Computer Vision and Image Understanding 249 (2024) 104144

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Multi-modal transformer with language modality distillation for early pedestrian action anticipation

Nada Osman, Guglielmo Camporese, Lamberto Ballan [*]

Department of Mathematics "Tullio Levi-Civita", University of Padova, Italy

ARTICLE INFO

Keywords:
Action anticipation
Pedestrian intent prediction
Language-vision models
Image-to-video captioning
Multi-modal transformers

ABSTRACT

Language-vision integration has become an increasingly popular research direction within the computer vision field. In recent years, there has been a growing recognition of the importance of incorporating linguistic information into visual tasks, particularly in domains such as action anticipation. This integration allows anticipation models to leverage textual descriptions to gain deeper contextual understanding, leading to more accurate predictions. In this work, we focus on pedestrian action anticipation, where the objective is the early prediction of pedestrians' future actions in urban environments. Our method relies on a multi-modal transformer model that encodes past observations and produces predictions at different anticipation times, employing a learned mask technique to filter out redundancy in the observed frames. Instead of relying solely on visual cues extracted from images or videos, we explore the impact of integrating textual information in enriching the input modalities of our pedestrian action anticipation model. We investigate various techniques for generating descriptive captions corresponding to input images, aiming to enhance the anticipation performance. Evaluation results on available public benchmarks demonstrate the effectiveness of our method in improving the prediction performance at different anticipation times compared to previous works. Additionally, incorporating the language modality in our anticipation model proved significant improvement, reaching a 29.5% increase in the F1 score at 1-second anticipation and a 16.66% increase at 4-second anticipation. These results underscore the potential of language-vision integration in advancing pedestrian action anticipation in complex urban environments.

1. Introduction

Human action anticipation is an essential task in many applications and is gaining much interest in the computer vision field. In particular, predicting pedestrians' intentions and actions is crucial in the context of autonomous driving and smart surveillance systems. Therefore, a recent research direction focuses on predicting pedestrians' intention to cross or not cross the street (Rasouli et al., 2017; Kotseruba et al., 2020, 2021). Moreover, as early as the anticipation of the human behavior, more proactive planning could be provided upon the predicted future, which opened the door for more interest in early anticipation of pedestrians' actions (Osman et al., 2022, 2023). In this work, we focus on early pedestrian action anticipation, employing a multi-modal transformer-based model that utilizes different input modalities extracted from the observed history. Mainly, we aim to combine visual and language modalities in order to enrich the comprehensive ability of the model to interpret the observed events and anticipate future actions. Notably, language can serve as a beneficial tool in enriching visual content by providing comprehensive contextual descriptions of the observed images and videos in many visual tasks. Textual descriptions can convey supplementary context, providing more details that surpass those derived solely from visual feature extraction. Therefore, with the rapid advancements in natural language processing (NLP), the integration of language and vision is becoming increasingly relevant, offering significant value to various visual tasks. Recent vision-language models such as CLIP (Radford et al., 2021) exemplify the potential of such integration, enhancing visual understanding through textual input.

In the domain of action anticipation, recent models have introduced innovative approaches that incorporate the generation of textual descriptions, seamlessly integrating them into the anticipation models (Manousaki et al., 2023; Zhao et al., 2024; Ghosh et al., 2022; Huang et al., 2023; Das and Ryoo, 2022). However, the current integration of language into the action anticipation task primarily addresses the anticipation of first-person actions, where the observed contexts often share similarities among actions, and language descriptions can offer concise and insightful representations. However, in this work, we

[*] Corresponding author.
E-mail address: lamberto.ballan@unipd.it (L. Ballan).

https://doi.org/10.1016/j.cviu.2024.104144
Received 13 May 2024; Received in revised form 18 July 2024; Accepted 2 September 2024
Available online 10 September 2024

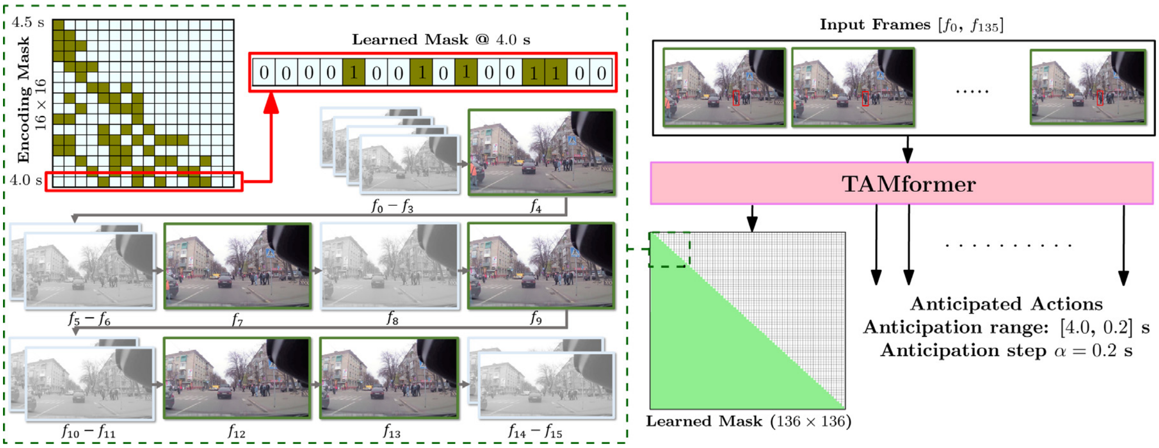


Fig. 1. **TAMformer**: **T**emporal **A**daptive **M**asking Trans**former**. We introduce a multi-modal transformer model that dynamically adjusts its attention to prioritize the most relevant frames within the observed sequence, by learning an attention mask. The figure provides a close-up view of the first half-second of a learned mask by TAMformer, showcasing the model's proficiency in utilizing frames considered relevant for processing. Our model minimizes the number of frames used, where it tends to select those closest to the event, as they are more correlated with the anticipated action. Due to the redundancy in the input frames, the model chooses frames with significant differences to reduce redundancy and maximize information extraction. This enables our TAMformer model to adaptively learn a mask that exploits only the most important frames during the anticipation process.

aim to incorporate language into the third-person pedestrian action anticipation task. In our case, pedestrian decisions are influenced by many factors, including their current actions, demographic profiles, collective behaviors of nearby individuals, and the overall street conditions. Consequently, language integration holds the potential to capture these diverse factors more accurately compared to solely relying on visually extracted representations. Nevertheless, this direction presents challenges, particularly in providing textual ground truth, especially in the complex urban environment, where we address this challenge, seeking to overcome obstacles associated with providing accurate textual descriptions in urban settings.

Furthermore, pedestrians' actions are not limited to crossing or not crossing, where many diverse behaviors can be observed in the urban environment. Anticipating these varied behaviors not only allows for a deeper comprehension of pedestrian actions but also enhances the accuracy of predicting crossing intentions. Therefore, we expand the conventional binary task of crossing/not crossing intention prediction into multi-action anticipation of five different pedestrian activities: ***walking***, ***stopping/standing***, ***waiting to cross***, ***crossing***, and ***other***, denoting uncategorized actions.

To summarize, the main contributions of this paper are as follows:

1. We extend the pedestrian action anticipation task to anticipate multiple actions instead of the conventional crossing/not crossing binary task.
2. We integrate vision and language to enhance the performance of action anticipation in complex urban scenarios, leveraging the complementary nature of these modalities.
3. We conduct an in-depth exploration of various techniques for generating per-frame textual captions, facilitating a deeper understanding of pedestrian behavior.
4. We evaluate our multi-modal transformer model on three well-known datasets, validating the efficacy of our model in anticipating pedestrians' actions.
5. We conduct extensive experiments on a novel large-scale multi-actions dataset to demonstrate the effectiveness of vision-language coupling in our action anticipation task.

This paper builds upon our previous work, TAMformer (Osman et al., 2023), which is a multi-modal transformer model, with the ability to measure the relevant importance of observed frames throughout the temporal axis, selecting the most relevant ones and adapting to the anticipation needs at each anticipation time, as illustrated in Fig. 1. In this work, we introduce two major extensions to TAMformer: firstly, we broaden the evaluation scope of TAMformer by assessing its performance on the LOKI dataset (Girase et al., 2021), a recent large-scale pedestrians dataset that allows for multi-action anticipation. We are the first to provide action anticipation performance on LOKI by comparing different anticipation models. Secondly, we introduce language integration into TAMformer, exploring the impact of incorporating textual modality alongside visual inputs on the anticipation performance. To this end, our experiments with TAMformer are substantially expanded to investigate the implications of language integration and to identify the optimal approach for generating appropriate textual modality.

