Summarize the Past to Predict the Future: Natural Language Descriptions of Context Boost Multimodal Object Interaction Anticipation

Razvan-George Pasca[1*] Alexey Gavryushin[1*] Muhammad Hamza[2]
Yen-Ling Kuo[3] Kaichun Mo[4] Luc Van Gool[1, 5, 6] Otmar Hilliges[1] Xi Wang[1]
[1] ETH Zurich [2] Univ. of Zurich [3] Univ. of Virginia [4] NVIDIA [5] KU Leuven [6] INSAIT, Sofia

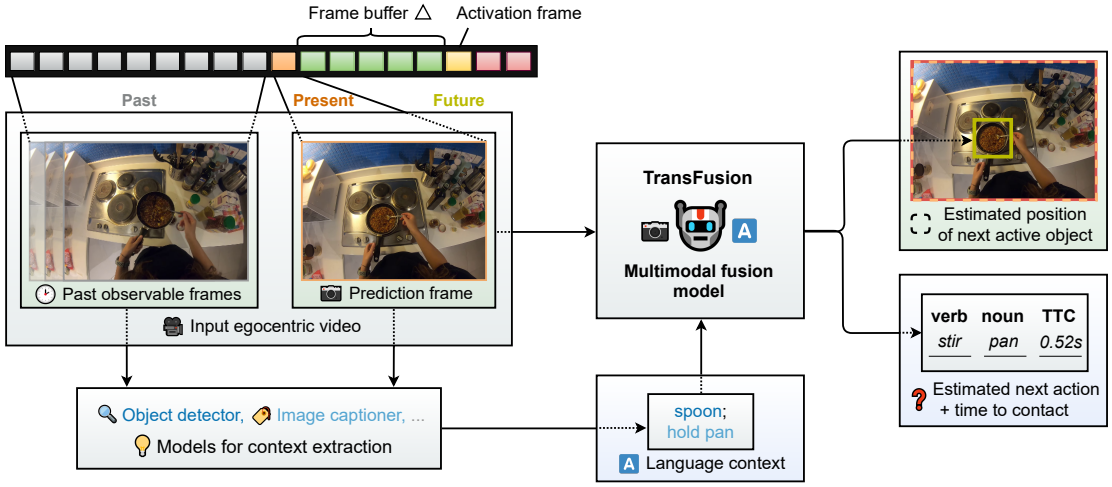


Figure 1. **TransFusion: Multimodal fusion transformer for short-term object interaction anticipation.** Given a video sequence of past observations, the task aims to predict a set of objects visible in the current frame that will be interacted with in the future, i.e. in the activation frame that is Δ frames away from the current prediction frame. Additionally, the task requires estimating the bounding box, the associated action described by a verb-noun pair, and the time to contact for each predicted object. We propose TransFusion, a multimodal fusion architecture that uses language summaries of past actions and salient objects to effectively predict future object interactions.

Abstract

We study object interaction anticipation in egocentric videos. This task requires an understanding of the spatio-temporal context formed by past actions on objects, coined action context. *We propose TransFusion, a multimodal transformer-based architecture for short-term object interaction anticipation. Our method exploits the representational power of language by summarizing the action context textually, after leveraging pre-trained vision-language foundation models to extract the action context from past video frames. The summarized action context and the last observed video frame are processed by the multimodal fusion module to forecast the next object interaction. Experiments on the Ego4D next active object interaction dataset show the effectiveness of our multimodal fusion model and highlight the benefits of using the power of foundation models and language-based context summaries in a task where vision may appear to suffice. Our novel approach outperforms all state-of-the-art methods on both versions of the Ego4D dataset. A project video and code are available at* https://eth-ait.github.io/transfusion-proj/.

1. Introduction

The ability to predict plausible future human-object interactions is important for many assistive technologies. The task of *short-term object interaction anticipation* [18, 22] is defined as predicting *which* object a user will interact with next, *what* action will be performed and *when* in the future, given an egocentric video input. Providing an effective solution would help artificial agents, such as care robots, assist humans in their daily activities, as well as enable safety applications such as child and elderly monitoring.

This task is challenging because it requires reasoning about user intent [6, 35, 43], which is not directly observable. For example, if the video shows the user holding

[*] Authors contributed equally.

a vegetable, the most likely next action is to slice it, but only with a knife and cutting board present. Without those, there is significantly more uncertainty. Therefore, understanding what has happened in the past and what relevant objects are present is critical for predicting the next object interaction. We collectively refer to these concepts as *action context*. Existing interaction forecasting methods do not explicitly model action context and rely on neural networks to extract that information from fixed-sized video chunks [5, 18, 22, 32].

In this paper, we propose to use language summarization of the past as an explicit representation of action context. This filters out ambiguities caused by visual clutter. It also puts focus on objects and actions directly relevant to the ongoing and likely future activities. For the object interaction anticipation, we next introduce a multimodal fusion approach. Specifically, we propose a transformer-based multimodal fusion model, *TransFusion*. As shown in Figure 1, our model takes language summaries of the action context and the last observed single frame of a video as input and predicts future object interactions. We employ an existing image captioning system [53] to generate a concise description of past actions given by pairs of verbs and nouns. We use CLIP [38], a vision-language model trained on pairs of images and captions, to detect salient objects in the prediction frame of interest, and deem these objects task-relevant. The verb-noun pairs together with the list of salient objects form the summary of the past action context. Leveraging the vast knowledge contained in pre-trained vision-language foundation models yields more generic representations of the past action context and enables us to better generalize towards various scenarios, in particular rare ones not well-represented in the dataset.

Our experiments show that our language-based summaries can effectively represent the past action context. We evaluate on both versions of the challenging egocentric dataset Ego4D [22]. Our approach improves interaction prediction accuracy and outperforms state-of-the-art models pre-trained on large video datasets [7, 40]. Such improvements show that large vision-language models can be useful even for a task that seems purely visual.

