

Can't make an Omelette without Breaking some Eggs: Plausible Action Anticipation using Large Video-Language Models

Himangi Mittal^{1,2*} Nakul Agarwal¹ Shao-Yuan Lo¹ Kwonjoon Lee¹

¹Honda Research Institute, USA ²Carnegie Mellon University

hmittal@andrew.cmu.edu {nakul.agarwal, shao-yuan.lo, kwonjoon.lee}@honda-ri.com

Abstract

We introduce *PlausiVL*, a large video-language model for anticipating action sequences that are plausible in the real-world. While significant efforts have been made towards anticipating future actions, prior approaches do not take into account the aspect of plausibility in an action sequence. To address this limitation, we explore the generative capability of a large video-language model in our work and further, develop the understanding of plausibility in an action sequence by introducing two objective functions, a counterfactual-based plausible action sequence learning loss and a long-horizon action repetition loss. We utilize temporal logical constraints as well as verb-noun action pair logical constraints to create implausible/counterfactual action sequences and use them to train the model with plausible action sequence learning loss. This loss helps the model to differentiate between plausible and not plausible action sequences and also helps the model to learn implicit temporal cues crucial for the task of action anticipation. The long-horizon action repetition loss puts a higher penalty on the actions that are more prone to repetition over a longer temporal window. With this penalization, the model is able to generate diverse, plausible action sequences. We evaluate our approach on two large-scale datasets, *Ego4D* and *EPIC-Kitchens-100*, and show improvements on the task of action anticipation.

1. Introduction

Having the ability to predict future events is a critical component in the decision-making process of an AI agent. For example, for an autonomous driving car, being able to anticipate the next sequence of actions for cars, pedestrians, and other agents in the scene can ensure safety of pedestrians as well as vehicles. To enable this, the model should be able to reason effectively from the spatial as well as temporal in-

* This work was done as Himangi Mittal's internship project at Honda Research Institute, USA.

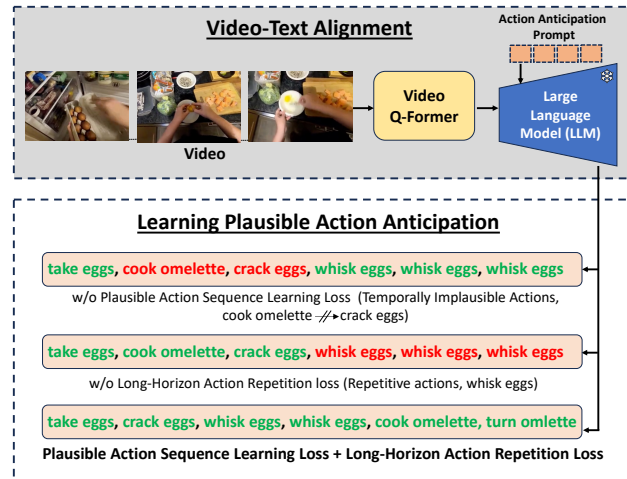


Figure 1. We present a large video-language model for learning to anticipate action sequences that are *plausible* in the real-world. We show an example of a kitchen-based environment. By using a large video-language model, we leverage their generative capabilities to anticipate future actions and further train the model with two devised objective functions: plausible action sequence learning loss and long-horizon action repetition loss. Without the plausible action sequence learning loss, the model has less temporal understanding and generates a temporally implausible action sequence of *cook omelette* \rightarrow *crack eggs*. Similarly, without the long-horizon action repetition loss, the model generates less diverse actions and repeats the same action, *whisk eggs* \rightarrow *whisk eggs* \rightarrow *whisk eggs*. When training the model with the two objective functions combined, our method is able to generate plausible action sequences which are temporally accurate, *crack eggs* \rightarrow *cook omelette* and more diverse with less repetition, *whisk eggs* \rightarrow *whisk eggs* \rightarrow *cook omelette*.

formation of the visual scene. This has led to a growing interest in the task of *Action Anticipation*. Action anticipation refers to the predictive task of forecasting future actions or activities given a sequence of visual data, typically videos. For example, in a kitchen-based environment, if a human has performed the following series of actions, *open fridge* \rightarrow *take eggs* \rightarrow *close fridge*, the model should be able to

reason that *crack eggs* could be one of the plausible future actions.

However, action anticipation is challenging because the uncertainty in precisely predicting the future makes the task non-deterministic in nature. In other words, given what has happened so far, there are infinitely many possibilities for what future actions might happen. Moreover, action anticipation is accompanied by an additional challenge of understanding the implicit temporal information present in an action sequence, which makes the sequence *plausible* in the real-world. For example, the model should be able to understand that an action like *crack eggs* will always happen *before* *cook omelette* as shown in Figure 1.

To this end, there has been some progress for the action anticipation task. Earlier works have explored an LSTM based approach by summarizing the past and inferring the future [17, 43], by logging the past history actions in text [38], or using RNN-based approaches [51, 54] by learning goals. However, such LSTM/RNN-based approaches are unable to effectively capture the temporal relations among the actions over a long horizon due to their sequential nature. There have been efforts to resolve this limitation by implementing different types of attention mechanism [25], skip-connection [30], and message-passing framework [56]. However, they show marginal improvement in modeling longer horizon. Recent works have also explored transformer-based approaches [24, 25, 52], with a memory-based system [65] or leveraging multiple-modalities [70]. While transformer-based approaches are able to model longer temporal understanding, they can still become confined to the information present in the training data and cannot model the diverse nature of the future actions. They rely on the ability of the transformer encoder to learn from the given training data which limits their generalization and scaling capability.

To overcome the above challenges, recent methods [3, 31, 32, 61] have attempted to leverage the autoregressive text generation capabilities of generative large-language models (LLMs) to improve generalizability for various vision tasks. Taking inspiration from these works and to address the challenges present in anticipating plausible actions, we introduce **PlausiVL**, **Plausible** action anticipation through a large **Video-Language** model.