2. Related works

2.1. Action anticipation

Action anticipation, whether in first-person settings (Fan et al., 2018; Furnari et al., 2017; Rhinehart and Kitani, 2020; Zhang et al., 2017; Camporese et al., 2021; Furnari and Farinella, 2021; Osman et al., 2021) or third-person videos (Abu Farha et al., 2018; Felsen et al., 2017; Gao et al., 2017; Mahmud et al., 2017; Zeng et al., 2017; Camporese et al., 2023), is gaining significant traction due to its relevance in various real-world applications such as robotic assistance and autonomous driving. In urban environments especially, pedestrian action anticipation plays a crucial role in enabling safe navigation for self-driving cars, resulting in a surge of research efforts in this domain (Neogi et al., 2020; Kotseruba et al., 2020; Liu et al., 2020; Kotseruba et al., 2021; Yang et al., 2021; Lorenzo et al., 2021, 2020; Gesnouin et al., 2021; Razali et al., 2021; Rasouli et al., 2021; Osman et al., 2022; Correia et al., 2022; Rasouli et al., 2022; Achaji et al., 2022; Osman et al., 2023; Yang et al., 2023; Rasouli and Kotseruba, 2023). Many techniques for pedestrian behavior and action prediction have been proposed, varying in the features extracted from observed images and the temporal backbones used to process time. Most related works rely on visual features extracted from images to describe the scene, along with pedestrian trajectories, typically employing an LSTM for temporal processing, as in Neogi et al. (2020), Kotseruba et al. (2020), Liu et al. (2020), Kotseruba et al. (2021), Yang et al. (2021), Lorenzo et al. (2021, 2020), Gesnouin et al. (2021), Rasouli et al. (2021), Osman et al. (2022) and Correia et al. (2022). More recent works have proposed alternative techniques, such as Rasouli et al. (2022), which employs two sets of Conv2D layers. An interesting approach in Achaji et al. (2022) uses pedestrian bounding box history alone to describe pedestrian movement and achieve high performance. Other recent techniques transform the observed scene into a graph and apply graph convolution to encode it, as in Yang et al. (2023). With the rise of transformer-based architectures, which have shown impressive performance across various tasks, recent works have employed transformers for pedestrian action prediction, such as Rasouli and Kotseruba (2023) and Osman et al. (2023). In this work, we also rely on a transformer-based architecture.

Pedestrian action prediction is mainly formed to predict the action of crossing or not crossing the street, as in JAAD (Rasouli et al., 2017), and PIE (Rasouli et al., 2019) datasets. However, recent work in pedestrian motion prediction (Girase et al., 2021) expanded the behavioral descriptions of pedestrians to include a broader range of actions such as walking, standing, crossing, and waiting to cross. However, while the aforementioned work primarily focuses on the trajectory prediction of observed agents, we aim to utilize such wider actions in the pedestrian action anticipation task.

2.2. Multi-modal modeling

Multi-modal modeling is becoming very popular due to its effectiveness in enriching the exacted information by merging multiple views about the observation. Each modality catches distinctive details on the observed input. Often, different modalities are complementary to each other, where fusing the multi-modal feature spaces would comprehensively describe the observed scene (Wang, 2021; Hu et al., 2022). Many previous works adopted a multi-modal design and proved an improved performance in the action anticipation and detection task (Furnari and Farinella, 2021; Liu et al., 2018; Kotseruba et al., 2021). Many techniques have been proposed for the fusion of multiple modalities; for example, in Furnari and Farinella (2021) and Osman et al. (2022), an attention-based approach is used for fusing the different modalities. Another interesting and popular technique targeting transformer-based models is proposed in Hu and Singh (2021).

In the Pedestrian action prediction task, many fusion techniques have been proposed. Most famously is the late fusion approach between encodings of the fused modalities, as in Kotseruba et al. (2021), Yang et al. (2021), Lorenzo et al. (2021), Gesnouin et al. (2021), Lorenzo et al. (2020) and Yang et al. (2023). Other works applied different fusion techniques, such as Rasouli et al. (2021, 2022), which fuses the different modalities at different levels inside the prediction model. Most recently, Rasouli and Kotseruba (2023) employed a transformer module to cross-encode the input modalities. Similarly, in Osman et al. (2023), a transformer-based architecture is utilized to crossly encode and decode the different modalities. In this work, we adopt a transformer-based fusion technique of multiple modalities in our TAMformer model.

2.3. Temporal and spatial scaling

Video-based models need to process temporal and spatial information. Throughout the temporal axis, the used processing frame rate affects the amount of extracted information. Consequently, many works proposed to model the temporal axis at different frame rates (Feichtenhofer et al., 2019; Sener et al., 2020; Osman et al., 2021; Wu et al., 2019). However, relying on fixed frame rates requires a pre-step of hyperparameter tuning, in addition to the possibility of losing important information during the sampling process. Therefore, our approach utilizes an adaptive mechanism to weigh the importance of the observed frames and learn the applied attention masks inside our transformer model.

Similar to temporal modeling, spatial scale affects the information extracted from an image. Larger scales allow for more details but could incorporate more noise. On the other hand, smaller spatial scales could miss the key-hidden input in the image. Consequently, many works adopted multi-spatial-scale design to create a more robust model to scale changes, especially in image classification and object detection problems (Nah et al., 2017; Niu et al., 2019; Lin et al., 2017; Zhang et al., 2023). Therefore, we consider multi-spatial-scale during the generation of per-frame language modality to allow for more robust textual descriptions.

2.4. Vision-language integration

Language provides a rich and contextual way to describe visual content, where it can provide additional context and details that may not be immediately apparent by visual feature extraction. Therefore, coupling language features and visual features has gained massive attention in computer vision, where numerous studies have proposed innovative techniques for such coupling and for extracting suitable language features for visual tasks. For example, in Burns et al. (2019), a study is conducted to measure the effectiveness of different language models and text embedding approaches on different visual tasks. While in Lu et al. (2019), VilBERT presents a co-attention transformer model that couples text and images into a multi-modal transformer to solve different vision tasks, including visual question answering and caption-based image retrieval.

The significant interest in this field has led to the production of breakthrough models, such as CLIP (Radford et al., 2021), a transformer-based model aiming at coupling textual inputs with images for zero-shot text/image retrieval. In addition to the impressive advances in natural language processing (NLP), with the witnessed improvements of the GPT models (Floridi and Chiriatti, 2020), and the image captioning models, such as BLIP (Li et al., 2022, 2023), that applies a two-stage of training: representation learning stage of the image, and a generative stage of the textual caption to the image.

Given such advances in the language-visual models, visual tasks have started integrating language and textual input to enhance their performance. Specifically, for the task of action anticipation, recent models proposed the generation of textual labels and their integration in the anticipation model (Manousaki et al., 2023; Zhao et al., 2024; Ghosh et al., 2022; Huang et al., 2023; Das and Ryoo, 2022). VLMAH (Manousaki et al., 2023) anticipates first-person actions, integrating images and textual descriptions of the observed actions and applying bidirectional LSTM encoders for visual and textual inputs. While in Zhao et al. (2024), a large language model is utilized to predict future actions based on extracted textual descriptions of the input images, where the visual task is fully transformed into a lingual task. A similar idea is applied in Ghosh et al. (2022), proposing knowledge distillation from the lingual anticipation model to the visual anticipation model. Again, a large language model is used for anticipation in Huang et al. (2023) using image captions generated for the input images. Another interesting technique is proposed in Das and Ryoo (2022), where CLIP encodes images and textual labels of the observed actions, followed by a transformer aggregation anticipation model.

In this work, we aim to couple a lingual modality with the visually extracted features for third-person action anticipation of pedestrians' behaviors, where we generate per-frame textual descriptions projected into the feature space of our multi-modal model to enrich the encoded representation inside the model and enhance the anticipation performance.

3. TAMformer model

We extend the TAMformer architecture (Osman et al., 2023), a multi-modal action anticipation model, to include textual prompts. As depicted in Fig. 2, TAMformer comprises four key modules: the value encoder, the query encoder, the decoder, the anticipation head, and the temporal adaptive mask module.

3.1. Value encoder

Raw images are first projected into different modalities to extract features $\mathbf{x}_m \in [\mathbf{x}_1, \ldots, \mathbf{x}_M]$. Here, M represents the number of modalities, $\mathbf{x}_m \in \mathbb{R}^{T \times D_m}$, where D_m denotes the feature size, and T signifies the time sequence length. Each projected modality $\mathbf{x}_m$ is augmented with positional embeddings $\mathbf{p}$ to preserve temporal causality before being passed through a transformer block TE_m to produce an encoded

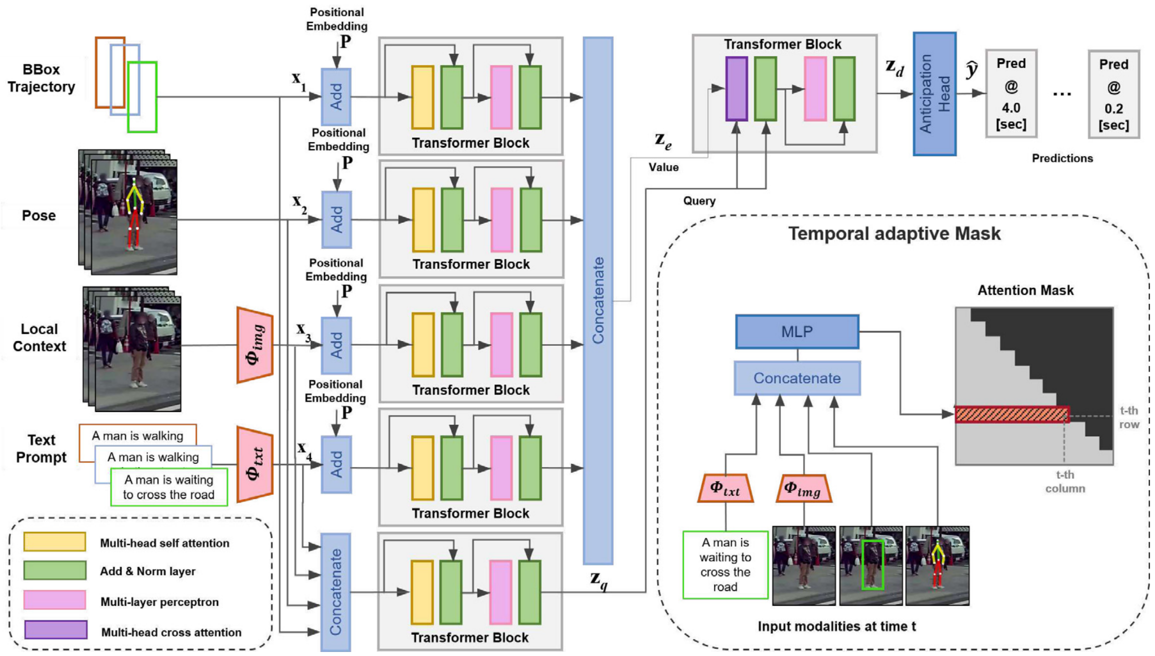


Fig. 2. TAMformer model.