In summary, our contributions are:

- A strategy for generating concise language-based summaries of action context over longer timespans from frame-wise language-based summaries, for instance obtained using vision-language foundation models.
- TransFusion, a multimodal fusion model for short-term object interaction anticipation combining action context descriptions and visual features.
- Experiments on both versions of the Ego4D dataset showing improvements over state-of-the-art methods on the short-term object interaction anticipation task, in particular for long-tailed classes.

2. Related work

Object interaction anticipation. Several works have been proposed for human-object interaction prediction. Furnari et al. [18] first introduced *next-active-object* (NAO) prediction, which predicts the active/not active label for each object using its motion trajectory features, but doesn't consider action anticipation. Closely related to NAO, Bertasius et al. [5] introduced *action-objects*, which considers objects that capture the human actor's visual attention or tactile interaction. Bertasius et al. use the data from a stereo camera system in a two-stream network, integrating RGB and depth information. They show that certain aspects of human actions can be predicted by exploiting the spatial configuration of the objects and the actor's head. Their approach is not easily scalable due to the need for stereo data.

In addition to different tasks, Liu et al. [32] propose a model that improves NAO prediction by incorporating future hand trajectory modeling. They obtain impressive results using 3D CNNs, confirming that hand motion cues and scene dynamics have important explanatory power in predicting the short-term future. Nonetheless, their approach also encounters scaling limitations because additional ground-truth labels for hand trajectories are needed.

The Ego4D dataset [22] proposed *object interaction anticipation*. Apart from locating the NAO, the task in Ego4D also requires the prediction of associated nouns and verbs, as well as the time to contact (TTC) for the future interacted objects. The official Ego4D baseline uses the current frame and past video frames as input and employs a two-stage approach to first learn noun and box prediction, and afterwards learn verb and TTC prediction. InternVideo [7] similarly leverages a DINO [62] model to detect objects in the first stage, and a VideoMAE [49]-pretrained ViT [14] spacetime attention model to predict verbs and TTC conditioned on the object predictions in the second stage. StillFast [40] presents an end-to-end method for NAO prediction, fusing multi-scale high-resolution image and low-resolution video features and combining local RoI features with global scene features to boost the interaction anticipation performance. GANOv2 [46, 47] and its predecessor [48] extract features from objects detected in past observed frames and fuse them with global frame features, forwarding them to a StillFast predictor and a motion-object-separating encoder-decoder architecture, respectively.

Action anticipation. Besides localizing future object interactions, predicting the next action steps is an equally important task that intelligent systems need to perform to interact with humans. A longer prediction timespan requires a better grasp of the structure behind the succession of actions. A pioneering advancement in this field is the introduction of the EPIC-Kitchens 100 (EK100) [9] dataset, which facilitates research in action anticipation within egocentric

videos. Significant research efforts have been dedicated to exploring the action anticipation task utilizing EK100. Anticipative Video Transformer (AVT) [20] is one of the best-performing architectures on the EK100 [9] action anticipation benchmark. AVT uses a pre-trained Vision Transformer [14] as the image feature backbone and attends to the previously observed video frames to anticipate future actions. MeMViT [54] improves AVT by exploiting a longer temporal context, highlighting the importance of modeling a long sequence of previous actions. Other model architectures such as recurrent neural networks [17, 55], multimodal temporal CNNs [25] and transformers [19, 51] are also used to model the temporal contexts. Note that these methods are designed for action classification using video data and adapting them for object detection is not trivial. It is worth noting that the action anticipation task from the EK100 benchmark is a subtask of our short-term object interaction anticipation task. Our task not only predicts the action label of verb-noun pairs, but we also predict the location of the object and the time to contact, which is more challenging than the action anticipation task itself. TransFusion focuses on detecting the next active objects, as all evaluation metrics are conditioned on the box IoU. Therefore, comparing our work with the state-of-the-art methods for the action anticipation task on EK100 is out of the scope of this work. The Ego4D dataset itself further introduces a long-term action anticipation benchmark, which [61] tackles using pre-trained visual grounding models.

Vision-language models. Our approach combines the visual and language modalities to improve object interaction anticipation. Most recent multimodal vision-language architectures are based on Transformer [50] fusion schemes applied on features extracted by various encoders [4, 12, 26]. We follow their spirit by employing self-attention-based modality fusion. Our work differs in one significant aspect: the language features represent the past action context. This serves as an additional signal to disambiguate the actor's intent given similar visual scenes. The model must learn to predict from the multimodal fusion of language context summaries and visual cues in a considerably smaller data regime.

More broadly, there are numerous transformer-based works aimed at learning cross-modal representations between vision and language [24, 28, 36, 38, 44, 45, 58, 59, 64] using terabyte-scale datasets crawled from multiple internet sources. When combined with large language models such as GPT-2 [37], BERT [13], Sentence-BERT [41], and the T5 variants [8, 39], these vision-language models can be applied to several downstream tasks such as VQA [3] and image captioning [56, 57]. Similar to other vision-language models that are trained with language input, our approach leverages state-of-the-art image captioning models [53] to provide a consistent description of past actions.

3. Language-guided prediction

We use language-based context summaries to represent the spatial-temporal context formed by past actions and object relationships. The context summary offers an explicit yet compact representation, which allows us to focus on task-relevant content and filter out unimportant information from the likely cluttered environment. To this end, we propose TransFusion, a transformer-based model that leverages a joint-attention fusion mechanism to effectively learn to predict object interactions in the future from the past action context. In the following, we describe our framework, which takes video clips as input and anticipates the next object interaction. We begin by defining the object interaction anticipation task (Sec. 3.1). We then show how to generate a summary for a prediction frame from an egocentric video to represent the action context provided by the past frames (Sec. 3.2). Finally, we present our multimodal fusion model's architecture (Sec. 3.3).

3.1. Task definition

Given a video $\mathcal{V} = \{f_1, \cdots, f_T\}$ of length T as input, our goal is to predict the next object interaction $\mathcal{O}$ in the last observed video frame f_T, possibly using older frames to extract context information. We call f_T the *prediction frame*.

