Referring Atomic Video Action Recognition

Kunyu Peng[1,*], Jia Fu[3,4,*], Kailun Yang[2,†], Di Wen[1], Yufan Chen[1], Ruiping Liu[1], Junwei Zheng[1], Jiaming Zhang[1], M. Saquib Sarfraz[1,6], Rainer Stiefelhagen[1], and Alina Roitberg[5]

[1]Karlsruhe Institute of Technology, [2]Hunan University,
[3]RISE Research Institutes of Sweden, [4]KTH Royal Institute of Technology,
[5]University of Stuttgart, [6]Mercedes-Benz Tech Innovation

Abstract. We introduce a new task called **R**eferring **A**tomic **V**ideo **A**ction **R**ecognition (RAVAR), aimed at identifying atomic actions of a particular person based on a textual description and the video data of this person. This task differs from traditional action recognition and localization, where predictions are delivered for all present individuals. In contrast, we focus on recognizing the correct atomic action of a specific individual, guided by text. To explore this task, we present the RefAVA dataset, containing 36,630 instances with manually annotated textual descriptions of the individuals. To establish a strong initial benchmark, we implement and validate baselines from various domains, *e.g.*, atomic action localization, video question answering, and text-video retrieval. Since these existing methods underperform on RAVAR, we introduce `RefAtomNet` – a novel cross-stream attention-driven method specialized for the unique challenges of RAVAR: the need to interpret a textual referring expression for the targeted individual, utilize this reference to guide the spatial localization and harvest the prediction of the atomic actions for the referring person. The key ingredients are: (1) a multi-stream architecture that connects video, text, and a new location-semantic stream, and (2) cross-stream agent attention fusion and agent token fusion which amplify the most relevant information across these streams and consistently surpasses standard attention-based fusion on RAVAR. Extensive experiments demonstrate the effectiveness of `RefAtomNet` and its building blocks for recognizing the action of the described individual. The dataset and code will be made publicly available at RAVAR.

1 Introduction

Referring scene understanding [50, 52, 66, 93] aims to solve the underlying computer vision task for the particular scene element specified via a natural language *referring expression*. Context-driven queries are vital in information retrieval, assistive systems, and multimedia analysis, but incorporating them into computer vision models is challenging due to a complex entanglement of linguistic descriptions, localization, and visual recognition itself. Several benchmarks for referring scene understanding have been successfully established for multi-object

* Equal contribution
† Correspondence: kailun.yang@hnu.edu.cn

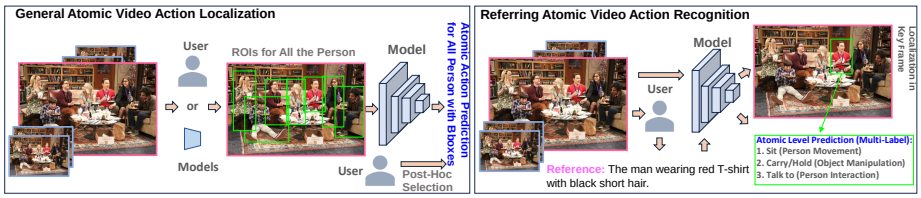


Fig. 1: A comparison of AAL Task (left) and RAVAR task (right).

tracking [86], semantic segmentation [30, 46, 72, 75, 76], medical imaging [71], and object detection [12, 64].While these efforts focus on object-centric referring expressions, many applications require a human-centric understanding of human actions, *e.g.*, in rehabilitation assistance [35, 70] and human-robot interaction [27,37]. Our work addresses a new problem – reference-driven recognition of atomic actions performed by the human described via a textual reference, which has been overlooked in the past.

The field of human action recognition has made exciting advances, from new approaches stemming from the rise of visual transformers [23,47,63,69,81,83,85] to new large-scale datasets [4, 22, 34, 73, 78]. Datasets utilized for the prediction of atomic actions often feature recordings with multiple individuals [24]. Nevertheless, the vast majority of previously published works [23, 31, 68, 69, 81] either rely on manually window crop of a particular person of interest or automatically generated region of interests to predict the atomic actions for all the individuals in one video, which results in large pre- or post-processing effort to focus on one specific person. This workflow is also very inconvenient in certain applications, such as assistive systems for users with visual impairments [51,62,97], who must understand the state of each present in the scene for effective interactions. Leveraging succinct textual descriptions, which may include broad positional indicators (*e.g.*, left, center, and right), appearance attributes (*e.g.*, hair color and clothing), or gender, to steer the model towards delivering atomic action recognition outcomes for the targeted individual, presents a promising solution to the outlined challenges. Employing these textual cues as a reference, the model can facilitate an end-to-end retrieval of the specified instance across the entire video, subsequently providing precise atomic action recognition results.