Given the generative capabilities of large language models, in this work, we introduce a video-large-language model which can efficiently model and leverage the temporal cues present in a video to generate plausible action sequences for the task of action anticipation. We use a Q-former [31] based transformer architecture to embed videos into spatio-temporal visual representations. This architecture ensures an effective alignment between the visual features and the desired text in the LLM embedding space. In addition to the alignment, we try to address the challenges

that are specifically present in the task of action anticipation and thus, introduce a method with the following important characteristics: 1). The ability to understand the temporal correlations present among the actions in a sequence which in turn makes the action sequence temporally *plausible* in the real-world, 2). Being able to model the diverse, possible actions that can happen in the future. For example, for the former characteristic, a model should follow a temporal constraint that *an action X has to happen before for the action Y to happen* to make the sequence *action X \rightarrow action Y* plausible in the real-world.

To build such temporal understanding required for generating plausible action sequences, we design a counterfactual-based plausible action sequence learning loss where we create temporal logic constraints and train the model to be able to differentiate between the plausible and not plausible action sequences. Additionally, we also use verb-noun action logical constraints to further improve the model’s understanding about which verbs are possible with which nouns to create a plausible action in the real-world (for example, *cook spoon* is not a plausible action). To our knowledge, the aspect of plausibility in generating an action sequence has not been explored for the task of action anticipation. While this loss is helpful for efficient temporal understanding, we also aim for the model to be able to understand the diverse nature of actions and generate plausible action sequences with less repeated actions as language models are prone to the issue of repetition. To resolve this, we devise a long-horizon action repetition loss where the later actions that are more prone to repetition have a higher penalty and the earlier, immediate actions have lower penalty. We summarize our contributions as follows:

1. We present **PlausiVL**, a large video-language model which leverages the spatial-temporal information present in videos for anticipating plausible future action sequences.
2. To learn the temporal cues and understand the temporal dependencies among actions in a plausible sequence, we design a counterfactual-based plausible action sequence learning loss. We create temporal logic rules and verb-noun action pair logic constraints for the model to be able to understand plausibility in action sequences.
3. To be able to generate less diverse future actions with less repetition, we devise a long-horizon action repetition loss by penalizing the longer-horizon actions more.

2. Related Works

Large Language Models. Language Modeling is a method to model the generative likelihood over the word token sequences and predict the probabilities of the next/future tokens. Large language models (LLMs) [5, 10, 58, 59] are transformers with billions of parameters that have been trained on massive amounts of data and have shown im-

pressive capabilities on the task of question-answering and chat-conversation with humans. Methods like in-context learning [5], prompt tuning [64], chain-of-thought reasoning [63], and reinforcement learning with human feedback [11, 44] have improved the language models to perform very well on few-shot tasks. While these models show great capabilities in understanding the input and solving complex tasks via text generation, these models can only understand the text modality and are at a loss of the rich information that is present in other modalities like video, audio. In our work, we utilize videos as input and learn from the visual and temporal information present in them.

Large Vision-Language Models. Recent strides in this domain have seen diverse pre-training methods leveraging extensive multimodal datasets driving the progress of large vision-language models. Some models [20, 29, 35, 48, 62] merge visual and linguistic modalities by co-training text and image encoders using contrastive loss on large datasets containing image-caption pairs. Meanwhile, other approaches [3, 7] integrate visual input directly into language model decoders through a cross-attention mechanism, eschewing the use of images as additional prefixes. Another category of vision-language models [9, 33, 34, 37, 55, 57] leverage Masked-Language Modeling (MLM) and Image-Text Matching (ITM) objectives to align image segments with text. BLIP-2 [31] was one of the works which proposed a Qformer-based method to ensure visual-text alignment. Since these works explore the image-text alignment, they are unable to model and understand the temporal information that is present in videos. There have been efforts towards video-text alignment by using a linear layer to project the video space to the LLMs textual space [6] in Video-LLM or by using a Q-former based module [68] in Video-LLaMA. While these works explore video-text alignment, these models can be ineffective for the task of action anticipation as they do not understand the temporal correlations among the actions in a sequence.

Temporal and symbolic logic reasoning. Symbolic logic reasoning is a method to create a system of rules and symbols in the form of logical expressions. Temporal logic reasoning specifically designs logical expressions for representing and reasoning about time. Linear temporal logic [46], metric temporal logic [42], signal temporal logic [15], and interval temporal logic [27] are some methods for develop temporal logical rules. We take inspiration from the work DTL [66] to generate temporal logic rules and create counterfactual sequences of actions.

Action Anticipation. This task has been explored for third-person videos [2, 8, 22, 49, 60] as well as egocentric videos [12, 13, 18, 24, 26, 47, 50]. Standard approaches for this task can be divided into LSTM/RNN-based [13, 54] approaches and transformer-based approaches. LSTM-based approaches [17, 43] mainly use a rolling LSTM to encode

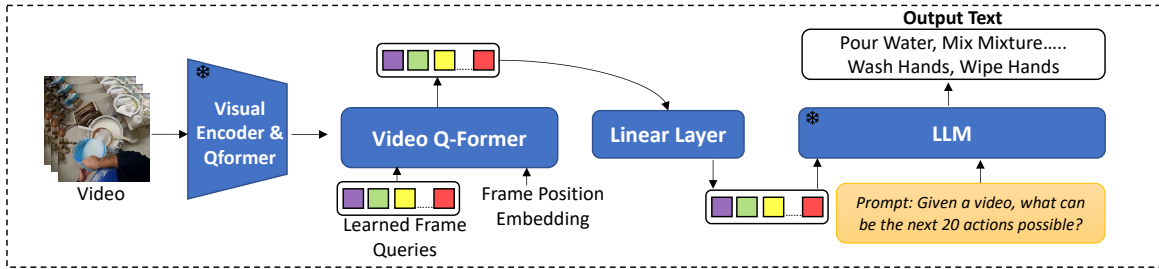
the observed video and store an updated summary. For inference, an unrolling LSTM is initialized with the hidden and cell state of the rolling LSTM to predict the next action. While LSTM/RNNs have shortcomings in modeling long-horizon temporal dependencies, some approaches mitigate this issue via goal-based learning [51], diverse attention mechanism [25], skip-connections [30], message passing framework [56], memory-based modules [38, 65] or similarity metric [16]. Recent works have explored transformer-based [24, 26] approaches with global attention [25], modelling appearance change in human-object interactions [52], conditioning on intention [40], hierarchical feature aggregation [40]. While most of the works explore it in a unimodal setting by using the visual modality, other works also present a multi-modal approach for this task by using optical flow [17, 43], object-based features [17, 43, 67] or audio [41, 70]. Other works explore uncertainty-based methods [1, 19] and GAN-based approach [21]. We take inspiration from the object detection [36] literature for the repetition loss. Concurrent to our work, there have been text-based LLM approaches [28, 69] which explore the task of action anticipation, however, they only operate in the textual space and lose the visual-temporal information present in video.