As illustrated in Figure 1, the task is to anticipate the object interaction in f_T. $\mathcal{O}_T = \{o_1^T, \cdots, o_N^T\}$ denotes the set of future object interactions, with $N > 1$ whenever multiple objects are interacted with simultaneously. In other words, multiple o_i^T can be detected in the prediction frame f_T. For example, $N = 2$ when one hand is holding two mugs in a kitchen scene, or the left hand is touching a plant while the right hand is moving a pot in a garden scene. Each object interaction $o_i^T = (b_i^T, n_i^T, v_i^T, t_i^T)$ consists of the object bounding box b_i^T, the semantic object label (noun) n_i^T, the action label (verb) v_i^T, and the time to contact (TTC) t_i^T. The task is to predict the next active object location b_i^T in the prediction frame f_T, the associated action described by the verb-noun pair (n_i^T, v_i^T), and the time t_i^T to the point of contact when the interaction is going to start. The task is set up such that the actual object interaction takes place after a buffer time Δ on frame $f_{T+\Delta}$ and the predicted time to contact $t_i^T = \Delta$. Ego4D uses $\Delta \geq 0.033s$. This requires anticipating future actions and increases the task difficulty.

3.2. Summarizing the past

Anticipating object interactions in the future requires an understanding of what happened in the past. We first infer the per-frame *action context* from each egocentric video. This process consists of extracting *action descriptions* and *held objects*. These information sources are generated and aggregated independently across frames and finally forwarded as the action context to our anticipation model.

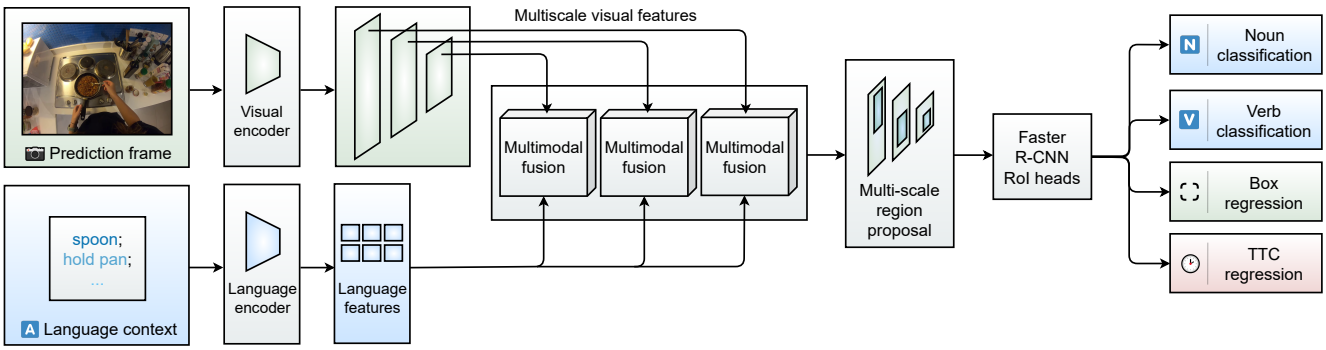


Figure 2. **Overview of the TransFusion model.** TransFusion takes the prediction frame and the action context summary as input and predicts the bounding box of the next-active object, the noun-verb pair describing the associated action, and the time to contact (TTC). Feature maps of different scales are extracted from the visual encoder and then fused via a multimodal fusion module with the encoded language features. Their output is then processed by a regular feature pyramid network (FPN), denoted as multi-scale region proposal networks, before being fed into the Faster R-CNN detector.

Extracting frame-wise *action descriptions*. Inspired by the label format used in egocentric video datasets [10, 22], we aim to describe the past actions by sequences of verb-noun pairs (e.g. "wash tomato", "cut tomato"), so-called *action descriptions*. For each video frame f, we extract the action currently performed by the agent, represented by a single pair (v, n) of a verb v and a noun n.

We use an off-the-shelf image captioning model [53] to generate *multiple* full-sentence captions for each frame of interest, so as to find repeatedly occurring outputs and reduce noise: we perform part-of-speech tagging [2] on these captions and collect *candidate pairs* consisting of verbs followed by nouns within some cutoff distance $d \in \mathbb{N}_0^+$. Setting $d > 0$ allows us to account for additional words appearing between a verb-noun pair (e.g. "*eat apple*" from "A person *eating* a red *apple*") and thereby obtain more valid candidate pairs from the captions. However, larger cutoff distances may introduce more spurious detections (e.g. "*eat gathering*", from "A person *eating* while at a *gathering*"). We use $d = 4$ in our experiments. The most frequent candidate pair across the frame's captions is selected as the verb-noun action description of this frame. If no valid verb-noun pair is found, the frame will have no action description.

Extracting frame-wise *salient objects*. Another characteristic shared by many daily action sequences is that the next active object is likely to have already appeared in the actor's view before the upcoming interaction. The current environment in the recording may also provide model conditioning. Therefore, we select a set of *salient objects* in each frame as (1) the potential candidates for future active objects and (2) a proxy of the surroundings.

Encouraged by its previous successful use in action-related tasks [30, 52], we use CLIP [38], a pre-trained open-vocabulary zero-shot classifier, to compute the cosine similarities between the CLIP image embedding of a frame and the embeddings of the object categories from the Ego4D domain. We choose to keep the highest-scoring k objects as the set of salient objects for a given frame. A small k may fail to provide sufficient information, while a large k makes capturing the true salient objects more likely, yet introduces more noise due to false detections. We set k=3.

Aggregating summaries over multiple frames. After the frame-wise summarization step, we have, for each frame t, at most one verb-noun pair describing the frame-wise action $a_t = (n, v) \in \mathcal{A}_t^F$. We then apply a *cross-frame aggregation* scheme to all $\mathcal{A}_t^F$ so as to identify contiguous segments of frames exhibiting the same activity, while accounting for individual noisy frames possibly not matching the frame-wise summaries of their neighbors. The aggregation processes the frames in temporal order. We consider frames belonging to the same atomic action as an action segment, which is initiated by observing a number of reoccurrences of some verb-noun pair and terminated by a lack of its occurrence within a number of contiguous frames. After the aggregation, we obtain $\mathcal{A}$, the sequence of all action segments in the video. For each prediction frame, we construct its action context from a number of action segments preceding the frame, and the objects from $\mathcal{N}_s$ in that frame.