representation $\mathbf{z}_m \in \mathbb{R}^{T\times D_m}$. Finally, the value encoding $\mathbf{z}_e$ is obtained by concatenating the encoded modalities:

$$\mathbf{z}_m = TE_m(\mathbf{x}_m + \mathbf{p}), \quad \mathbf{z}_e = Cat\left[\mathbf{z}_1, \ldots \mathbf{z}_M\right] \tag{1}$$

TAMformer can accommodate various modalities. However, in this work, we focus on four distinct modalities: bounding box trajectories, pose features derived from OpenPose, RGB features obtained from cropped pedestrian images to encapsulate the local scene context, and linguistic features. When the language modality is excluded, the image feature extractor Φ_{img} is represented by a VGG16 (Simonyan and Zisserman, 2015) network. However, when we include the textual part, to align the visual and textual modalities, we utilize a pre-trained CLIP framework (Radford et al., 2021) for extracting visual features ($\Phi_{img} = CLIP_{img}$) and textual features ($\Phi_{txt} = CLIP_{txt}$).

3.2. Query encoder

Next, TAMformer employs a query encoder to further enrich the extracted representation by allowing a step of early interaction between the input modalities. Here, the projected features are concatenated and then encoded using the transformer block TQ:

$$\tilde{\mathbf{z}}_q = Cat\left[\mathbf{x}_1, \ldots, \mathbf{x}_M\right], \quad \mathbf{z}_q = TQ(\tilde{\mathbf{z}}_q) \tag{2}$$

3.3. Decoder

The decoder utilizes a multi-head cross-attention transformer TD, where the value encodings serve as the value and the query encodings serve as the query. The final decoding $\mathbf{z}_d$ is passed to the anticipation head to produce the predictions at the corresponding anticipation times, employing a multi-layer perceptron followed by a sigmoid activation function:

$$\mathbf{z}_d = TD(\mathbf{z}_e, \mathbf{z}_q), \quad \hat{\mathbf{y}} = Sigmoid(MLP(\mathbf{z}_d))$$

3.4. Temporal adaptive mask

To filter out redundant frames while retaining those deemed most significant during processing, TAMformer incorporates a learning module that adapts attention masks within transformer blocks to reflect the temporal importance of observed frames. As depicted in Fig. 2, input features at time t are concatenated and fed into a feed-forward network, which predicts attention masks up to the tth column. Frames beyond t are masked to prevent the utilization of future information in present predictions. The mask $\mathbf{M}$ is predicted up to t as follows:

$$\begin{aligned} \tilde{\mathbf{z}} &= Cat\left[\mathbf{x}_1, \ldots, \mathbf{x}_M\right], \\ \mathbf{M}_{[:t]} &= Sigmoid(MLP(\tilde{\mathbf{z}}_{[:t]})). \end{aligned}$$

The predicted mask $\mathbf{M}$ is passed to the value encoders TE_m, the query encoder TQ, and the decoder TD, employed as a masking strategy inside the transformer blocks. In this way, we were able to allow the model to learn and choose only the most influential frames at each time step to be utilized during the anticipation process.

3.5. Auxiliary loss

Typically, as anticipation models approach the time of the anticipated action, their performance tends to improve. To capitalize on this observation, TAMformer utilizes an auxiliary regularization loss function:

$$\mathcal{L}_r = \sum_t \left\| \mathbf{z}_d[t] - \mathbf{z}_d[T] \right\|^2$$

which aims to minimize the disparity between the current decoder embedding $\mathbf{z}_d[t]$ and the final embedding $\mathbf{z}_d[T]$. It employs a two-stage training approach: initially, pre-training the model using only the cross-entropy loss $\mathcal{L}ce$ for action anticipation. Subsequently, the regularization term $\mathcal{L}r$ is incorporated into the total loss ($\mathcal{L} = \mathcal{L}ce + \mathcal{L}r$), thereby encouraging earlier anticipation predictions to leverage the comprehensive representation provided by the final decoder embedding, which encapsulates the entire sequence preceding the onset of the action.

3.6. Data augmentation

In contrast to the standard protocol outlined in Kotseruba et al. (2021), TAMformer departs from the use of overlapped samples and instead employs the protocol proposed in Osman et al. (2022), where each pedestrian is treated as a single sample. Consequently, this approach results in a significant reduction in the number of samples

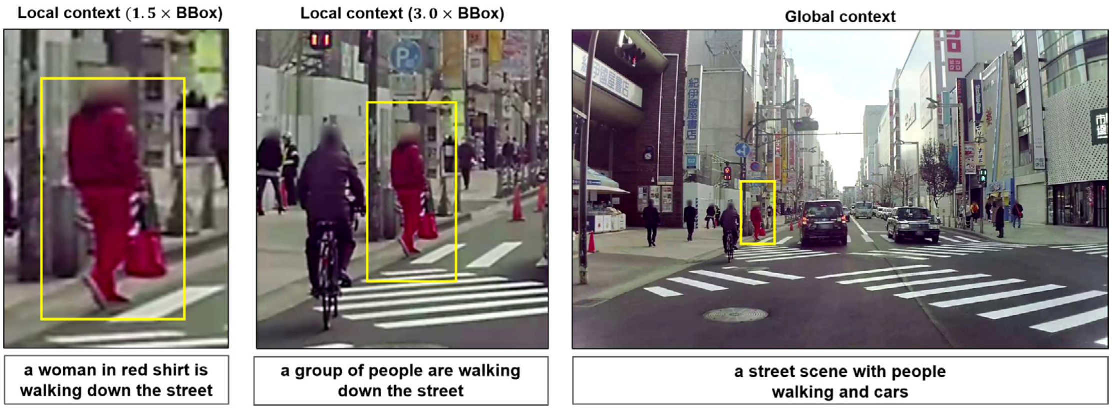

Local context (1.5 × BBox)
Local context (3.0 × BBox)
Global context
a woman in red shirt is walking down the street
a group of people are walking down the street
a street scene with people walking and cars

Fig. 3. An example of the effect of spatial scale on the generated text descriptions. The **Global Context** is the complete scene, without cropping. While a **Local Context** means a cropped image with a ratio/scale of the size of the pedestrian bounding box (*BBox*).

Table 1
Predefined per-frame captions.

Description category	List of predefined captions
Demographic description	A woman is A man is A girl is A boy is
Basic activity	walking - standing
Location description	on the sidewalk on zebra line next to a bus station next to a car next to traffic lights in front of a building
Behavioral description	while talking to someone while talking in the phone while looking at the phone and raising his hand while carrying a baby while pushing a strolling holding a crutch
Interaction description	with a child with a group of people with a dog

ex. A man is walking on the sidewalk.
ex. A woman is standing in front of a building while talking on the phone with a child.

compared to the methodology presented in Kotseruba et al. (2021). However, given the substantial data requirements of transformer models for optimal performance, a data augmentation strategy is proposed to augment the training dataset in case of limited training samples.

Following the methodology described in Osman et al. (2022), the observation length is set to 4.5 s, disregarding any frames occurring before this duration in each sample. Leveraging these history frames to augment the data, TAMformer replaces the encoding window with history frames whenever available, thereby generating multiple versions of the same sample with different encoding windows. This augmentation strategy effectively enriches the training data, enhancing the model's ability to learn robust representations.

4. Language modality generation approaches

The main challenge in integrating language into the pedestrian action anticipation task is the absence of textual labeling for the images or the scene. In our work, we investigate different techniques for generating textual descriptions for the input images. Herein, we discuss the investigated techniques for language generation.

4.1. Predefined per-frame captions

As our task addresses action anticipation in urban scenarios, this could limit the range of possible activities that can be performed in an urban environment. A straightforward technique to couple an image with a textual description is to predefine a set of likely descriptions and then pair a given image with its closest description in feature space. We split a textual description into four parts: Demographic description, basic activity description, behavioral description, location description, and interaction description. Table 1 summarizes the possible captions in each category, where each category aims to cover a possible aspect in a given scene. The final list of predefined captions represents all possible combinations between the pre-defined texts. The chosen caption C_I for an image I should be the closest to the image in feature space, as depicted in (3), where K is the number of predefined captions.

$$C_I = \operatorname*{argmin}_{k=1}^{K} \left\| CLIP_{img}(I) - CLIP_{txt}(C_k) \right\| \tag{3}$$

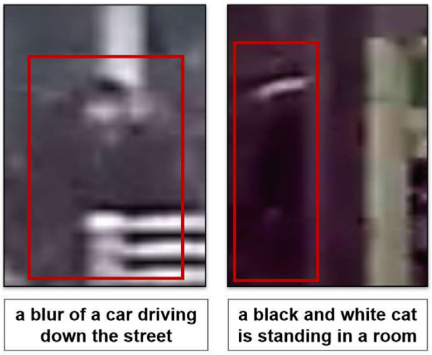
4.2. Pre-trained image captioner

A more reasonable approach for getting a textual description for an image is using an image captioner. BLIP-2 (Li et al., 2023) is a recent image captioning framework that produces robust textual descriptions in multiple domains. We rely on a pre-trained BLIP-2 model to generate our per-frame captions. However, many factors can affect the generated text, such as the scale of the context enclosed around the pedestrian in the image or the amount of noise incorporated. Therefore, we conduct multiple steps to ensure the accuracy and clarity of the generated text, including a multi-spatial-scale captioning procedure, in addition to a text cleaning step.