3. Method

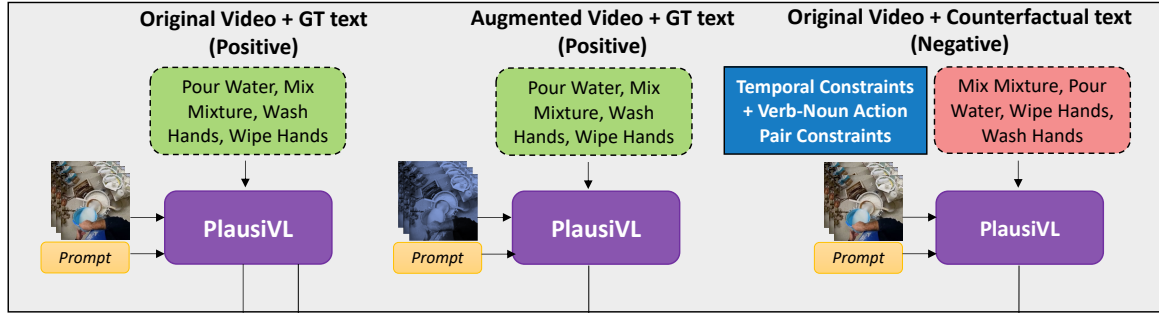
In the following sections, we present the details of our method, PlausiVL, to learn the temporal cues for plausible action sequence generation.

3.1. Model Architecture

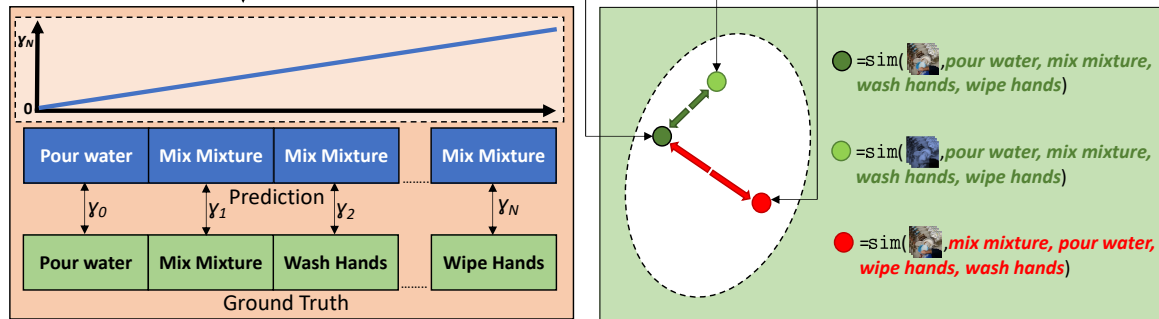
Given a video clip of N frames, $V = [v_1, v_2, v_3 \dots v_N]$, we use a frozen visual encoder (ViT) to extract video-frame-level representations, $V = [v'_1, v'_2, v'_3 \dots v'_N]$. After this, each frame feature is passed through a Q-former [31] with k number of query tokens, to get the d_q -dimensional visual representation as $v''_i \in \mathbb{R}^{k \times d_q}$. These queries are helpful in extracting the visual features with the most information aligned to the text. For the frames to have an understanding of the temporal relations among them, a frame position embedding layer is applied to each Q-former feature. At the same time, we also apply a clip-position embedding layer to infuse more grouping information about the frames that belong to a clip. These features are then passed through a video Q-former to aggregate the spatio-temporal information of the video. Finally, a linear projection layer is used to project these output representations to the LLM text embedding space of d_l dimension, $v_i \in \mathbb{R}^{k_l \times d_l}$. These video embeddings can be considered as *visual prompts* which are concatenated with the input text embeddings t_i to make the LLM generate text conditioned on the video content.



(a) PlausiVL



(b) Augmentation



Long-Horizon Action Repetition Loss

Plausible Action Sequence Learning Loss

(c) Objective Functions and Training

Figure 2. **Model diagram:**(a) **PlausiVL**: Given a video, a frozen visual encoder a Q-former with k number of query tokens is used to extract frame level representations which are further concatenated with a frame position embedding layer to add temporal understanding. Next, the representations are passed through the video Q-former and a linear layer is added to project these features into the LLM space. These visual embeddings (visual prompts) and are concatenated with text-prompts to get the desired output text (Sec 3.1), (b) **Augmentation**: For plausible action anticipation, we use logical rules to create counterfactual implausible action sequences. Given an input video, we create a positive augmentation of the video and a negative augmentation by using temporal logical and verb-noun action pair constraints (Sec 4.1). (c) **Objective Functions and Training**: We train our model with two losses: (i) Plausible Action Sequence Learning Loss ($\mathcal{L}_{\text{plau}}$) which aligns the original video-plausible text pair closer to the positive augmentation of video-plausible text, and brings the original video-plausible text far apart from the video-counterfactual text. (Sec 4.1), (ii) long-horizon action repetition loss that ensures diverse and less repetitive actions by adding a higher penalty to the later tokens (mix mixture and wipe hands) and lower penalty to immediate future actions (pour water, pour water). The graph shows the linearly increasing γ penalty for the tokens over the long-horizon (Sec 4.2).