By aggregating across frames, we thus capture dominant actions in the past, reduce noise in the generated action context and represent long timespans of similar activity by short textual descriptions. The complete action context is provided to our model by concatenating the textual representations of the aggregation results of any desired subset of the two context representations $\{\mathcal{A}, \mathcal{N}_s\}$, where we drop the index t for convenience. For more details, see the Supp. Mat.

3.3. TransFusion: Multimodal fusion

Figure 2 visualizes the architecture of the proposed multimodal TransFusion model. TransFusion employs two different base encoders, for language and image input respectively. The input prediction frame f_T is processed by a regu-

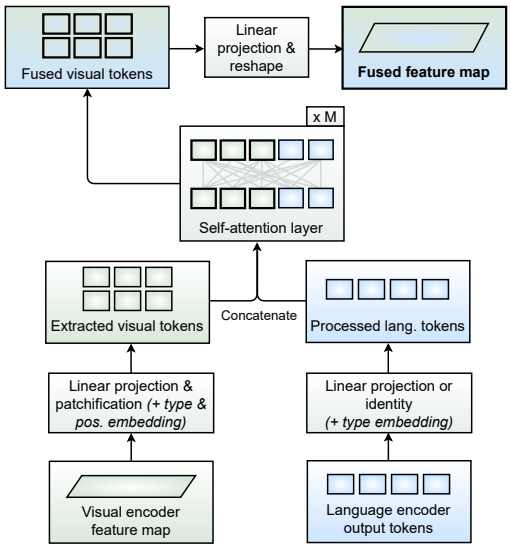


Figure 3. **Multimodal fusion module.** We first project the CNN feature map and the language tokens to a common dimensionality before adding the specific embeddings. We concatenate the visual and language tokens and feed them to M self-attention layers. At the output, the fused visual tokens are projected back into the initial feature map shape.

lar CNN-based visual encoder [23], producing a set of multiscale feature maps $\mathcal{F}^s_{\mathcal{V}}(f_T)$, with s indicating the scale. The frame's action context is encoded by a language encoder [41] and yields $\mathcal{F}_{\mathcal{L}}(\mathcal{C}_L)$. The feature maps belonging to different scale levels from the visual encoder are fused with the encoded language features via a joint-attention scheme. The result is then processed by a regular feature pyramid network (FPN) [29] before being forwarded to the Faster R-CNN [42] detector.

Multimodal fusion. For simplicity, we describe the multimodal fusion module for a single input scale s, which we hence omit from notation. See Figure 3 for a detailed view of a single Transformer encoder layer together with the input projection stages. For the visual modality, we first split the image into a sequence of tokens. Specifically, the visual features $\mathcal{F}_{\mathcal{V}}(f_T) \in \mathbb{R}^{C\times H\times W}$ are reshaped to $\mathbb{R}^{N\times(P^2\cdot C)}$, where P is the token size and $N = \frac{H\times W}{P^2}$ the number of tokens. We add unique type embeddings as well as a positional encoding to each modality. Both visual and language tokens are first mapped to $\mathbb{R}^D$ via a linear projection before being fed to the fusion module in a concatenated form. The fusion block consists of a regular Transformer encoder layer replicated M times. Finally, the visual tokens are regrouped into their initial shape to compose the feature map for the feature pyramid network (FPN).

Multiscale fusion. The multimodal fusion block outputs multiple fused feature maps of varying scales: $\mathcal{F}_{fi} \in \mathbb{R}^{C_i\times H_i\times W_i}$, where i is the scale index. They are fed into the FPN, which has different parameters for each feature map, allowing the network to learn specialized features for each downsampling ratio. The output of the FPN is processed by a Faster R-CNN detector.

Prediction head networks. The Faster R-CNN detector outputs the bounding box b through a regression layer, an associated object n and verb v, and time to contact t.

Training objective. We train TransFusion using the standard Faster R-CNN [42] objective $\mathcal{L}_{box}$ for bounding box prediction, which consists of localization and objectness terms; further, cross-entropy losses $\mathcal{L}_{noun}$ and $\mathcal{L}_{verb}$ for noun and verb prediction respectively, and an L1 loss $\mathcal{L}_{ttc}$ for time to contact prediction. The overall objective is

$$\mathcal{L} = \mathcal{L}_{box} + \mathcal{L}_{noun} + \mathcal{L}_{verb} + \mathcal{L}_{ttc}. \tag{1}$$

As defined in Faster R-CNN, $\mathcal{L}_{box}$ is

$$\mathcal{L}_{box} = \frac{1}{N_{\text{cls}}}\sum_i \mathcal{L}_{\text{cls}}(p_i, p_i^*) + \frac{\lambda}{N_{\text{reg}}}\sum_i p_i^* \cdot L_{\text{reg}}(b_i - b_i^*), \tag{2}$$

where p_i is the predicted probability of a box being classified as a foreground object (true positive), p_i^* is the ground-truth label; b_i and b_i^* are the predicted and ground-truth box coordinates. N_{reg} is the number of bounding boxes used in training, N_{cls} the detection network's image batch size. We leave $\lambda = 11$ as in the reference implementation.

4. Experiments

Dataset. We use both versions (v1 and v2) of the dataset provided for the short-term object interaction anticipation task in **Ego4D** [22]. It contains various activities ranging from gardening work to cooking or car parts changing. There are in total 64,798 (v2: 165,451) annotated video clips of $\sim$ 5 minutes each with frame rate $fps = 30$. The dataset has a highly imbalanced long tail distribution, including 87 nouns and 74 verbs (v2: 128 and 81 resp.). For instance, the "take" class encompasses almost 40% of all the verb labels in version 1, while "mold" appears only in one video. The train set consists of about 28k (v2: 98k) samples, while the rest are roughly equally split between validation and test.