4.2.1. Multi-spatial-scale captioning

A single frame in our input represents a wide-view urban scene; however, as we focus on a single pedestrian in the scene, the frame is cropped, centering the pedestrian in the cropped image while adding a reasonable amount of context to the observation. However, the amount of context added to the cropped frame affects the interpretation of the image and, consequently, the generated caption, as illustrated in Fig. 3. To allow more informative descriptions, we employ different cropping scales, capturing different contexts and information from the scene. Let S be the number of considered cropping scales, and C^s_{BLIP} is the description generated by the BLIP-2 for scale s, then the projected lingual modality $\mathbf{x}_{BLIP}$ for the generated text description is given by (4).

$$\mathbf{x}_{BLIP} = \sum_{s=1}^{S} CLIP_{txt}(C^s_{BLIP}) \tag{4}$$



a blur of a car driving down the street

a black and white cat is standing in a room

Fig. 4. Examples of corrupted captions due to unclear and noisy images.

4.2.2. *Captions cleaning and refinement*

The generated captions incur a large amount of noise. Firstly, the pre-trained BLIP-2 model is not specialized in urban scenarios; therefore, the captions could contain unrelated descriptions to the urban scenes. Additionally, we have a high percentage of noisy images due to zooming in for centering a relatively far pedestrian, which leads to corrupted descriptions, as exemplified in Fig. 4. To overcome such noisy captions, we first ignore all captions containing any special characters, numbers, specific dates, or places that might be guided by a misleading interpretation of the images, leading to such specific and unrelated information. Additionally, to further filter out unrelated captions, we create a dictionary of words related to the urban environment and pedestrians' activities, *e.g., street, car, walk, etc.* Any generated caption that seems to be unrelated is overlooked and replaced by a general descriptive text, such as ***"A person is observed in the street"***.

4.3. *Template-based per-frame captions*

Template captions represent a more stable technique for providing textual descriptions for the observed frames. A template sentence is pre-defined to be filled with some relevant inquiries about the observed pedestrian. In our work, we pre-defined a textual template, given by: "*A/An **person_des** is **action_des***". The keyword "***person_des***" is replaced with the corresponding pedestrian description (*age/gender*), and "***action_des***" is replaced by the detected activity in the corresponding frame, *i.e., walking, standing, crossing, etc*. For example, one description could be "***An adult female is walking in the street***".

The chosen attributes reflect the most important factors affecting the future activity of the pedestrian. For example, both age and gender have a considerable influence on the pedestrians' decisions and, consequently, their future actions. The most influential attribute is the history of actions taken by the pedestrian, providing a clear description of the observed events. The tremendous evolution in human detection and action recognition models concedes the feasibility of obtaining these descriptive attributes. However, Herein, we rely on the ground truth existence of these attributes in our evaluation dataset.

4.4. *Text prompting*

Much like human perception, where different synonyms of a single word can influence the comprehension of a sentence, language models are similarly sensitive to this phenomenon. Depending on the popularity of a given word and its contextual prevalence during the language model's training, the obtained representation would be more/less informative. Considering this effect, we leverage text prompting within captions to optimize the model's perception of the context and enhance its understanding of the described events.

Table 2
List of synonyms used in manual prompting.

Age synonyms	Adult - grownup Child - little child
Gender synonyms	Female Male
Age/Gender synonyms	Woman - Lady Man - Gentleman Girl - Schoolgirl Boy - Schoolboy
Action synonyms	Walking - Moving Standing - Stopped - Was walking and stopped Waiting to cross - Wanting to cross Crossing the street - Crossing the road Does not want to cross Observed - Seen

ex. C^1_{temp}: An adult female is walking in the street.
C^2_{temp}: A woman is walking in the street.
C^3_{temp}: A grownup female is moving in the street.
C^4_{temp}: A lady is walking in the street and does not want to cross.

4.4.1. *Manual prompting*

We investigate the effect of promoting in our template-based captioning approach. For both ***"person_des"*** and ***"action_des"***, we prompt the text by alternating their synonyms, producing multiple synonym descriptions for the same image, allowing for a richer language representation. Table 2 summarizes the synonyms used for the prompting process. Given an initial template-based caption C^1_{temp}, a number P of prompted descriptions are generated, where the collective representation of the projected captions $\mathbf{x}_{temp}$ is as follows:

$$\mathbf{x}_{temp} = Cat\left[CLIP_{txt}(C^1_{temp}), \dots, CLIP_{txt}(C^P_{temp})\right] \tag{5}$$

4.4.2. *ChatGPT prompting*

Recently, GPT models achieved a significant jump in the field of natural language processing, whereas ChatGPT showed impressive capabilities in multiple language tasks, including paraphrasing. We asked ChatGPT to prompt our template sentence into a set of possible paraphrased captions. To permit the production of more prompted textual descriptions through ChatGPT, the concatenation fusion approach in (5) is replaced with a summation.

4.4.3. *Image captioner prompting*

The descriptions generated by an image captioner represent a broader interpretation of the observed frames compared to the predefined template. In contrast, template-based captioning provides a more to-the-point description of the event. To take advantage of both captioning techniques, the generated BLIP-2 text is prompted by our predefined template, specifically, "***person_des***" and "***action_des***". Given a generated BLIP-2 caption C_{BLIP}, we first prompt the text with the demographic age/gender attributes, providing C_{BLIP_D}. Then C_{BLIP_D} is prompted with the action attributes, generating C_{BLIP_AD}. The projected features of the prompted text $\mathbf{x}_{BLIP\_AD}$ is given by (6), where P is the number of manually-prompted captions, S is the number of the considered spatial scales, while β and γ are uncertainty values to reflect the possibility of erroneous during the prompting process.

$$\begin{aligned}\mathbf{x}_{BLIP\_AD} = \sum_{s=1}^{S}(\beta \times CLIP_{txt}(C^s_{BLIP}) + \gamma \times CLIP_{txt}(C^s_{BLIP\_D}) \\ + CLIP_{txt}(C^s_{BLIP\_AD})) + \sum_{p=1}^{P} CLIP_{txt}(C^p_{temp})\end{aligned} \tag{6}$$

5. Experiments

5.1. *Datasets and evaluation metrics*

The performance of TAMformer is evaluated on three datasets: JAAD (Rasouli et al., 2017), containing two subsets: JAAD$_{beh}$ with

Table 3
Architectural and performance Comparison of different SOTA models in the standard anticipation range [2−1] s. We compare three architectural aspects: (1) **Visual Backbone**, the backbone model used in extracting context features from images; (2) **Blocks**, the time sequence processing framework; (3) **Fusion**, the merging technique of the multi-modalities, where **L-ATT** stands for **L**ate **ATT**ention, **EC** is **E**arly **C**oncatenation, and **LC** is **L**ate **C**oncatenation. For a fair comparison of TAMformer with previous works, we overlook the language modality in this table, where JAAD uses BBOX, pose, and VGG16 features; PIE adds in the velocity features; and LOKI employs only the BBOX and the VGG16 features.

	Visual backbone	Blocks	Fusion	t_a	S_{temp}	LOKI		PIE		JAAD_{all}		JAAD_{beh}	
						AUC	F1	AUC	F1	AUC	F1	AUC	F1
PCPA (Kotseruba et al., 2021)	C3D	GRU	L-ATT	[2−1] s	No sampling	0.51	0.54	**0.86**	0.77	0.79	0.57	0.50	0.67
RU-LSTM (Furnari and Farinella, 2021)	VGG16	LSTM	L-ATT	[4−0.1] s	Manual	0.56	0.55	0.84	0.77	0.78	0.62	0.62	0.78
G-RULSTM (Osman et al., 2022)						–	–	–	–	0.80	0.63	0.65	0.80
TAMformer (ours)	VGG16	TF	EC+LC	[4−0.1] s	Adaptive	**0.59**	**0.58**	**0.86**	**0.79**	**0.83**	**0.68**	**0.69**	**0.8**

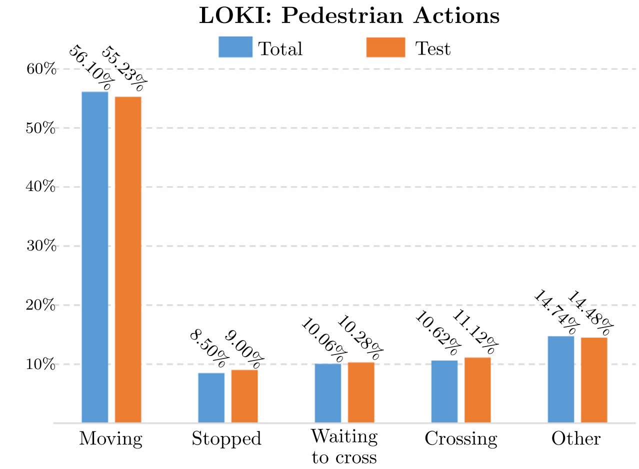


Fig. 5. Percentage of different actions in LOKI dataset.

only behaviorally annotated subjects (495 crossing samples and 191 not crossing samples), and JAAD_{all} with an additional 2100 not crossing samples. Then, we have PIE (Kotseruba et al., 2020), a larger dataset containing 1842 behaviorally annotated pedestrians (519 crossing pedestrians and 1323 not crossing pedestrians), in addition to more annotations of the ego-vehicle, i.e., speed.