4. Training

While the above backbone network ensures the alignment of the visual features with the LLM textual space, we also focus on making the model learn to better understand long-horizon temporal dependencies among the actions which is crucial for plausible action anticipation. To develop such

temporal understanding in a model, we train our system to optimize two losses, (1). Plausible Action Sequence Learning loss $\mathcal{L}_{\text{plau}}$ and (2). Long-horizon action repetition loss \mathcal{L}_{rep} . With these two losses, the model can understand the temporal cues better to be able to generate a *plausible* and diverse sequence of future actions.

4.1. Plausible Action Sequence Learning loss

For a model to be able to understand the plausible nature of an action sequence, it should be able to leverage the implicit temporal information present in input videos. Thus, we design a self-supervised plausible action sequence learning loss, $\mathcal{L}_{\text{plau}}$. The key idea is to create counterfactuals based on temporal logical constraints as well as verb-noun action pair logical constraints and optimize the network by minimizing a loss with two negative log-likelihood terms: (1) *increase* the probability of associating the visual modality with the temporally correct and plausible sequence of actions, and (2) *decrease* the probability of associating the video with the action sequences that are not plausible in the real-world and temporally incorrect. Here, sequences of action that satisfy the temporal as well as verb-noun action pair logic constraints are considered as logically correct.

Temporal logical constraints: In our work, we define a temporal constraint for an action sequence as follows: *an action X that has to happen before an action Y* to make it a plausible sequence in the real-world. Consider for example, given a sequence of *take eggs* \rightarrow *crack eggs* \rightarrow *whisk eggs* \rightarrow *cook omelette*, a counterfactual of this sequence of actions would be, *take eggs* \rightarrow *cook omelette* \rightarrow *whisk eggs* \rightarrow *crack eggs* since *crack eggs* would always happen before *cook omelette*. Mathematically, we define it as follows:

$$CF^{\text{temp}}(a_i, a_j) = \begin{cases} 1, & \text{if } \forall t \in T (t_{a_i} \rightarrow t_{a_j}) \wedge \neg(t_{a_j} \rightarrow t_{a_i}), \\ -1 & \text{if } \forall t \in T (t_{a_j} \rightarrow t_{a_i}) \wedge \neg(t_{a_i} \rightarrow t_{a_j}), \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

where $CF^{\text{temp}}(a_i, a_j)$ is an action pair matrix with a value of 1 if a_i always happens before a_j for all the ground truth sequences $t \in T$, a value of -1 if a_i always happens after a_j , and 0 otherwise if there is no relation between the two actions. With this temporal logical constraint, given a text sequence t , we perform a swap operation if there is a forward or backward relation between an action pair. Hence, given a ground truth text sequence t , we define the operation if a_j always happens before a_p as follows:

$$t^{cf}(a_i, a_j, a_p, a_n) = \begin{cases} a_i, a_p, a_j, a_n, & \text{if } CF^{\text{temp}}(a_j, a_p) = 1, \\ a_i, a_j, a_p, a_n, & \text{otherwise.} \end{cases} \quad (2)$$

Similarly, we define the operation if a_j always happens after a_i as follows:

$$t^{cf}(a_i, a_j, a_p, a_n) = \begin{cases} a_j, a_i, a_p, a_n, & \text{if } CF^{\text{temp}}(a_j, a_i) = -1, \\ a_i, a_j, a_p, a_n, & \text{otherwise.} \end{cases} \quad (3)$$

Next, we define the another logical constraint - verb-noun action pair constraint.

Verb-Noun Action pair constraints: For this, we create a counterfactual where a verb-noun action pair is not plausible in the real-world, for example, *cook spoon*. We define a verb-noun action constraint as follows: *a verb-noun pair consisting of an action verb that is plausible with the object noun in the real-world*. Mathematically, we define it as follows:

$$CF^{\text{act}}(a_i, a_j) = \begin{cases} 1, & \text{if } \forall t \in T \neg(a_i^v \wedge a_j^n), \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

where $CF^{\text{act}}(a_i, a_j)$ is a verb-noun pair matrix with a value of 1 if for a verb, the corresponding noun is not plausible or vice-versa in all the ground truth actions $t \in T$ and 0 otherwise if the verb-noun pair is plausible. Similar to the temporal constraints mentioned above, with this verb-noun action pair constraint, given an action, we swap either the verb or noun with a uniform probability to create implausible verb-noun action pairs. Given a text action pair t , we define the operation of a counterfactual, implausible verb-noun action pair as follows:

$$t^{cf}(a_i^v, a_i^n) = \begin{cases} (a_i^v, a_j^n) || (a_j^v, a_i^n), & \text{if } CF^{\text{act}}(a_i^v, a_j^n) = 1, \\ (a_i^v, a_i^n), & \text{otherwise.} \end{cases} \quad (5)$$

Loss: With this, for every video-text action sequence pair (V_i, T_i) in the dataset \mathcal{D} , we create a temporal as well as verb-noun action pair counterfactual T_i^{cf} for every textual ground truth text sequence and collect it as a dataset, \mathcal{D}_{vtcf} . Finally, we define plausible action sequence learning loss ($\mathcal{L}_{\text{plau}}$) as follows:

$$\mathcal{L}_{\text{plau}} = \mathbb{E}_{(v_i, t_i) \in \mathcal{D}_{vtcf}} \left[-\log(z(v_i, t_i, v'_i)) - \log(1 - z(v_i, t_i, t_i^{cf})) \right] \quad (6)$$

In the above equation, $z(v_i, t_i, v'_i)$ and $z(v_i, t_i, t_i^{cf})$ probabilities are computed as follows:

$$z(v_i, t_i, v'_i) = \sigma(\text{sim}(\Delta p(v_i, t_i), \Delta p(v'_i, t_i)) / \tau) \quad (7)$$

$$z(v_i, t_i, t_i^{cf}) = \sigma(\text{sim}(\Delta p(v_i, t_i), \Delta p(v_i, t_i^{cf})) / \tau) \quad (8)$$

where v_i and v'_i are the visual embeddings of the original video and augmented video (respectively), t_i and t_i^{cf} are the text embeddings of the ground truth text sequence and counterfactual text (respectively), τ is the temperature, σ is the sigmoid function, $\Delta p(\cdot, \cdot)$ is the cross-modal video-text representation from LLM after passing through a MLP projection layer (absorbed in the equation for better readability), and sim is the similarity function.