To close this gap, we formalize the new task of **R**eferring **A**tomic **V**ideo **A**ction **R**ecognition (RAVAR). The differences between RAVAR and atomic action localization are shown in Fig. 1. The conventional atomic action recognition in multi-person scenarios is often formalized in an action localization manner, and requires the region of interest to localize the human available a priori, after which atomic action recognition is carried out for each individual. Post-processing is required for individuals of interest. In contrast, RAVAR assumes a textual reference and a video as inputs, delivering the referring individual's atomic actions along with the location. This comparison shows the efficiency of RAVAR, leveraging textual information for less labor-intensive subject identification to assist atomic action analysis for the individual of interest. We further

introduce the new RefAVA dataset for RAVAR. RefAVA is established using the 17,946 video clips of the public AVA dataset [24] initially designed for atomic action localization, which we extend with 36,630 textual descriptions of the corresponding individuals used as the referring expression. The dataset covers diverse settings, *e.g.*, in/out door and day/night time scenarios, and numerous multi-person scenes, constituting an ideal testbed for our task.

To establish an initial benchmark, we assess 15 well-established approaches from multiple related research domains. Our experiments reveal that none of the benchmarked methods delivers sufficient performance. For example, previous methods within the domains of VQA and VTR typically yield predictions of a coarse analysis regarding actions, whereas strategies in AAL face challenges in spatially localizing the correct referring individual. We attribute this phenomenon to the massive irrelevant information provided by the visual data, distracting the model from the referring location. How to suppress the referring irrelevant information becomes a critical challenge to pursue better RAVAR performances.

To address this, we introduce the new `RefAtomNet` approach specialized in simultaneously solving the tasks of understanding textual `ref`erring expressions and using them as guides for spatial localization and classification of `atom`ic actions. We begin by computing textual, visual, and newly proposed *location-semantic tokens* using a large foundation model and well-established object detector. The latter stream is designed to harvest semantic cues of different scene entities: it leverages a pre-trained object detector to localize and identify objects, after which the location-semantic tokens are computed as a fusion of visual bounding box coordinates and text embeddings of the object category. Another novel aspect of our work is the introduction of *cross-stream agent attention and agent token fusions*, which builds on the concept of agent attention [26] and re-defines it for cross-stream compatibility: agent tokens are reformulated into a 1D sequence format through the use of fully connected layers, bypassing the initial need for 2D pooling, removing the depthwise convolution branch, and forgoing 2D positional encodings. This mechanism enhances the model's ability to filter and emphasize relevant information from multiple data sources effectively. `RefAtomNet` achieves the highest performance for RAVAR, while a detailed ablation study demonstrates the effectiveness of each component.

Our contributions can be summarized as follows:

- We introduce the new **R**eferring **A**tomic **V**ideo **A**ction **R**ecognition (RAVAR) task and the RefAVA benchmark with 36,630 atomic action instances manually annotated with suitable language expressions.
- To establish a competitive benchmark, we examine 11 well-established models from different related fields, namely atomic action localization, video question answering, and video-text retrieval.
- We propose `RefAtomNet`, a new RAVAR approach that uses a novel agent-based semantic-location aware attentional fusion to integrate multi-modal tokens, suppressing the irrelevant visual cues. `RefAtomNet` delivers improve-

ments of 3.85% and 3.17% of mAP, and 4.33% and 4.03% of AUROC, on val and test sets, respectively, compared with the leading baseline, BLIPv2 [42].

2 Related Work

Referring Scene Understanding. Referring scene understanding aims to locate parts of interest within images or videos guided by natural language, exhibiting the utility in many computer vision tasks including autonomous driving [86] and video editing [6]. The development of this field cannot be advanced without the contribution of high-quality open-source datasets and benchmarks [2, 12, 30, 48, 72, 80, 86, 91]. For example, the CLEVR-Ref+ benchmark is proposed by Liu *et al*. [52] to achieve visual reasoning with referring expressions. Li *et al*. [46] proposed referring image segmentation by using a recurrent refinement network. Wu *et al*. [86] proposed referring multi-object tracking benchmark. However, there are no referring understanding works focusing on atomic video action analysis, whereupon we first conduct RAVAR benchmarking with our RefAVA dataset. We further clarify that Referring Video Object Segmentation (RVOS) task [21,53,58,79] differs us from multiple perspectives, *e.g.*, RVOS includes action name in input textual reference.

Video Text Retrieval. Video Text Retrieval (VTR) aims at matching relevant video content with text. Amidst the rise of Vision-Language foundation models like CLIP [67] and BLIP [43], efforts [9,42,49,55–57,77,82,87,95] are being made to apply powerful pre-trained models for promoting their competencies in VTR. XCLIP [56] introduced multi-grained contrastive learning for end-to-end VTR, by counting all the video-sentence, video-word, sentence-frame, and frame-word contrasts. BLIPv2 [42] pre-trained a Querying Transformer (QFormer) to bootstrap vision-and-language representation learning and vision-to-language generative learning. The key difference between VTR and RAVAR lies in their inference tasks: VTR queries text or video to localize elements, while RAVAR integrates textual and visual data for fine-grained subject retrieval and atomic action prediction in videos.