Additionally, we evaluate TAMformer on a more recent large-scale urban dataset LOKI (Girase et al., 2021). The dataset provides a larger-scale behavioral and demographical pedestrian annotation compared to older datasets. It comprises 9226 pedestrians, annotated, at 5 FPS, with age, gender, 2D bounding boxes, 3D bounding boxes, and intended action in the next 0.8 s (4 frames). The action labels for pedestrians consist of 5 actions: *Moving, Stopping, Waiting to cross, Crossing, and Other*, where Fig. 5 illustrates the percentage of the different actions throughout the whole dataset, as well as for the test set.

Due to the considerable imbalance in the dataset, we rely on the AUC and F1-score metrics for evaluation. Following Rasouli et al. (2017) and Rasouli et al. (2019), we split the dataset into 60% training and 40% testing.

5.2. Implementation details

The training procedure includes two phases (500 epochs each): a pre-training phase on action anticipation and a tuning phase with auxiliary loss. We used the SGD optimizer with learning rates $l_r = \{10^{-5}, 10^{-2}, 10^{-3}, 10^{-3}\}$ for PIE, JAAD_{all}, JAAD_{beh}, and LOKI, respectively. Each transformer block has $N_h = 6$ heads, $ff_{dim} = 1024$, and the MLP producing the learned masks consists of $N_l = 3$ layers with sizes $\{128, 64, 32\}$. The earliest anticipation time is set to 4.0 s, and the warm-up window is 0.5 s for JAAD and PIE, while it is 1.0 s for LOKI. The time step α is set to $\{0.1, 0.2, 0.2\}$ for JAAD, PIE, and LOKI, respectively.

For multi-spatial-scale captioning, local scales are given by $[1.5\times, 5.0\times]$ with step $0.5\times$, in addition to the global scale, where the overall number of scales used is $S = 9$. Furthermore, for manual text prompting, we set the number of prompts P to 4 prompted captions, while it is set to 10 for ChatGPT prompting.

We consider the following modalities in TAMformer: The bounding box trajectory (BBOX), the cropped images with a ratio $1.5\times$ of the bounding box (local context), the pose features extracted with OpenPose, the velocity of the vehicle, and the generated textual descriptions. For both JAAD and PIE, we rely on BBOX, local context, and pose modalities, and we add the velocity for PIE. Meanwhile, we rely on BBOK, local context, and language modalities for LOKI.

5.3. TAMformer evaluation results

We compare our model with PCPA (Kotseruba et al., 2021), which represents the SOTA work in intent prediction and an adapted PCPA version that can produce earlier anticipations. Although we are not applying the overlapping protocol in Kotseruba et al. (2021), we align with it on the used samples and anticipation range during evaluation to allow for a fair comparison. Additionally, we compare with RULSTM (Furnari and Farinella, 2021), and G-RULSTM (Osman et al., 2022). Following Kotseruba et al. (2021), Table 3 reports the comparison in the anticipation range of [2−1] s and the main architecture differences. We observe an F1-score out-performance gap that reaches +3% on LOKI, +2% on PIE, and +5% on JAAD_{all} when comparing our TAMformer to the best model in the table. Moreover, we reported a comparison on different anticipation times from 4 s to 1 s in Table 4 and, depending on the dataset, we notice two trends: for LOKI and PIE, TAMformer outperforms previous models in all anticipation times, where it reaches +3% increase in F1-score at 1 s in LOKI, and achieves +2% F1-score improvement at almost all anticipation times for PIE. Nevertheless, on JAAD, our model suffers a degraded performance at early anticipation ([4−3] s) while maintaining the improvements on JAAD_{all} (maximum +9%) and on JAAD_{beh} (maximum +2%). The reduction in training samples in early anticipation (>50% on JAAD) could explain this degradation, as transformers need lots of training samples.

To measure the effect of the integrated auxiliary loss in TAMformer, Fig. 6 illustrates the F1-score performance of the model with and without applying the $\mathcal{L}_r$ loss throughout the anticipation phase. As observed across all datasets, anticipation performance degrades as the anticipation time increases, with the best performance at 1 s and the worst at 4 s. To address this, we introduce an auxiliary loss aimed at minimizing the gap between the best anticipation time and earlier anticipations. The performance is generally enhanced for all four datasets, especially at the earlier anticipation. This is reasonable; as we get closer to the event, the performance typically gets higher, and the addition of our loss will not add much value. This effect is especially pronounced in the JAAD dataset, both JAAD*all* and JAAD*beh*, which are smaller and imbalanced, making the model more reliant on the auxiliary loss for better performance. The model still benefits from the auxiliary loss for larger datasets like PIE and LOKI. However, the effect is less dominant since these datasets are sufficiently large to allow effective training without the auxiliary loss.

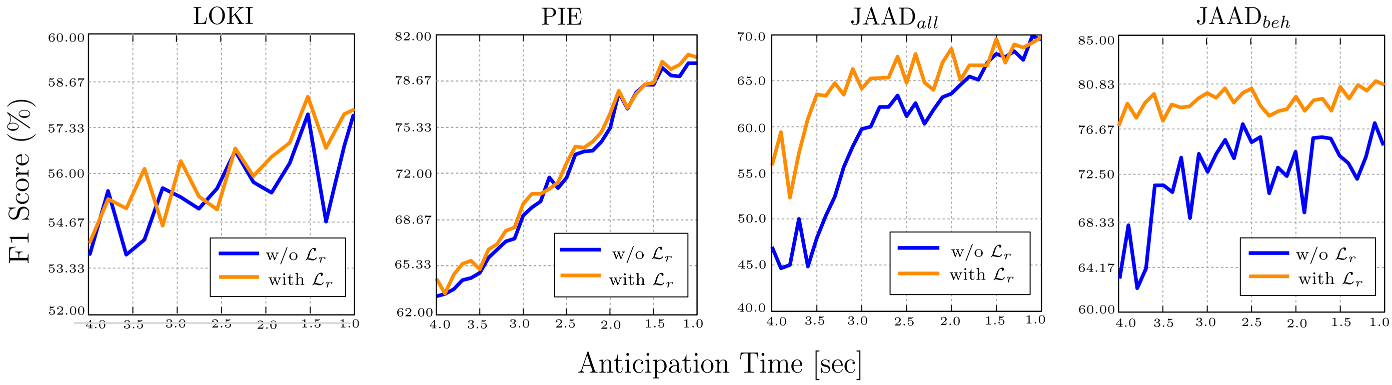


Fig. 6. Effect of using the $\mathcal{L}_r$ loss.

Table 4
Performance at different anticipation times [4−1] s, where † denotes that the reported results are our run of PCPA.

	LOKI							
	4 s		3 s		2 s		1 s	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
PCPA† (Kotseruba et al., 2021)	0.50	0.53	0.50	0.53	0.50	0.55	0.50	0.56
RU-LSTM (Furnari and Farinella, 2021)	0.54	0.54	0.54	0.55	0.56	0.55	0.56	0.55
TAMformer (ours)	**0.58**	**0.55**	**0.59**	**0.57**	**0.58**	**0.57**	**0.59**	**0.59**
	PIE							
PCPA† (Kotseruba et al., 2021)	0.75	0.62	0.76	0.63	0.84	0.73	0.85	0.77
RU-LSTM (Furnari and Farinella, 2021)	0.76	0.63	0.79	0.68	0.82	0.74	0.85	0.79
TAMformer (ours)	**0.77**	**0.65**	**0.81**	**0.7**	**0.84**	**0.76**	**0.88**	**0.8**
	JAAD$_{all}$							
PCPA† (Kotseruba et al., 2021)	0.75	0.50	0.76	0.53	0.75	0.52	0.79	0.55
RU-LSTM (Furnari and Farinella, 2021)	**0.76**	**0.57**	0.78	**0.64**	0.76	0.59	0.78	0.62
TAMformer (ours)	0.75	0.56	**0.79**	**0.64**	**0.82**	**0.68**	**0.82**	**0.7**
	JAAD$_{beh}$							
PCPA† (Kotseruba et al., 2021)	0.51	0.62	0.46	0.54	0.45	0.61	0.52	0.63
RU-LSTM (Furnari and Farinella, 2021)	**0.67**	**0.79**	0.64	**0.81**	0.62	0.78	0.63	0.79
TAMformer (ours)	0.62	0.77	**0.68**	0.80	**0.7**	**0.79**	**0.69**	**0.81**

Table 5
The performance comparison of TAMformer, with and without language integration, employing various captioning techniques, across different anticipation times $\{4.0, 3.0, 2.0, 1.0\}$ s on LOKI dataset.