In summary, training the model to optimize the $\mathcal{L}_{\text{plau}}$ loss helps the model to differentiate between the plausible and counterfactual/implausible action sequences by aligning the visual modality closer to the temporally correct, plausible action sequence. By learning this alignment, it is able to understand the implicit temporal information that defines the dependencies and correlations among actions in a plausible sequence.

4.2. Long-Horizon Action Repetition Loss

While the plausible action sequence learning loss $\mathcal{L}_{\text{plau}}$ helps the model to understand the implicit temporal information present in the action sequences, we consider another aspect of plausibility by reducing the repetition of actions and in turn generating more diverse actions. We observe that while the model is able to generate accurate, temporally correct, and diverse actions over a short temporal window, it starts repeating the same actions over a longer horizon. To mitigate this, we train the model by enforcing a larger penalty on the actions that happen over a longer horizon in the temporal window and lesser penalty to the actions that are immediately near to the observed video. We add a penalty of γ_t over the negative log-likelihood of the probability as follows:

$$p_t = \frac{\exp(\hat{y}_t)}{\sum_j \exp(\hat{y}_j)}, \quad (9)$$

$$\mathcal{L}_{\text{rep}}(p_t) = -\gamma_t \log(p_t) \quad (10)$$

where \hat{y}_t is the output from the language model for the t 'th token over which softmax operation is applied to get the probability p_t . γ_t is the γ value temporally unique to the t 'th token following the order, $\gamma_0 < \gamma_1 < \gamma_2 < \dots < \gamma_N$. In summary, by optimizing the \mathcal{L}_{rep} loss, the model is penalized more for the actions that happen over a longer horizon and less penalized for immediate actions. This is helpful in regulating repetition and ensuring more diverse actions in the generated text.

Finally, we train our model with the overall loss as:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{plau}} + \beta \mathcal{L}_{\text{rep}} \quad (11)$$

where α and β are the weight hyper-parameter for the two losses.

5. Experiments

5.1. Implementation Details

We process the videos of size 224×224 with Ego4D containing 8 clips with 4 frames, making a total of 32 frames, and EPIC-Kitchens-100 with 32 frames as well. We use the pretrained Qformer model, BLIP2-FlanT5xxl from BLIP2 [31] with number of query tokens as 32 and ViT-G/14 as our vision encoder. We train our method end-to-end

with a learning rate of $1e^{-5}$, for 100 epochs, and $\alpha = 0.5$ and $\beta = 0.5$. We use LLaMA-2-7B as our language model. For long-horizon action repetition loss, \mathcal{L}_{rep} , we use γ in the uniform distribution from $[0, 2]$ with number of steps equal to the number of output tokens from the language model. For plausible action sequence learning loss $\mathcal{L}_{\text{plau}}$, we use video augmentation of color jitter, random horizontal flip, and a random rotation of 10 degrees.

5.2. Experimental Setup

Datasets: We evaluate on two action anticipation datasets: Ego4D [26] and EPIC-Kitchens-100 [13]. Ego4D is a large-scale egocentric dataset covering diverse indoor and outdoor scenarios like home, workplace, etc. It consists of 3670 hours of videos with 115 verbs and 478 nouns. To evaluate our method on Ego4D, we use videos from the Forecasting and Hand-Object interaction subset and show results on the validation set. In Ego4D, a video and a stopping time is given, and the model predicts N sets of sequences having Z number of actions in the form of verb-noun pairs, $\{(\hat{v}_{z,n}, \hat{n}_{z,n})\}_{z=1}^Z\}_{n=1}^N$, where, $\hat{v}_{z,n}$ is the predicted verb and $\hat{n}_{z,n}$ is the predicted noun.

EPIC-Kitchens-100 [13] is an egocentric dataset of a kitchen-based environment. It consists of 100 hours of egocentric videos with 97 verbs and 300 nouns. For this dataset, given an action segment that starts at time τ_s , the model has to predict the anticipated action by observing a video segment of duration $[\tau_s - (\tau_o + \tau_a), \tau_s - \tau_a]$ where τ_o is the observation time and τ_a is the anticipation time. The anticipation time τ_a means how much in advance the model has to anticipate the action.

Baselines: We compare our method with large video-language models, Video-LLaMA [68] and Video-LLM [6]. We also compare our method with the transformer and LSTM-based approaches for action anticipation along with text-based large language models for a more exhaustive analysis.

Ablation Study: In the ablation study, we present results of PlausiVL with and without $\mathcal{L}_{\text{plau}}$ and \mathcal{L}_{rep} objective functions to show the effect of each component on the final performance of the model.

5.3. Discussion of Results

Referring to Table 1 and Table 2, we can observe that PlausiVL is able to perform better when compared with the baselines. This can be attributed to its ability to be able to understand the plausibility in the action sequences and leverage the temporal correlations among the actions in a sequence. We present a closer analysis of the results in our discussion following next.