Evaluation. As in the official Ego4D benchmark [22], we use the Top-5 mean Average Precision (mAP) to evaluate the performance of individual predictions. It considers two constraints: The **IoU** constraint requires that the predicted boxes are counted as hits only if there is an intersection over union IoU ≥ 0.5 with the ground-truth box. The **TTC** constraint requires that the predicted time-to-contact t_i is close enough to the ground-truth $\hat{t}_i$, i.e. $|t_i - \hat{t}_i| < T_\delta, T_\delta = 0.25$. The following metrics are considered:

- **Noun** refers to the Noun-Box Top-5 mAP. In addition to the IoU constraint, a correct match with the ground-truth noun is required.

Ver.	Model	N ↑	N-V ↑	N-T ↑	A ↑
v1	FRCNN+Rnd. [22]	20.45	2.22	3.86	0.44
	FRCNN+SF [22]	20.45	6.78	6.17	2.45
	InternVideo [7]	24.60	9.18	**7.64**	3.40
	StillFast [40]	19.51	9.95	6.45	3.49
	TF (ours)	**24.90**	**10.06**	7.57	**3.54**
v2	FRCNN+SF [22]	26.15	9.45	8.69	3.61
	StillFast [40]	25.06	13.29	9.14	5.12
	GANOv2 [46]	25.67	**13.60**	9.02	5.16
	TF (ours)	**30.43**	13.45	**10.38**	**5.18**

Table 1. **Comparison with the state-of-the-art.** Our method TransFusion (TF) outperforms all state-of-the-art approaches on 3 out of 4 metrics, on both versions of the Ego4D test set. We report the top-5 mean average precision for the Noun (N), Noun-Verb (N-V), Noun-TTC (N-T), and Noun-Verb-TTC (A) metrics.

- **Noun-Verb** refers to the Noun-Verb-Box Top-5 mAP. Correct noun and verb match with the IoU constraint.
- **Noun-TTC** refers to the Noun-Box Top-5 mAP. Correct noun matches with the TTC constraint.
- **Noun-Verb-TTC** ***("Overall")*** refers to the overall Top-5 mAP, intersecting correct noun-verb matches with IoU and TTC constraints.

Note that since detecting the next-active objects is the main focus of the task, all evaluation metrics are conditioned on the IoU constraint.

Implementation details. We make use of the pre-trained Torchvision [34] Faster R-CNN. ResNet-50 [23] is used as the visual encoder and Sentence-BERT (SBERT) [41] is used as the language encoder. We freeze ResNet-50 and apply the multimodal fusion scheme on top. We use the RAdam [31] optimizer with a weight decay of $2e^{-4}$, and an effective batch size of 32. Besides the default Faster R-CNN background class in nouns, we also add it in the verb class, which improves the Noun-Verb mAP by almost 2 points. We reduce the region proposal network's sampling batch size and the detection network's image batch size and use data augmentation techniques to reduce overfitting. The number of verb-noun pairs used in the input is defined as the context length L_c. On average, an action segment described by one verb-noun phrase corresponds to one second with 30 frames, as computed from the GT annotations. Unless specified otherwise, we use L_c=3. We perform postprocessing involving non-maximum suppression on the predicted bounding boxes, as elaborated in the Supp. Mat.

4.1. Comparison with state-of-the-art

We first compare our method against the baselines proposed in the Ego4D work [22]. The **FRCNN + SF** method employs a two-stage approach: a ResNet-based Faster R-CNN detector first produces bounding box regressions and noun predictions from the prediction frame, without using any video features. In the second stage, a SlowFast [16] 3D

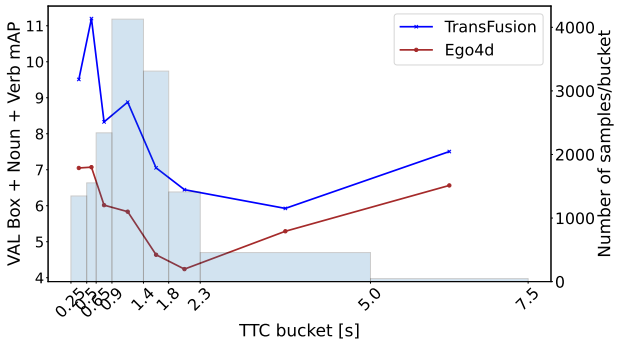


Figure 4. **Prediction of Noun-Verb in relation to time-to-contact.** Histogram of the time-to-contact labels in Ego4Dv1 is shown in blue bars. Performance measured in Noun-Verb is plotted as a function of time-to-contact. We compare our method to Ego4D's official FRCNN + SF [22] baseline.

CNN video processing module performs region-of-interest (RoI) pooling [21] on the corresponding SlowFast 3D CNN video features. The pooled video features are then used by RoI heads to predict the verb and TTC values. The **FRCNN + Rnd.** baseline, with results published on v1 only, drops the second stage of FRCNN + SF and uses randomly drawn verb and TTC predictions, following their respective distributions in the training set.

Another baseline on v1 is given by InternVideo [7], a method achieving strong results on the short-term object interaction anticipation task with a two-stage strategy similar to FRCNN+SF: bounding boxes and noun labels are first obtained using a DINO–based object detector [62], after which a VideoMAE-pretrained [49] ViT [14] model leverages space-time attention to predict the corresponding verb classes and TTC values.

We further compare TransFusion to the recent StillFast [40] approach on both dataset versions. StillFast uses a Faster-RCNN–based spatial stream operating on a high-resolution version of the input image, and an X3D–based [15] temporal stream operating on a low-resolution version of the input video. The two streams each produce a multiscale feature map. Both feature maps are fused into a combined feature pyramid, from which local RoI features and global scene features are extracted, merged by a fusion network, and finally used by linear layers to predict each boxes' noun, verb and TTC.

GANOv2 [46] enhances StillFast by adding a multi-head guided attention mechanism to its temporal stream, achieving favorable performance on version 2 of the dataset. The mechanism attends to object detections produced from the network's temporal branch, together with patches from the 3-D temporal feature stack, in a cross-fusion manner, and uses the fused tokens as the input to StillFast's feature pyramid and subsequent architecture.