	4 s		3 s		2 s		1 s	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
W/o Captioning	53.51	55.5	53.92	55.83	54.71	57.60	55.74	58.41
Predefined captions	56.18	54.99	58.09	55.51	58.37	57.78	60.40	58.32
Image captioner	56.34	57.65	57.96	58.18	61.28	57.97	63.98	60.39
Template captions	70.10	64.51	70.77	65.14	72.98	70.10	80.80	77.68
Manual prompting	77.30	71.05	80.86	75.95	84.64	80.92	89.67	87.40
ChatGPT prompting	77.07	70.11	80.94	**76.24**	**85.35**	**81.34**	89.95	87.43
Captioner prompting	**77.77**	**72.16**	**80.95**	75.98	85.24	**81.34**	**90.16**	**87.92**

5.4. Language modality evaluation

Given the larger scale of the LOKI dataset and its relatively low anticipation performance, we focus on evaluating TAMformer on the non-binary anticipation of multiple actions in LOKI when language is incorporated into the model.

Table 5 provides a comparative analysis of TAMformer's performance using various captioning techniques. As the anticipation space expands to more actions, the anticipation becomes increasingly complicated, resulting in poor performance. Yet, the integration of any form of textual description into the model significantly improves anticipation performance. Notably, not all captioning approaches yield equally impactful enhancements. The predefined captioning approach, which selects the nearest predefined caption to the image in CLIP feature space, shows only marginal improvement over the model without textual input. This may be attributed to the possible limitation in added value compared to using image features alone, in addition to the restricted range of events and activities permitted in the predefined captions limiting the capabilities of the provided textual descriptions. In contrast, employing a pre-trained image captioner (BLIP-2) allows for broader descriptions and interpretations, enriching the model with more valuable representations and yielding much-improved anticipations. However, our template-based captioning proves quite advantageous, offering more precise and accurate descriptions and achieving notably higher performance.

Furthermore, textual prompting of all types shows impressive performance enhancement. Almost all prompting techniques are on par,

Table 6
The effect of changing the spatial scale of the cropped image in the generated BLIP-2 captions and the anticipation performance.

	4 s		3 s		2 s		1 s	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
1.5×	55.62	55.74	56.53	56.7	56.47	**58.36**	58.81	57.84
3.0×	54.61	56.64	**60.23**	54.33	56.75	57.96	59.40	58.28
5.0×	55.54	56.73	56.61	57.41	59.29	56.82	59.80	57.80
Global	55.43	56.62	59.70	54.82	58.18	56.3	59.40	57.13
Fusion	**56.34**	**57.65**	57.96	**58.18**	**61.28**	57.97	**63.98**	**60.39**

yet the best performance is achieved by prompting the generated BLIP-2 text using the predefined template captions, which takes advantage of both the generality of the pre-trained image captioner and the precise template-based captions. ChatGPT prompting allowed for increasing the number of prompted textual descriptions, richer representations, and high performance.

5.5. Ablation study on the language modality

We conduct a set of ablation experiments to assess the effect of the different parameters in the captioning techniques utilized on the LOKI dataset.

5.5.1. Spatial scaling

Table 6 reports the effect of changing the spatial scale of the cropped image on the generated captions, leading to a shift in the anticipation performance. Each spatial scale in the image allows the captioner to capture a different perspective about the observed scene, as depicted in Fig. 3. Therefore, combining different spatial scales widens the range of the captured information, offering better and more robust performance.

5.5.2. Manual text prompting

We test the effect of using different words describing the age/gender attributes on the anticipation performance, as shown in Table 7. In the first Text Prompt (TP1), the word "person" is used in replacement of the age/gender attributes, *i.e.,* "***a person is walking in the street***", leading to degraded anticipation. Therefore, TP[2:6] show higher performance, as they employ the age/gender description. To measure the effect of changing the word synonyms on the information extracted from the textual description and, hence, on the anticipation performance, we create five test cases in TP[2:6]. In each test case, we alternate the person$_{des}$ using different words and synonyms from Table 2 to describe the age and gender of the person. for example, TP2 uses "***adult/child***" for age and "***male/female***" for gender, while TP3 represents *age/gender* with "***woman/man/girl/boy***", respectively. Therefore, the TP2 caption to describe an adult female person walking is "***An adult female is walking in the street***", while TP3 replaces "***adult female***" with "***woman***" to e "***a woman is walking in the street***". Similarly, for TP[4:6], we alternate and combine different words from Table 2 to create the employed captions. Notably, alternating the words used in the captions affects the perception of the sentence in the language model, adding more/less information to the produced representation and influencing the performance.

Fusing multiple prompted captions improves the extracted information, compared to using a single caption, which leads to enhanced performance, as shown in Table 7. Additionally, as more captions are fused, the anticipation performance is better. Finally, adding TP$_{act}$, which prompts the "***action_des***", pushes the model to further performance improvement.

Table 8 reports the results of applying different fusion schemes on the prompted captions. As depicted, concatenating the language descriptions in the text space and extracting the features of the concatenated texts, as in (7), shows poor performance. Furthermore, increasing the number of concatenated captions decreases the performance even more. This degraded performance of the Cat_{txt} approach could be attributed to the increased complexity of the integrated text, making it harder for the CLIP model to provide a clear representation of the given text. Therefore, a finer approach is to apply the fusion in the feature space instead. Using the Avg_{feat} function, defined in (8), reports a better performance compared to the first approach, where the model is able to benefit from the integrated information. The best performance is achieved through the concatenation of features extracted from the integrated captions, as in (5).

$$Cat_{txt}(P) = CLIP_{txt}(Cat\left[C^{1}_{temp}, \dots, C^{P}_{temp}\right]) \tag{7}$$

$$Avg_{feat}(P) = \frac{\sum_{p=1}^{P} CLIP_{txt}(C^{p}_{temp})}{P} \tag{8}$$

5.5.3. ChatGPT prompting

Using ChatGPT for prompting introduces a broader range of prompts to the captions. In Table 9, we study the impact of increasing the number of prompts generated by ChatGPT. Firstly, comparing our template-based caption C_{temp} and a single randomly generated ChatGPT prompt demonstrates a clear preference for C_{temp}. ChatGPT tends to produce complex and general synonyms, such as "***strolling***" or "***wandering***" for the word "***walking***", potentially blurring the straightforward meaning of the caption and posing challenges for the feature extractor in providing informative representations. However, with increasing the number of prompts generated by ChatGPT, more informative descriptions are incorporated, resulting in improved performance. The peak performance is achieved with 10 generated prompts, beyond which the inclusion of additional prompts introduces numerous complex words, contributing to a decline in performance.

5.5.4. Image captioner prompting

In this ablation study, we examine the impact of different levels of refining the image captions generated by the pre-trained BLIP-2. The first two rows in Table 10 establish a comparison baseline using our template-based captioning approach. The first row represents the

Table 7
The effect of prompting the template-based captions with different synonyms representing the "***person_des***" and the "***action_des***". The Text Prompts (TP[1:6]) prompts the "***person_des***", and **TP**$_{act}$ includes "***action_des***" prompting as well. The fusion technique used is given by (5).

	4 s		3 s		2 s		1 s	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
TP1	67.56	62.65	71.04	66.95	72.52	68.72	80.36	74.46
TP2	70.10	64.51	70.08	65.14	72.98	70.10	80.80	77.68
TP3	69.74	63.79	74.50	68.96	76.74	72.63	80.16	75.37
TP4	69.28	62.97	73.25	67.83	76.50	73.54	79.84	74.79
TP5	65.68	62.84	72.27	65.95	76.97	72.16	82.33	77.42
TP6	68.70	62.60	73.88	67.52	73.63	72.25	80.80	77.20
Fusion (TP2, TP3)	73.42	67.90	77.43	72.66	82.31	78.50	88.00	85.54
Fusion (TP[2:4])	76.35	70.58	80.25	74.37	83.59	79.17	**89.77**	**87.70**
Fusion (TP[2:4], TP$_{act}$)	**77.30**	**71.05**	**80.86**	**75.95**	**84.64**	**80.92**	89.67	87.40

Table 8
Comparing different fusion approaches for the prompted captions. $Cat_{txt}(P)$, defined in (7), denotes text concatenation, where the number of concatenated texts is P. $Avg_{feat}(P)$, given by (8), averages the extracted features from the captions, and $Cat_{feat}(P)$, defined in (5), concatenates the extracted features.

	4 s		3 s		2 s		1 s	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
$Cat_{txt}(P=2)$	65.87	63.63	73.52	63.26	76.06	69.68	78.70	70.37
$Cat_{txt}(P=3)$	62.24	61.69	68.81	64.60	71.09	63.28	72.72	73.69
$Avg_{feat}(P=2)$	65.70	63.42	72.25	65.31	74.68	70.29	81.97	80.53
$Cat_{feat}(P=2)$	**73.42**	**67.90**	**77.43**	**72.66**	**82.31**	**78.50**	**88.00**	**85.54**

Table 9
Ablation on ChatGPT prompting, where P is the number of the prompted captions, GPT denotes ChatGPT prompting, and C_{temp} is the template caption on the form "*A/An* ***{adult/child} {male/female}*** *is* ***{moving/stopping/waiting to cross/crossing/observed}***".