PlausiVL shows performance gain towards action anticipation: Prior large video-language models [6, 68] have only explored the visual-text alignment and lack the tem-

| Method | ED@(Z=20) ↓ | |
|------------------|--------------|--------------|
| | Verb | Noun |
| RepLAI [41] | 0.755 | 0.834 |
| SlowFast [26] | 0.745 | 0.779 |
| ICVAE [39] | 0.741 | 0.739 |
| HierVL [4] | 0.723 | 0.734 |
| Video+CLIP [14] | 0.715 | 0.748 |
| AntGPT [69] | 0.700 | 0.717 |
| Video LLM [6] | 0.721 | 0.725 |
| Video LLaMA [68] | 0.703 | 0.721 |
| PlausiVL | 0.679 | 0.681 |

Table 1. Performance on Long-term action anticipation on Ego4D ↓: Lower is better. This shows that our method, PlausiVL is able to outperform all the previous baselines for verb, noun, and action.

| Method | Class-mean | | |
|---------------------------|--------------------|--------------|--------------|
| | Top-5 recall (%) ↑ | | |
| | Verb | Noun | Action |
| RU-LSTM [13] | 23.20 | 31.40 | 14.70 |
| Temporal Aggregation [53] | 27.80 | 30.80 | 14.00 |
| Video LLM [6] | - | - | 15.40 |
| AFFT [70] | 22.80 | 34.60 | 18.50 |
| AVT [24] | 28.20 | 32.00 | 15.90 |
| MeMViT [65] | 32.20 | 37.00 | 17.70 |
| RAFTformer [23] | 33.80 | 37.90 | 19.10 |
| InAViT [52] | 52.54 | 51.93 | 25.89 |
| Video LLaMA [68] | 52.90 | 52.01 | 26.05 |
| PlausiVL | 55.62 | 54.23 | 27.60 |

Table 2. Performance of action anticipation on EPIC-Kitchens-100 on class-mean Top-5 recall (%) ↑: Higher is better. Our method is able to outperform all the previous baselines.

| | | Ego4D | | EPIC-Kitchens-100 | | |
|-----------------------------|----------------------------|-------------|-------|-------------------------------|-------|--------|
| $\mathcal{L}_{\text{plau}}$ | \mathcal{L}_{rep} | ED@(Z=20) ↓ | | Class-mean Top-5 recall (%) ↑ | | |
| | | Verb | Noun | Verb | Noun | Action |
| ✓ | ✓ | 0.679 | 0.683 | 55.62 | 54.23 | 27.60 |
| ✓ | | 0.686 | 0.698 | 54.50 | 53.60 | 26.67 |
| | ✓ | 0.691 | 0.707 | 54.15 | 53.03 | 26.21 |
| | | 0.703 | 0.721 | 52.90 | 52.01 | 26.05 |

Table 3. Ablation study of modules, $\mathcal{L}_{\text{plau}}$ and \mathcal{L}_{rep} in our method on Ego4D ↓: Lower is better, and EPIC-Kitchens-100 on class-mean Top-5 recall (%) ↑: Higher is better. The analysis that starting from our method, row (1), there is a dip in the performance as each module is removed showing that the losses, $\mathcal{L}_{\text{plau}}$ and \mathcal{L}_{rep} are helpful in improving the performance.

poral understanding needed for the action anticipation. To show that our model is able to learn the temporal understanding, we compare PlausiVL with Video-LLM and Video-LLaMA in Table 1 and observe an improvement of

| | BLEU Score ↑ | Repetition Score ↓ |
|------------------|--------------|--------------------|
| Video-LLaMA [68] | 37.89 | 7.12 |
| PlausiVL | 45.54 | 5.87 |
| Ground Truth | 100.00 | 4.33 |

Table 4. BLEU score and Repetition Score on the Ego4D dataset. For BLEU score, ↑: Higher is better, and for repetition score, ↓: lower is better. We can observe that both the BLEU score and repetition score are better for PlausiVL than Video-LLaMA.

4.2% and 2.4%, respectively on verbs. Similarly, we observe an improvement of 2.72% and 2.22% on verbs for EPIC-Kitchens-100 in Table 2. The improvement in the performance emphasizes that the model is able to learn the temporal dependencies among the actions to generate more accurate and plausible action sequences. Qualitative results presented in Figure 3 also show the quality of our generated sequence in comparison to the ground truth. We can see that our method is able to understand the activity happening the video and anticipate the temporal future action sequence accordingly. We also exhaustively compare PlausiVL with prior approaches in Table 1 and Table 2 that utilize transformer and LSTM-based architectures and show that our method is able to perform better.

$\mathcal{L}_{\text{plau}}$ helps the model to learn plausible future action sequences: We hypothesize that for generating accurate future action sequences, a model should have an understanding about the temporal plausibility of an action sequence in the real-world. To assess if our devised loss function, plausible action sequence learning loss, $\mathcal{L}_{\text{plau}}$ is able to create such understanding in the model, we compare our method, row (1) and our method without $\mathcal{L}_{\text{plau}}$, rows (3) and (4) in Table 3. We observe by removing this module, there is a drop in the performance of 1.2 % on verbs for Ego4D and 1.47 % for verbs of EPIC-Kitchens-100 (row(1) and row(3) are compared). This shows that training a model with $\mathcal{L}_{\text{plau}}$ as an objective function helps the model to learn the implicit temporal information of action correlations in a sequence. Through learning to differentiate between the plausible and not plausible action sequences and aligning the video representations closer to the plausible action sequences, the model learns an effective video-text alignment which helps in generating more accurate, plausible future action sequences.

\mathcal{L}_{rep} helps with lesser repetition and more diversity over long horizons: We also try to address another aspect of plausibility in action sequences by making the model learn to generate sequences with less repetitive actions and more diverse actions via our devised objective function, long-horizon action repetition loss, \mathcal{L}_{rep} . To assess the efficacy of this module, we compare our method, row (1) and our method without \mathcal{L}_{rep} , row (2) and row (4) in Table 3. We observe that there is performance dip of 1.5 % on Ego4D nouns and 0.63 % on EPIC-Kitchens-100 nouns.