Table 1 demonstrates that TransFusion outperforms all state-of-the-art methods in terms of the N, N-V and A met-

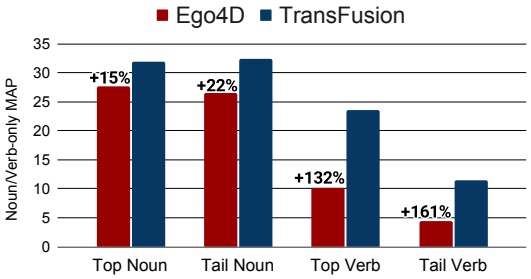


Figure 5. **Classification performance on top/tail categories.** We show the relative and absolute gains of TransFusion over Ego4D's FRCNN + SF baseline [22] on the validation set, for Noun-Only and Verb-Only mAP (without any box IoU constraint). Relative improvements are written on top of the red bars.

rics on version 1 of the dataset, and approaches the state-of-the art performance on the N-T metric while requiring a fraction of the competitor's resources during training [7] (see the Supp. Mat. for details). On the dataset's second version, TransFusion performs best in terms of N, N-T and A, while approaching the runner-up method on N-V. These favorable results validate the advantages of modeling action context explicitly for the interaction anticipation task.

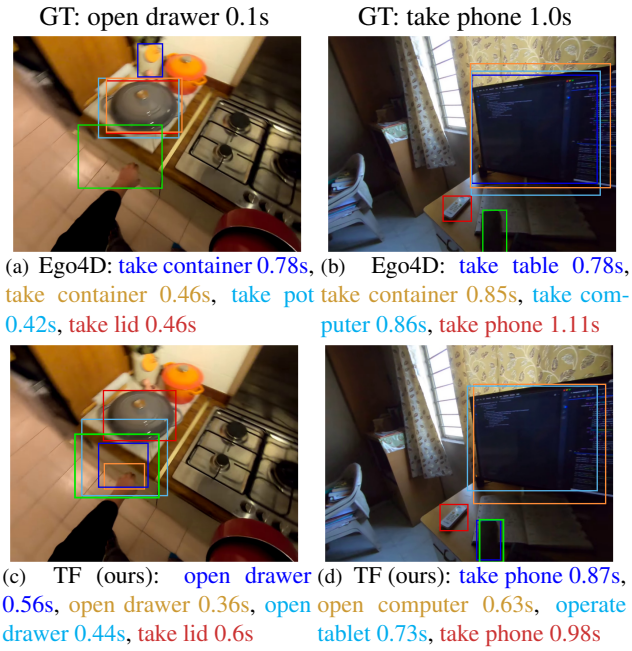
We henceforth analyze our performance gains over the official Ego4D FRCNN + SF baseline on the first version of the Ego4D dataset. Figure 4 shows the long tail distribution of the time-to-contact labels in the validation set, split into buckets. We see that TransFusion consistently outperforms the baseline on the N-V metric for all TTC ranges. The results in Figure 5 indicate that our method not only outperforms the Ego4D method on top noun and verb classes, but also achieves bigger gains in the tail classes, demonstrating the generalization ability of using our language-based context summaries. Top and tail buckets represent roughly 50% of the samples in the validation set.

Figure 6 shows that the Ego4D approach manages to detect the most salient objects in the environment, but fails to reliably associate the bounding boxes to the correct action phrase. In contrast, TransFusion is able to correctly associate the drawer bounding box to "open drawer" and to predict the correct action associated with the phone. More examples are provided in the Supp. Mat.

4.2. Ablation studies

Evaluation of action context representation. In Table 2, we ablate different variants of context representation and compare against the representation that uses the ground-truth (GT) action labels provided by Ego4D in the action context, to assess the quality of our generated context summaries. The observed TTC Average Precision difference measures around 2% and was thus omitted from the comparison. We report on the Noun and Noun-Verb metrics as we expect major differences in semantic-related tasks.

Experiments show that we achieve results approaching GT, demonstrating the effectiveness of our context repre-



(a) Ego4D: take container 0.78s, take container 0.46s, take pot 0.42s, take lid 0.46s (b) Ego4D: take table 0.78s, take container 0.85s, take computer 0.86s, take phone 1.11s

(c) TF (ours): open drawer 0.56s, open drawer 0.36s, open drawer 0.44s, take lid 0.6s (d) TF (ours): take phone 0.87s, open computer 0.63s, operate tablet 0.73s, take phone 0.98s

Figure 6. **Qualitative results of the proposed approach.** Results of Ego4D's FRCNN + SF baseline [22] are shown in the top row and TransFusion in the bottom row. Green denotes the ground truth bounding box, with the labels shown on top of each column.

Rep.	$\emptyset$	$\mathcal{A}$	$\mathcal{N}_h$	$\mathcal{N}_s$	$\mathcal{A}$+$\mathcal{N}_h$	$\mathcal{A}$+$\mathcal{N}_s$	GT
N ↑	17.71	<u>20.09</u>	19.33	19.75	19.03	**20.19**	21.71
N-V ↑	6.14	<u>7.16</u>	6.86	7.05	7.12	**7.55**	7.86

Table 2. **Comparison of different action context representations.** We evaluate different variations of context representations on the predicted noun (N) and noun-verb (N-V), and compare them against ground-truth annotations (GT) provided by v1 of the Ego4D validation set. GT only contains action descriptions of verb-noun pairs. Action context represented by $\mathcal{A}$+$\mathcal{N}_s$ achieves the best N-V performance.

sentations. Using a vision-only version of our model ($\emptyset$) yields markedly worse results than when adding any type of language conditioning. Action descriptions $\mathcal{A}$ consistently boost the verb classification performance. Adding salient objects $\mathcal{N}_s$ to $\mathcal{A}$ further improves task performance. These together offer a reasonable picture of using language summaries to represent the action context. In addition, we notice a domain gap in the choice of words between the generated labels and the ground-truth ones. In some cases, even if the description is satisfactory for a human evaluator, the information may be too vague to be directly useful to the model. Phrasing differences between the generated captions and the training corpus of the language encoders may further help to improve the performance.