	4 s		3 s		2 s		1 s	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
C_{temp}	70.10	64.51	70.08	65.14	72.98	70.10	80.80	77.68
$GPT(P=1)$	61.08	58.20	63.42	57.88	62.76	60.53	68.12	64.14
$GPT(P=4)$	75.15	70.54	77.77	73.06	81.29	75.74	86.14	84.53
$GPT(P=10)$	77.07	70.11	**80.94**	**76.24**	85.35	**81.34**	**89.95**	87.43
$GPT(P=20)$	**77.02**	**71.29**	80.55	75.86	**85.54**	80.96	89.74	**87.93**

results of using a single template text, while the second row employs manual text prompting. The performance is notably subpar when directly using the generated BLIP-2 captions without any refinement. A marginal improvement is observed when cleaning up the captions, as outlined in Section 4. The improvement, though limited, becomes more pronounced with the incorporation of multiple spatial scales, although it does not yet match the performance achieved with the template-based captioning approach. This observation can be attributed to the non-negligible noise present in the images, leading to captions with inherent noise. Consequently, refining the captions by incorporating accurate age, gender, and action attributes yields a remarkable enhancement in performance.

Comparing the refined BLIP-2 captions with the templated textual descriptions demonstrates the effectiveness of generated BLIP-2 captions in capturing additional information from the scene beyond the "***person_des***" and the "***action_des***" attributes used in our templates. The integration of the manual prompting procedure with refined captions proves to be the most effective approach, surpassing the performance of all investigated captioning techniques in our work.

5.5.5. Per-action evaluation

As shown in Fig. 5, the most frequent action in the LOKI dataset is ***"Moving"***, which accounts for more than half of the observed actions, leading to significant data imbalance. To validate our model's ability to anticipate rare actions, we evaluate it with respect to the individual action classes. As indicated in Table 11, our model achieves stable performance across all anticipated actions. The highest performance is observed for the ***"Waiting to Cross"*** and ***"Stopped"*** actions, as these are the most recognizable and least likely to be confused with other behaviors. The lowest performance is for the ***"Other"*** category, which encompasses a variety of general activities, making it challenging to detect or anticipate.

5.6. Computational analysis

The inclusion of a captioning process in our anticipation model adds more complexity, affecting the computational and inference time. Table 12 reports the execution time consumed by the different modules of textual feature extraction tested during our experimental evaluation. Relying on the commonly applied process of image feature extraction, in our case using VGG16, as a reference point, the utilization of CLIP to extract both image features and textual features reduces the computational time by almost 58%. However, generating the textual captions using the pre-trained BLIP2 model increases the time by a considerable amount. This observation is reasonable with respect to the relatively large size of such generative models. Therefore, from a real-time application perspective, depending on either the pre-defined captioning technique or the template-based captioning technique is more feasible.

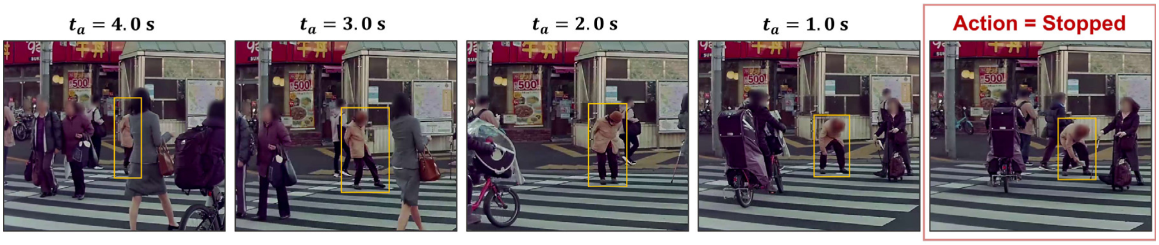
5.7. Qualitative examples

Fig. 7 presents a detailed case study on the generated BLIP-2 captions with their refinement process. In this example, we have a scenario where the pedestrian in question was initially crossing the street but was interrupted to pick something up and stopped. The template captions take advantage of the *age/gender/action* attributes at each frame, offering precise yet limited descriptions restricted to these attributes. In contrast, the generated BLIP-2 captions provide broader descriptions but are vulnerable to generating errors or conveying misleading information. For instance, in Fig. 7, due to the similarity between our pedestrian's body movements and those of someone engaged in skating, the captioner erroneously describes a skating activity in the images. Therefore, the incorporation of multi-scale information proves to be a necessity to capture a broader and more accurate understanding of the scene. The wider contexts in larger spatial scales help the captioner align more accurately with urban scenarios activities.

Moreover, the generated captions have the potential to include more general yet valuable information, describing possible interactions within the scene, such as "***A group of people walking down the street***". In the refinement process, new captions are generated and integrated with the original ones, as detailed in Section 4. These new captions aim to leverage the existing context in the original descriptions but are skewed more towards the accurate *age/gender/action* attribution. While the new captions might inherit some noisy contexts from the originals, they generally contribute to more accurate prompting in most cases, as evidenced by the enhanced performance outlined in Table 10.

Table 10
Studying the performance of the image captioning throughout the different levels of prompting and refinement. **TC** is the template-based captioning technique, and the **IC** reflects the usage of image captioner. For the image capioner, we define the refinement levels as: **CU** denotes the cleanup process of the captions, **MS** represents the multi-spacial-scale fusion, and **AR** is the *age/gender/action* attributes refinement process, as described in (6), but excluding the integration with the manually prompted captions $C^{1:P}_{temp}$. Finally, **MP** denotes the application of the manual prompts in either template-based captions or with the generated BLIP-2 captions.

Cap	CU	MS	AR	MP	4 s		3 s		2 s		1 s	
					AUC	F1	AUC	F1	AUC	F1	AUC	F1
TC	–	–	–	✗	70.10	64.51	70.08	65.14	72.98	70.10	80.80	77.68
TC	–	–	–	✓	77.30	71.05	80.86	75.95	84.64	80.92	89.67	87.40
IC	✗	✗	✗	✗	55.62	55.74	56.53	56.7	56.47	58.36	58.81	57.84
IC	✓	✗	✗	✗	55.65	55.73	56.53	57.27	56.2	58.5	61.15	57.27
IC	✓	✓	✗	✗	56.34	57.65	57.96	58.18	61.28	57.97	63.98	60.39
IC	✓	✓	✓	✗	77.25	71.4	78.82	75.56	85.22	**81.93**	88.9	87.11
IC	✓	✓	✓	✓	**77.77**	**72.16**	**80.95**	**75.98**	**85.24**	81.34	**90.16**	**87.92**



	Template Captions	BLIP2 Captions			
		1.5x context	Refined captions	Higher scale context	Refined captions
$t_a = 4.0$s	an adult female is crossing the road	a woman in a tan top	a woman in a tan top	a woman is walking across a crosswalk	a woman is crossing across a crosswalk
$t_a = 3.0$s	an adult female is crossing the road	a group of people are walking down the street	a woman is crossing down the street	a woman is crossing the street	a woman is crossing the street
$t_a = 2.0$s	an adult female is crossing the road	a man is pushing a woman on a skateboard	a woman is crossing on a skateboard	a man is running across the street	a woman is crossing across the street
$t_a = 1.0$s	an adult female is stopped	a man is standing on a skateboard	a woman is stopped on a skateboard	a man is walking across the crosswalk	a woman is stopped across the crosswalk

Fig. 7. A qualitative example of image captioning and the refinement process.

Table 11
Per-action evaluation.

	4 s		3 s		2 s		1 s	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Overall	77.77	72.16	80.95	75.98	85.24	81.34	90.16	87.92
Moving	71.21	78.16	78.30	81.60	81.40	85.31	90.41	91.66
Stopped	88.85	82.50	87.41	82.18	90.09	85.82	92.32	89.39
Waiting to Cross	86.96	80.48	88.45	84.11	92.67	89.31	96.00	94.12
Crossing	76.23	59.71	82.06	67.17	86.81	72.51	92.25	82.65
Other	66.12	43.04	68.68	48.56	74.38	57.14	79.22	66.80

Table 12
Execution time analysis of the different techniques utilized to extract features.

	VGG16	CLIP_{img}	CLIP_{text}	BLIP2 (Single Image)	BLIP2 (Multi-Scale)
Execution time (s)	0.031	0.013	0.007	0.16	0.69

6. Conclusion

In this work, we propose the integration of a language modality into pedestrian action anticipation. We study multiple textual generation approaches and tools, proving the effectiveness of coupling vision-language features to enrich the understanding of visual scenes and enhance anticipation performance. Additionally, we extended the binary crossing or not crossing pedestrian anticipation task into anticipating multiple actions, evaluating our task on a novel large-scale urban scenes dataset (LOKI). Our evaluation and ablation experiments demonstrate the out-performance of our language-aided anticipation approach, with an improvement over that reaches 29.5% F1-score at 1-s anticipation and 16.66% at 4-s anticipation.

Generating a textual description for an image is a fast-advancing field where more informative and accurate captions could be generated. For example, answering vision questions could guide the image captioners in furnishing tailored captions in response to posed questions, opening new dimensions for anticipatory understanding and leading to improved performance. Therefore, our future work will focus on utilizing such frameworks, aiming to provide more precise captions and better anticipations.

CRediT authorship contribution statement

Nada Osman: Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization. **Guglielmo Camporese:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Conceptualization. **Lamberto Ballan:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support. This research was also supported by an UniPD BIRD-2021 Project and by the PREVUE project from the PRIN National Research Program, MUR, Italy (CUP E94I19000650001).

References

Abu Farha, Y., Richard, A., Gall, J., 2018. When will you do what? - anticipating temporal occurrences of activities. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Achaji, L., Moreau, J., Fouqueray, T., Aioun, F., Charpillet, F., 2022. Is attention to bounding boxes all you need for pedestrian action prediction? In: Proc. of the IEEE Intelligent Vehicles Symposium. IV, pp. 895–902.