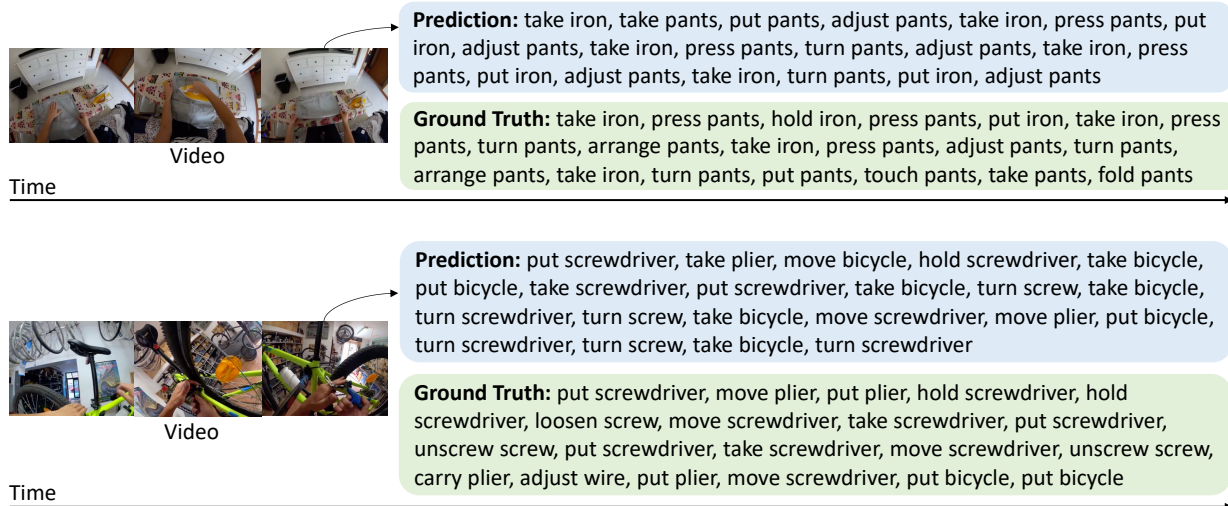


Figure 3. Qualitative Results: Given a video, the top blue box shows the prediction from PlausiVL and the green box contains the ground truth action sequence for reference. We can observe that PlausiVL is able to generate action sequences that satisfy the temporal logic constraints and are diverse with less repetitions. The predicted action sequence is also closer to the ground truth action sequence.

This indicates that by penalizing the actions more over the long horizon, \mathcal{L}_{rep} is able to reduce the repetition of actions in the sequence generation and hence, contribute towards plausible action anticipation sequences.

Large Video-language model vs Text-large-language-model: Given the exploration of text-only large language models, we also address the comparison between text-based LLM and large video-language models for the task of action anticipation. We compare PlausiVL with AntGPT [69] which is a text-based LLM and observe a performance gain of 2.1% on verbs and 3.6% on nouns for Ego4D from our method. We reason that a major drawback of text-based LLM for this task is that they completely discard the visual as well as temporal information present in the videos. Whereas, the task of action anticipation is highly dependent on the visual spatio-temporal information to understand the real-world temporal flow of actions and anticipate actions accurately. Incorporating visual modality can give crucial information such as the environment of the agent, the objects interacted with, and other objects in the scene which might be interacted with later in the future. Such vital information is lost when converting a video into textual actions [69] or into a summary [28]. Summarizing a video into text-based information can only provide the high-level details about a video, but it doesn't give a signal about the real-world temporal flow of the actions and objects in a video.

PlausiVL is able to generate plausible action sequences:

To further emphasize the plausibility, less repetition and quality our generated text, we compute the BLEU score [45] and repetition score. The repetition score is an average of the number of actions that are repeated in an action sequence and the BLEU score measures the similarity between our generated text and ground truth. We report the

results in Table 4. By having a better BLEU score than the baseline, we show that the generated text from our method is a more plausible action sequence, thus emphasizing the efficacy of our objective functions. Similarly, by having a lower repetition score than the baseline, we show that the model has lesser repetitive actions in the generated sequence. Our method repeats 5.87 actions in an action sequence on average whereas Video-LLaMA repeats an average of 7.12 actions. We also observe an average repetition of 4.33 actions in ground truth action sequences. Moreover, a lower edit distance metric in Table 1 also indicates less repetition and more plausibility in the generated text as a lower metric would mean less substitutions were made to bring the output text closer to the ground truth.

6. Conclusion

In this work, we leverage the generative capabilities of large video-language models for plausible action anticipation. In addition to the abilities of large video-language models, for the model to better understand the plausibility in an action sequence, we introduce a plausible action sequence learning loss which helps the model to differentiate between the plausible and not plausible action sequences, and thus learn anticipation related temporal cues. We further devise a long-horizon action repetition loss that puts a higher penalty on the actions that happen over a longer temporal window and are more prone to repetition, thus mitigating action repetition and ensuring more diverse actions. Experimental results show that our model is able to perform better by generating more plausible and accurate action sequences on Ego4D and EPIC-Kitchens-100. While our method is an initial step towards plausible action anticipation, there can be further exploration mitigating the issue of hallucinating implausible action sequences in the future work.

References

- [1] Yazan Abu Farha and Juergen Gall. Uncertainty-aware anticipation of activities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3
- [2] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018. 3
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2, 3
- [4] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23066–23078, 2023. 7
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2, 3
- [6] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023. 3, 6, 7
- [7] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022. 3
- [8] Junwen Chen, Gaurav Mittal, Ye Yu, Yu Kong, and Mei Chen. Github: Gated history unit with background suppression for online action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19925–19934, 2022. 3
- [9] Liunian Harold Chen, Yukun Zhu, Yen-Chun Shen, Heng Gao, Xiaodong Liu, Xiaohui Shen, Zhe He, Ricardo Henao, Renjie Miao, Yuan Guo, et al. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 3
- [10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2
- [11] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 3
- [12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 3
- [13] 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. *arXiv preprint arXiv:2006.13256*, 2020. 3, 6, 7
- [14] Srijan Das and Michael S Ryoo. Video+ clip baseline for ego4d long-term action anticipation. *arXiv preprint arXiv:2207.00579*, 2022. 7
- [15] Georgios E Fainekos and George J Pappas. Robustness of temporal logic specifications for continuous-time signals. *Theoretical Computer Science*, 410(42):4262–4291, 2009. 3
- [16] Basura Fernando and Samitha Herath. Anticipating human actions by correlating past with the future with jaccard similarity measures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13224–13233, 2021. 3
- [17] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020. 2, 3
- [18] Antonino Furnari and Giovanni Maria Farinella. Towards streaming egocentric action anticipation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1250–1257. IEEE, 2022. 3
- [19] Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3
- [20] Andreas Fürst, Elisabeth Rumetshofer, Johannes Lehner, Viet T Tran, Fei Tang, Hubert Ramsauer, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto, et al. Cloob: Modern hopfield networks with infolob outperform clip. *Advances in neural information processing systems*, 35: 20450–20468, 2022. 3
- [21] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the future: A jointly learnt model for action anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5562–5571, 2019. 3
- [22] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017. 3
- [23] Harshayu Girase, Nakul Agarwal, Chiho Choi, and Kartikeya Mangalam. Latency matters: Real-time action forecasting transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18759–18769, 2023. 7
- [24] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13505–13515, 2021. 2, 3, 7
- [25] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future transformer for long-term action

- anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3052–3061, 2022. 2, 3
- [26] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, 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*, pages 18995–19012, 2022. 3, 6, 7
- [27] Joseph Y Halpern and Yoav Shoham. A propositional modal logic of time intervals. *Journal of the ACM (JACM)*, 38(4): 935–962, 1991. 3
- [28] Daoji Huang, Otmar Hilliges, Luc Van Gool, and Xi Wang. Palm: Predicting actions through language models@ ego4d long-term action anticipation challenge 2023. *arXiv preprint arXiv:2306.16545*, 2023. 3, 8
- [29] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3
- [30] Qiuhong Ke, Mario Fritz, and Bernt Schiele. Time-conditioned action anticipation in one shot. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9925–9934, 2019. 2, 3
- [31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3, 6
- [32] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23119–23129, 2023. 2
- [33] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, Yiming Chang, and Kai Wang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3
- [34] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Fei Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 3
- [35] Yanguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 3
- [36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3
- [37] Jiasen Lu, Dhruv Batra, and Devi Parikh. Vibert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 3
- [38] Victoria Manousaki, Konstantinos Bacharidis, Konstantinos Papoutsakis, and Antonis Argyros. Vlmah: Visual-linguistic modeling of action history for effective action anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1917–1927, 2023. 2, 3
- [39] Esteve Valls Mascaro, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action forecasting@ ego4d challenge 2022. *arXiv preprint arXiv:2207.12080*, 2022. 7
- [40] Esteve Valls Mascaró, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6048–6057, 2023. 3
- [41] Himangi Mittal, Pedro Morgado, Unnat Jain, and Abhinav Gupta. Learning state-aware visual representations from audible interactions. *Advances in Neural Information Processing Systems*, 35:23765–23779, 2022. 3, 7
- [42] Angelo Montanari. *Metric and layered temporal logic for time granularity*. University of Amsterdam, 1996. 3
- [43] Nada Osman, Guglielmo Camporese, Pasquale Coscia, and Lamberto Ballan. Slowfast rolling-unrolling Istm for action anticipation in egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3437–3445, 2021. 2, 3
- [44] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 3
- [45] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 8
- [46] Amir Pnueli. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pages 46–57. IEEE, 1977. 3
- [47] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. Self-regulated learning for egocentric video activity anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3
- [48] 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. *arXiv preprint arXiv:2103.00020*, 2021. 3
- [49] Mamshad Nayeem Rizve, Gaurav Mittal, Ye Yu, Matthew Hall, Sandra Sajeev, Mubarak Shah, and Mei Chen. Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22992–23002, 2023. 3
- [50] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. Untrimmed action anticipation. In *International Conference on Image Analysis and Processing*, pages 337–348. Springer, 2022. 3

- [51] Debaditya Roy and Basura Fernando. Predicting the next action by modeling the abstract goal. *arXiv preprint arXiv:2209.05044*, 2022. 2, 3
- [52] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction visual transformer for egocentric action anticipation. *arXiv preprint arXiv:2211.14154*, 2022. 2, 3, 7
- [53] Fadime Sener, Dipika Singhania, and Angela 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*, pages 154–171. Springer, 2020. 7
- [54] Yuge Shi, Basura Fernando, and Richard Hartley. Action anticipation with rbf kernelized feature mapping rnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 301–317, 2018. 2, 3
- [55] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 3
- [56] Tsung-Ming Tai, Giuseppe Fiameni, Cheng-Kuang Lee, Simon See, and Oswald Lanz. Unified recurrence modeling for video action anticipation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3273–3279. IEEE, 2022. 2, 3
- [57] Hui Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 3
- [58] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022. 2
- [59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [60] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 98–106, 2016. 3
- [61] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023. 2
- [62] Lan Wang, Gaurav Mittal, Sandra Sajeev, Ye Yu, Matthew Hall, Vishnu Naresh Boddeti, and Mei Chen. Protégé: Untrimmed pretraining for video temporal grounding by video temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6575–6585, 2023. 3
- [63] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 3
- [64] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023. 3
- [65] 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*, pages 13587–13597, 2022. 2, 3, 7
- [66] Ziwei Xu, Yogesh Rawat, Yongkang Wong, Mohan S Kankanhalli, and Mubarak Shah. Don’t pour cereal into coffee: Differentiable temporal logic for temporal action segmentation. *Advances in Neural Information Processing Systems*, 35:14890–14903, 2022. 3
- [67] Ce Zhang, Changcheng Fu, Shijie Wang, Nakul Agarwal, Kwonjoon Lee, Chiho Choi, and Chen Sun. Object-centric video representation for long-term action anticipation. *arXiv preprint arXiv:2311.00180*, 2023. 3
- [68] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3, 6, 7
- [69] Qi Zhao, Ce Zhang, Shijie Wang, Changcheng Fu, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? *arXiv preprint arXiv:2307.16368*, 2023. 3, 7, 8
- [70] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6068–6077, 2023. 2, 3, 7