As an additional type of action context, we hypothesize that the sequence of objects held by the actor may be indicative of the overall task and can help the model better infer the immediate next steps. Motivated by this, we use an existing hand-object interaction predictor [11] jointly with an

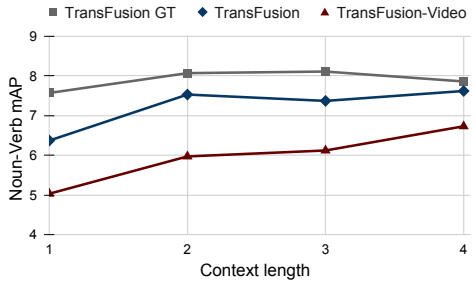


Figure 7. **Comparison to video features.** TransFusion outperforms TransFusion-Video over different context lengths measured by Noun-Verb mAP. TransFusion using GT-based context summaries as language input represents an upper performance bound.

object detector [65] to look for held objects and their semantic labels in each frame. Usually, such objects occur only in a subset of frames (e.g. due to limited visibility of the actor's hands), which results in a lower performance of the resulting $\mathcal{N}_h$ and $\mathcal{A}+\mathcal{N}_h$ representations. We hence use $\mathcal{A}$+$\mathcal{N}_s$ to represent the action context, considering the good performance achieved. Ablations on the hyperparameters d, the cutoff distance, and k, the number of salient objects per frame, show that the performance is less sensitive to their settings. More details can be found in the Supp. Mat.

Comparison to video features. We evaluate whether language summaries can better represent action context than video features for the object-anticipation task. To do so, we implement **TransFusion-Video**, which takes video features extracted from the Ego4D SlowFast [16] model as input and uses an identical fusion module as in TransFusion.

Figure 7 shows that TransFusion consistently outperforms TransFusion-Video. $L_c \times 30$ frames are taken in TransFusion-Video. For video features, the domain gap between pre-training and target datasets seems to be too large to allow effective generalization to a diverse dataset such as Ego4D. We suspect that the feature representations are more entangled due to the need to represent temporal, visual, and semantic aspects simultaneously. With language, temporal information is directly encoded in the syntax, i.e. the succession of words. Given a video clip of one second, language can summarize it in two words whereas corresponding video features use more than 12 times the space (e.g. smaller SBERT [41] features of 2×384 compared to SlowFast [16] features of 4×2304). Therefore, our language-based context representation has an advantage when it comes to longer video sequences. Overall, TransFusion obtains a 21% boost in N-V performance. See the Supp. Mat. for details.

Length of action context. A significant parameter in obtaining good performance is the length of action context, i.e., how many past actions to include in the input. An adequate context length is important to allow the model to pick up more specific activity patterns. Some activities that have a more structured nature or certain repeating patterns (e.g. cooking or repairing a bicycle) are easier to model with an increased context length compared to more random activities such as playing basketball. Figure 7 shows that the overall performance plateaus at $L_c = 2$ and higher.

Language Encoder	Token size	N ↑	N-V ↑
SBERT/BERT [13, 41]	384	**20.19**	7.55
SBERT/RoBERTa [33, 41]	768	19.78	**7.78**
GPT-2 [37]	768	18.98	7.08
Flan-T5 [8]	1024	18.26	6.61

Table 3. **Language encoder ablations.** We compare SBERT using BERT and RoBERTa backbones, with GPT-2 and Flan-T5.

Language encoder. A language encoder is used to process the context summaries in TransFusion. To understand the impact of different language encoders, we experiment with multiple LLMs including SBERT with a RoBERTa backbone [33, 41], GPT-2 [37], and FLAN-T5 [8], using version 1 of the validation set. Table 3 summarizes the performance of different language encoders, showing that the SBERT models perform best. We consider TTC less relevant here. We hypothesize that the difference comes from the training objectives of language models. SBERT is trained to map semantically similar sentences closer to each other in the embedding space. This is particularly useful when the model needs to understand descriptions with semantically close words, e.g. "cut carrots" and "slice carrots". GPT-2 and FLAN-T5 are optimized using generative objectives, which makes them overly sensitive to the input word choice.

Fusion module. We provide ablations regarding the fusion module in the Supp. Mat.

5. Conclusion and discussion

We propose *TransFusion*, a multimodal fusion model that anticipates object interactions by considering the past action context represented by language summaries. Our successful experiments on a challenging egocentric video dataset demonstrate how language summaries and the vast knowledge contained in vision-language foundation models can improve object interaction anticipation, highlighting the representational power of language. It appears promising to investigate fusing even more modalities, such as motion cues from optical flow. We emphasize that the proposed summarization approach is not limited to this task. Language provides a universal interface to complement the visual input with information encoded in foundation models or task-specific pipelines. Future work can leverage language summarization for other video reasoning tasks [60, 63], or extend it temporally to describe possible future activities [1, 27] for long-term interaction anticipation.

Acknowledgements. This work was supported by an ETH Zurich Postdoctoral Fellowship. We thank Adrian Spurr for the helpful comments and Velko Vechev for the voiceover.

References

[1] Chaitanya Ahuja and Louis-Philippe Morency. Language2Pose: Natural Language Grounded Pose Forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728, 2019. 8

[2] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649, 2018. 4

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 3

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. *CoRR*, abs/2104.00650, 2021. 3

[5] Gedas Bertasius, Hyun Soo Park, Stella X. Yu, and Jianbo Shi. First-Person Action-Object Detection with EgoNet. In *Proceedings of Robotics: Science and Systems*, 2017. 2

[6] Jerome S Bruner. Intention in the Structure of Action and Interaction. *Advances in Infancy Research*, 1981. 1

[7] Guo Chen et al. InternVideo-Ego4D: A Pack of Champion Solutions to Ego4D Challenges. *arXiv preprint arXiv:2211.09529*, 2022. 2, 6, 7

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models, 2022. 3, 8

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling Egocentric Vision. *CoRR*, abs/2006.13256, 2020. 2, 3

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-Kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. 4

