concerning the ways of generating intentions on the Ego4D dataset are reported in Table 4, of which *Visual features* represents generating latent intention embeddings solely from visual features using learnable tokens, *Action labels* means generating intention from observed action

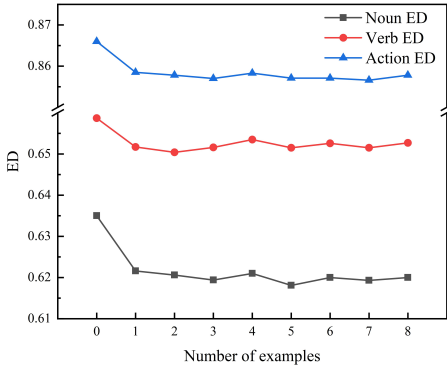


Figure 4: Ablation study on the number of the Selected Examples.

Method	Noun ↓	Verb ↓	Action ↓
Baseline	0.6927	0.6823	0.8944
Baseline w/ ES	0.6549	0.6759	0.8813
Baseline w/ ICAF	0.6287	0.6550	0.8643
ICVL	0.6194	0.6516	0.8570

Table 3: Ablation study on ICAF and Example Selection (ES). Baseline refers to fine-tuning LLM only with text-prompt input.

labels via LLM, and *VLM* refers to generating intentions through a VLM. Compared with *Baseline* without using intentions, results show that integrating intentions effectively enhances the discriminative power of visual features and improves LLM predictions. Using a VLM to infer intentions from video inputs achieves the best performance.

Table 5 demonstrates the impact of different integration strategy. *Concat* refers to concatenating the intention and visual embeddings. *CrossAttn (V)* denotes employing visual embeddings as the query in cross-attention method, whereas *CrossAttn (I)* utilizes intention embeddings as the query. Results illustrate that the cross-attention methods are superior to the method of simple concatenation. This indicates that cross-attention method successfully integrates intention embeddings into the visual embeddings, allowing the visual cues relevant to the intentions to be highlighted, thus obtaining intention-enhanced visual embeddings. Additionally, the choice of modality for the query affects performance, with using intention as the query yielding the best results.

Influence of the number of examples. Figure 4 shows the impact of the number of examples on the Ego4D dataset. As observed, providing a high-quality example significantly improves model performance. However, the relative performance improvement decreases as the number of examples increases. Once the number of examples reaches three, further increases have a negligible effect on performance, with optimal results attained when seven examples are provided. This is likely due to the decreasing relevance of subsequent examples, which fail to provide more meaningful guidance to the LLM. We report the final performance based on the

Derivation	Noun ↓	Verb ↓	Action ↓
Baseline	0.6469	0.6661	0.8773
Visual features	0.6454	0.6580	0.8733
Action labels	0.6383	0.6605	0.8680
VLM	0.6287	0.6550	0.8643

Table 4: Ablation study on intention generation. Baseline refers to using visual embeddings without intention integration.

Method	Noun ↓	Verb ↓	Action ↓
Concat	0.6329	0.6610	0.8676
CrossAttn (V)	0.6285	0.6580	0.8656
CrossAttn (I)	0.6287	0.6550	0.8643

Table 5: Ablation study on the integration between intention and visual embeddings.

Examples	Noun ED ↓	Verb ED ↓	Action ED ↓
Text Similarity	0.7299	0.6994	0.9076
Visual Similarity	0.7330	0.6948	0.8993
Fused Similarity	0.7128	0.6646	0.8912

Table 6: Ablation study on the example selection method.

use of three examples.

Effectiveness of example selection method. Table 6 illustrates the impact of example selection based on different modalities on the Ego4D dataset. The results indicate that examples selected based on the visual modality outperform those chosen based on textual similarity, likely due to the authenticity of the visual information. Our multimodal selection strategy, which considers both modalities, identifies the most relevant examples, proving the effectiveness of our approach.

Conclusion

In this study, we explore how to effectively utilize both visual and textual modalities through LLM to tackle LTA tasks. To make visual features more discriminative, we introduce an Intention-Context Attention Fusion mechanism that integrates visual embeddings with behavior intentions inferred by VLM. Furthermore, to improve the LLM’s understanding of the task and enhance its in-context learning capabilities, we propose a multi-modality example selection mechanism that provides more relevant examples. Extensive experiments on Ego4D, EPIC-Kitchens-55 and EGTEA GAZE+ datasets validate the effectiveness of our Intention-Conditioned Vision-Language model. While this work represents a preliminary investigation into multimodal fusion method using LLM, future research may focus on improving the logical consistency and coherence of action sequences predicted by the LLM.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62376217, 62576279, 62301434), Young Elite Scientists Sponsorship Program by CAST (No. 2023QNRC001), and the Joint Research Project for Meteorological Capacity Improvement (Grants 24NLTSZ003).