Burns, A., Tan, R., Saenko, K., Sclaroff, S., Plummer, B.A., 2019. Language features matter: Effective language representations for vision-language tasks. In: Proc. of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 7474–7483.

Camporese, G., Bergamo, A., Lin, X., Tighe, J., Modolo, D., 2023. Early action recognition with action prototypes. arXiv arXiv:2312.06598.

Camporese, G., Coscia, P., Furnari, A., Farinella, G., Ballan, L., 2021. Knowledge distillation for action anticipation via label smoothing. In: Proc. of the IAPR International Conference on Pattern Recognition. ICPR.

Correia, J., Moreno, P., Avelino, J., 2022. Pedestrian intention anticipation with uncertainty based decision for autonomous driving. In: Proc. of the IEEE International Conference on Robotic Computing. IRC.

Das, S., Ryoo, M.S., 2022. Video+ clip baseline for ego4d long-term action anticipation. arXiv preprint arXiv:2207.00579.

Fan, C., Lee, J., Ryoo, M.S., 2018. Forecasting hands and object locations in future frames. In: Proc. of the European Conference on Computer Vision Workshops.

Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. Slowfast networks for video recognition. In: Proc. of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 6202–6211.

Felsen, P., Agrawal, P., Malik, J., 2017. What will happen next? Forecasting player moves in sports videos. In: Proc. of the IEEE/CVF International Conference on Computer Vision. ICCV.

Floridi, L., Chiriatti, M., 2020. GPT-3: Its nature, scope, limits, and consequences. Minds Mach. 30, 681–694.

Furnari, A., Battiato, S., Grauman, K., Farinella, G., 2017. Next-active-object prediction from egocentric videos. J. Vis. Commun. Image Represent. 49, 401–411.

Furnari, A., Farinella, G., 2021. Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. IEEE Trans. Pattern Anal. Mach. Intell. 43 (11), 4021–4036.

Gao, J., Yang, Z., Nevatia, R., 2017. RED: Reinforced Encoder-Decoder Networks for Action Anticipation. In: Proc. of the British Machine Vision Conference. BMVC.

Gesnouin, J., Pechberti, S., Stanciulcscu, B., Moutarde, F., 2021. TrouSPI-Net: Spatio-temporal attention on parallel atrous convolutions and U-GRUs for skeletal pedestrian crossing prediction. In: Proc. of the IEEE International Conference on Automatic Face and Gesture Recognition.

Ghosh, S., Aggarwal, T., Hoai, M., Balasubramanian, N., 2022. Distilling knowledge from language models for video-based action anticipation. arXiv preprint arXiv:2210.05991.

Girase, H., Gang, H., Malla, S., Li, J., Kanehara, A., Mangalam, K., Choi, C., 2021. Loki: Long term and key intentions for trajectory prediction. In: Proc. of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 9803–9812.

Hu, X., Dai, J., Li, M., Peng, C., Li, Y., Du, S., 2022. Online human action detection and anticipation in videos: A survey. Neurocomputing 491, 395–413.

Hu, R., Singh, A., 2021. Unit: Multimodal multitask learning with a unified transformer. In: Proc. of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 1439–1449.

Huang, D., Hilliges, O., Van Gool, L., Wang, X., 2023. Palm: Predicting actions through language models@ ego4d long-term action anticipation challenge 2023. arXiv preprint arXiv:2306.16545.

Kotseruba, I., Rasouli, A., Tsotsos, J.K., 2020. Do they want to cross? Understanding pedestrian intention for behavior prediction.. In: Proc. of the IEEE Intelligent Vehicles Symposium. IV.

Kotseruba, I., Rasouli, A., Tsotsos, J.K., 2021. Benchmark for evaluating pedestrian action prediction. In: Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision. WACV.

Li, J., Li, D., Savarese, S., Hoi, S., 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597.

Li, J., Li, D., Xiong, C., Hoi, S., 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proc. of the International Conference on Machine Learning. ICML, pp. 12888–12900.

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Liu, B., Adeli, E., Cao, Z., Lee, K.-H., Shenoi, A., Gaidon, A., Niebles, J.C., 2020. Spatiotemporal relationship reasoning for pedestrian intent prediction. IEEE Robot. Autom. Lett. 5 (2), 3485–3492.

Liu, J., Li, Y., Song, S., Xing, J., Lan, C., Zeng, W., 2018. Multi-modality multi-task recurrent neural network for online action detection. IEEE Trans. Circuits Syst. Video Technol. 29 (9), 2667–2682.

Lorenzo, J., Alonso, I.P., Izquierdo, R., Ballardini, A.L., Saz, Á.H., Llorca, D.F., Sotelo, M.Á., 2021. CAPformer: pedestrian crossing action prediction using transformer. Sensors 21 (17).

Lorenzo, J., Parra, I., Wirth, F., Stiller, C., Llorca, D.F., Sotelo, M.Á., 2020. RNN-based Pedestrian Crossing Prediction using Activity and Pose-related Features. In: Proc. of the IEEE Intelligent Vehicles Symposium. IV.

Lu, J., Batra, D., Parikh, D., Lee, S., 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proc. of Advances in Neural Information Processing Systems. NeurIPS.

Mahmud, T., Hasan, M., Roy-Chowdhury, A., 2017. Joint prediction of activity labels and starting times in untrimmed videos. In: Proc. of the IEEE/CVF International Conference on Computer Vision. ICCV.

Manousaki, V., Bacharidis, K., Papoutsakis, K., Argyros, A., 2023. VLMAH: Visual-linguistic modeling of action history for effective action anticipation. In: Proc. of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 1917–1927.

Nah, S., Kim, T., Lee, K., 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Neogi, S., Hoy, M., Dang, K., Yu, H., Dauwels, J., 2020. Context model for pedestrian intention prediction using factored latent-dynamic conditional random fields. IEEE Trans. Intell. Transp. Syst. 22 (11), 6821–6832.

Niu, Y., Lu, Z., Wen, J.-R., Xiang, T., Chang, S.-F., 2019. Multi-modal multi-scale deep learning for large-scale image annotation. IEEE Trans. Image Process. 28 (04), 1720–1731.

Osman, N., Camporese, G., Ballan, L., 2023. TAMformer: Multi-modal transformer with learned attention mask for early intent prediction. In: Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP.

Osman, N., Camporese, G., Coscia, P., Ballan, L., 2021. SlowFast Rolling-Unrolling LSTMs for action anticipation in egocentric videos. In: Proc. of the IEEE/CVF International Conference on Computer Vision Workshops.

Osman, N., Cancelli, E., Camporese, G., Coscia, P., Ballan, L., 2022. Early pedestrian intent prediction via features estimation. In: Proc. of the IEEE International Conference on Image Processing. ICIP, pp. 3446–3450.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: Proc. of the International Conference on Machine Learning. ICML, pp. 8748–8763.

Rasouli, A., Kotseruba, I., 2023. PedFormer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning. In: Proc. of the IEEE International Conference on Robotics and Automation. ICRA.

Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J.K., 2019. PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. In: Proc. of the IEEE/CVF International Conference on Computer Vision. ICCV.

Rasouli, A., Kotseruba, I., Tsotsos, J.K., 2017. Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior. In: Proc. of the IEEE/CVF International Conference on Computer Vision Workshops.

Rasouli, A., Rohani, M., Luo, J., 2021. Bifold and semantic reasoning for pedestrian behavior prediction. In: Proc. of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 15600–15610.

Rasouli, A., Yau, T., Rohani, M., Luo, J., 2022. Multi-modal hybrid architecture for pedestrian action prediction. In: Proc. of the IEEE Intelligent Vehicles Symposium. IV, pp. 91–97.

Razali, H., Mordan, T., Alahi, A., 2021. Pedestrian intention prediction: A convolutional bottom-up multi-task approach. Transp. Res. C 130, 103259.

Rhinehart, N., Kitani, K.M., 2020. First-person activity forecasting with online inverse reinforcement learning. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 304–317.

Sener, F., Singhania, D., Yao, A., 2020. Temporal aggregate representations for long term video understanding. arXiv arXiv:2006.00830.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: Proc. of the International Conference on Learning Representations. ICLR.

Wang, Y., 2021. Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 17 (1s), 1–25.

Wu, Z., Xiong, C., Ma, C.-Y., Socher, R., Davis, L.S., 2019. AdaFrame: Adaptive Frame Selection for Fast Video Recognition. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Yang, B., Wei, Z., Hu, H., Wang, R., Yang, C., Ni, R., 2023. DPCIAN: A novel dual-channel pedestrian crossing intention anticipation network. IEEE Trans. Intell. Transp. Syst..

Yang, D., Zhang, H., Yurtsever, E., Redmill, K., Özgüner, Ü., 2021. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. arXiv preprint arXiv:2104.05485.

Zeng, K.-H., Shen, W., Huang, D.-A., Sun, M., Niebles, J.C., 2017. Visual forecasting by imitating dynamics in natural sequences. In: Proc. of the IEEE/CVF International Conference on Computer Vision. ICCV.

Zhang, G., Luo, Z., Tian, Z., Zhang, J., Zhang, X., Lu, S., 2023. Towards efficient use of multi-scale features in transformer-based object detectors. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Zhang, M., Ma, K., Lim, J., Zhao, Q., Feng, J., 2017. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 4372–4381.

Zhao, Q., Zhang, C., Wang, S., Fu, C., Agarwal, N., Lee, K., Sun, C., 2024. AntGPT: Can large language models help long-term action anticipation from videos? In: Proc. of the International Conference on Learning Representations. ICLR.