[11] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 7

[12] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-End Visual Grounding with Transformers. *CoRR*, abs/2104.08541, 2021. 3

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805, 2018. 3, 8

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, abs/2010.11929, 2020. 2, 3, 6

[15] Christoph Feichtenhofer. X3D: Expanding Architectures for Efficient Video Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020. 6

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. *CoRR*, abs/1812.03982, 2018. 6, 8

[17] Antonino Furnari and Giovanni Maria Farinella. Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. *CoRR*, abs/2005.02190, 2020. 3

[18] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-Active-Object Prediction from Egocentric Videos. *CoRR*, abs/1904.05250, 2019. 1, 2

[19] H. Girase, N. Agarwal, C. Choi, and K. Mangalam. Latency Matters: Real-Time Action Forecasting Transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18759–18769, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 3

[20] Rohit Girdhar and Kristen Grauman. Anticipative Video Transformer. In *ICCV*, 2021. 3

[21] Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015. 6

[22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video. *CoRR*, abs/2110.07058, 2021. 1, 2, 4, 5, 6, 7

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385, 2015. 5, 6

[24] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. *CoRR*, abs/2004.00849, 2020. 3

[25] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. *CoRR*, abs/1908.08498, 2019. 3

[26] Weicheng Kuo, Fred Bertsch, Wei Li, AJ Piergiovanni, Mohammad Saffar, and Anelia Angelova. FindIt: Generalized Localization with Natural Language Queries. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 502–520. Springer, 2022. 3

[27] Yen-Ling Kuo, Xin Huang, Andrei Barbu, Stephen G McGill, Boris Katz, John J Leonard, and Guy Rosman. Trajectory Prediction with Linguistic Representations. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2868–2875. IEEE, 2022. 8

[28] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling. *CoRR*, abs/2102.06183, 2021. 3

[29] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature Pyramid Networks for Object Detection. *CoRR*, abs/1612.03144, 2016. 5

[30] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard De Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen Clip Models Are Efficient Video Learners. In *European Conference on Computer Vision*, pages 388–404. Springer, 2022. 4

[31] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the Variance of the Adaptive Learning Rate and Beyond. *CoRR*, abs/1908.03265, 2019. 6

[32] Miao Liu, Siyu Tang, Yin Li, and James M. Rehg. Forecasting Human Object Interaction: Joint Prediction of Motor Attention and Egocentric Activity. *CoRR*, abs/1911.10967, 2019. 2

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692, 2019. 8

[34] TorchVision maintainers and contributors. TorchVision: PyTorch's Computer Vision Library. https://github.com/pytorch/vision, 2016. 6

[35] David W Orr. *The Nature of Design: Ecology, Culture, and Human Intention*. Oxford University Press, 2002. 1

[36] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. EGOVLPV2: Egocentric Video-Language Pre-training With Fusion in the Backbone, 2023. 3

[37] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 2019. 3, 8

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *CoRR*, abs/2103.00020, 2021. 2, 3, 4

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *CoRR*, abs/1910.10683, 2019. 3

[40] Francesco Ragusa et al. StillFast: An End-to-End Approach for Short-Term Object Interaction Anticipation. *arXiv preprint arXiv:2304.03959*, 2023. 2, 6

[41] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR*, abs/1908.10084, 2019. 3, 5, 6, 8

[42] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *CoRR*, abs/1506.01497, 2015. 5

[43] Paschal Sheeran. Intention-Behavior Relations: A Conceptual and Empirical Review. *European Review of Social Psychology*, 12(1):1–36, 2002. 1

[44] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive Bidirectional Transformer for Temporal Representation Learning. *CoRR*, abs/1906.05743, 2019. 3

[45] Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *CoRR*, abs/1908.07490, 2019. 3

[46] Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Enhancing Next Active Object-Based Egocentric Action Anticipation With Guided Attention. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1450–1454. IEEE, 2023. 2, 6

[47] Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Guided Attention for Next Active Object @ EGO4D STA Challenge. *arXiv preprint arXiv:2305.16066*, 2023. 2

[48] Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Leveraging Next-Active Objects for Context-Aware Anticipation in Egocentric Videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8657–8666, 2024. 2

[49] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training. *Advances in Neural Information Processing Systems*, 35:10078–10093, 2022. 2, 6

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *CoRR*, abs/1706.03762, 2017. 3

[51] Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, and Tong Lu. Memory-and-Anticipation Transformer for Online Action Understanding, 2023. 3

[52] Mengmeng Wang, Jiazheng Xing, and Yong Liu. ActionCLIP: A New Paradigm for Video Action Recognition. *arXiv preprint arXiv:2109.08472*, 2021. 4

[53] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. *CoRR*, abs/2202.03052, 2022. 2, 3, 4

[54] 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. *CoRR*, abs/2201.08383, 2022. 3

[55] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to Anticipate Egocentric Actions by Imagination. *CoRR*, abs/2101.04924, 2021. 3

[56] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation With Visual Attention. In *International Conference on Machine Learning*, pages 2048–2057. PMLR, 2015. 3

[57] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting Image Captioning With Attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4894–4902, 2017. 3

[58] Moon Ye-Bin, Jisoo Kim, Hongyeob Kim, Kilho Son, and Tae-Hyun Oh. TextManiA: Enriching Visual Feature by Text-driven Manifold Augmentation, 2023. 3

[59] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal Neural Script Knowledge Models. *CoRR*, abs/2106.02636, 2021. 3

[60] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging Video Descriptions to Learn Video Question Answering. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 8

[61] Ce Zhang, Changcheng Fu, Shijie Wang, Nakul Agarwal, Kwonjoon Lee, Chiho Choi, and Chen Sun. Object-Centric Video Representation for Long-term Action Anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6751–6761, 2024. 3

[62] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 6

[63] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video Summarization With Long Short-Term Memory. In *European Conference on Computer Vision*, pages 766–782. Springer, 2016. 8

[64] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning Video Representations From Large Language Models, 2022. 3

[65] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple Multi-Dataset Detection. *CoRR*, abs/2102.13086, 2021. 8