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Cao, C.; Sun, Z.; Lv, Q.; Min, L.; and Zhang, Y. 2024. 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*.
- Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.-D.; et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 958–979.
- Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2020. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11): 4125–4141.
- Damerau, F. J. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3): 171–176.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Furnari, A.; and Farinella, G. M. 2020. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11): 4021–4036.
- Gong, D.; Lee, J.; Kim, M.; Ha, S. J.; and Cho, M. 2022. Future transformer for long-term action anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3052–3061.
- Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18995–19012.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Hussein, N.; Gavves, E.; and Smeulders, A. W. 2019a. Time-ception for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 254–263.
- Hussein, N.; Gavves, E.; and Smeulders, A. W. 2019b. Videograph: Recognizing minutes-long human activities in videos. *arXiv preprint arXiv:1905.05143*.
- Ishibashi, T.; Ono, K.; Kugo, N.; and Sato, Y. 2023. Technical report for ego4d long term action anticipation challenge 2023. *arXiv preprint arXiv:2307.01467*.
- Kim, S.; Huang, D.; Xian, Y.; Hilliges, O.; Van Gool, L.; and Wang, X. 2024. Palm: Predicting actions through language models. In *European Conference on Computer Vision*, 140–158. Springer.
- Koppula, H. S.; and Saxena, A. 2015. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1): 14–29.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, Y.; Liu, M.; and Rehg, J. M. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, 619–635.
- Li, Y.; Wang, C.; and Jia, J. 2024. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, 323–340. Springer.
- Lin, K. Q.; Wang, J.; Soldan, M.; Wray, M.; Yan, R.; Xu, E. Z.; Gao, D.; Tu, R.-C.; Zhao, W.; Kong, W.; et al. 2022. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35: 7575–7586.
- Mittal, H.; Agarwal, N.; Lo, S.-Y.; and Lee, K. 2024. Can’t make an Omelette without Breaking some Eggs: Plausible Action Anticipation using Large Video-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18580–18590.
- Nagarajan, T.; Li, Y.; Feichtenhofer, C.; and Grauman, K. 2020. Ego-topo: Environment affordances from egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 163–172.
- Nawhal, M.; Jyothi, A. A.; and Mori, G. 2022. Rethinking learning approaches for long-term action anticipation. In *European Conference on Computer Vision*, 558–576. Springer.
- Pasca, R.-G.; Gavryushin, A.; Hamza, M.; Kuo, Y.-L.; Mo, K.; Van Gool, L.; Hilliges, O.; and Wang, X. 2024. Summarize the Past to Predict the Future: Natural Language Descriptions of Context Boost Multimodal Object Interaction Anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18286–18296.
- Pei, B.; Chen, G.; Xu, J.; He, Y.; Liu, Y.; Pan, K.; Huang, Y.; Wang, Y.; Lu, T.; Wang, L.; et al. 2024. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*.

- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Rodin, I.; Furnari, A.; Mavroeidis, D.; and Farinella, G. M. 2021. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 211: 103252.
- Sadegh Aliakbarian, M.; Sadat Saleh, F.; Salzmann, M.; Fernando, B.; Petersson, L.; and Andersson, L. 2017. Encouraging lstms to anticipate actions very early. In *Proceedings of the IEEE International Conference on Computer Vision*, 280–289.
- Sener, F.; Singhanian, D.; and Yao, A. 2020. 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*, 154–171. Springer.
- Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; et al. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.
- Wang, J.; Chen, G.; Huang, Y.; Wang, L.; and Lu, T. 2023. Memory-and-anticipation transformer for online action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13824–13835.
- Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; et al. 2024. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.
- Yao, H.; Tang, X.; Wei, H.; Zheng, G.; and Li, Z. 2019. Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 5668–5675.
- Yuan, H.; Chen, Y.; Ji, Z.; Zheng, Z.; Gu, Y.; and Zhou, J. 2025. Throughout procedural transformer for online action detection and anticipation. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Zhang, Y.; Zhang, J.; Zhang, J.; Wang, J.; Lu, K.; and Hong, J. 2020. A novel learning framework for sampling-based motion planning in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1202–1209.
- Zhao, Q.; Wang, S.; Zhang, C.; Fu, C.; Do, M. Q.; Agarwal, N.; Lee, K.; and Sun, C. 2023. Antgpt: Can large language models help long-term action anticipation from videos? *arXiv preprint arXiv:2307.16368*.
- Zhong, Z.; Martin, M.; Diederichs, F.; and Beyerer, J. 2024. Querymamba: A mamba-based encoder-decoder architecture with a statistical verb-noun interaction module for video action forecasting@ ego4d long-term action anticipation challenge 2024. *arXiv preprint arXiv:2407.04184*.