Video Question Answering. Video Question Answering (VQA) focuses on generating answers to questions posed in natural language for a given video. Depending on the emphasis of the question, factoid VQA [1,5,7,19,20,25,28,36,39, 45,54,88,89,96] straightforward queries visual facts. Lei *et al.* [38] proposed that a single-frame trained transformer-based model, with large-scale pre-training and a frame ensemble at the inference stage, can perform better than existing multi-frame trained models in factoid VQA tasks. On the other hand, inference VQA [18, 41] delves into logical reasoning. Li *et al.* [44] released a video-centric instruction dataset and leveraged a neural interface to integrate video foundation models and Large Language Models (LLM), showcasing capability in temporal reasoning, causal inference, and event localization. In addition, some multimodal VQA frameworks [20, 40, 90] explore information-invoking scenarios that incorporate visual, audio, subtitle, and external knowledge. However, compared with the RAVAR task, most of the existing VQA approaches do not particularly focus

Table 1: An overview of the referring scene understanding datasets and our RefAVA dataset; Our task is Referring Atomic Video Action Recognition (RAVAR), whereas existing benchmarks focus on Referring Object Detection (ROD), Referring Video-based Object Segmentation (RVOS), and Referring Multi-Object Tracking (RMOT).

Dataset	RefCOCO [91]	RefCOCO+ [91]	RefCOCOg [91]	Talk2Car [13]	VID-Sentence [10]	Refer-DAVUS17 [30]	Refer-YV [94]	Refer-KITTI [86]	**RefAVA**
Task	ROD	ROD	ROD	ROD	ROD	RVOS	RMOT	RMOT	**RAVAR**
$\mathbf{N}_{Frames}$	26,711	19,992	26,711	9,217	59,238	4,219	93,869	6,650	**1,615,140**
$\mathbf{N}_{Instance}$	26,711	19,992	26,711	10,519	7,654	3,978	7,451	-	**36,630**

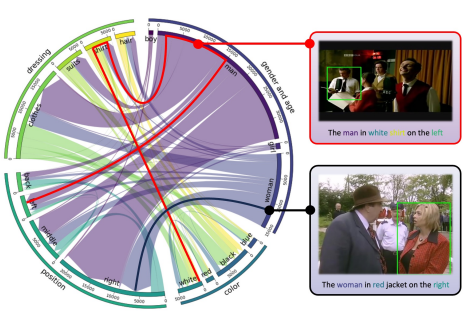
on atomic action recognition and will only tend to give a coarse textual description, which limits the applications requiring precise prediction, *e.g.*, human-robot assistance [27, 37].

Atomic Video Action Recognition and Localization. Atomic video-based action recognition [11, 23] and localization [24] involves the identification and analysis of the most fundamental, indivisible actions or movements performed by humans for single and multiple-person scenarios. Compared with the general video action recognition task, atomic video-based action localization is much more fine-grained and is always formulated in a multi-label manner with bounding box predictions. Most of the existing convolutional neural networks (CNN) [4, 16, 17, 81] and the transformer [23, 47, 63, 69, 81, 83, 85] networks for human action recognition are commonly used in the atomic video-based action localization by changing the classification head into multi-label manner, adding additional bounding box prediction head, and integrating region of interest features from human detector. Ryali *et al.* [69] proposed a hierarchical vision transformer with high efficiency and precision within the realm of the existing methods using video as input. Wang *et al.* [81] proposed scaling video-masked autoencoders with dual masking. Current methods for predicting individual actions in multi-person scenarios often require manual video cropping for atomic action recognition or generate predictions for all detected individuals, necessitating further human selection and reducing practicality [65, 68, 84]. Most existing AAL methods are not specifically designed for the RAVAR task. We thereby introduce `RefAtomNet`, a novel method that utilizes location semantics atop predicted location and scene semantic information derived from the scenario together with the textual reference and visual cues, then further uses cross-stream agent attention and agent token fusions to suppress redundant information.

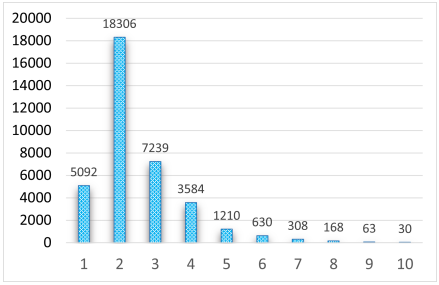
3 RAVAR: Established Benchmark

3.1 Introduction of the RefAVA Dataset

Textual Annotations. To acquire precise textual annotations for the individuals of interest, 7 annotators manually provided the textual annotations according to the key frame bounding boxes presented in the AVA dataset [24]. Cross-checking among all the annotators for the annotated individuals is conducted to deliver high textual annotation quality.



(a) Chord visualization of several keywords.



(b) Annotation statistics: #The number of annotated persons contained in the scenario (x-axis); #The number of instances (y-axis).

Fig. 2: An chord visualization of several keywords from the textual references, shown on the left, and the instance amount in video clips containing different numbers of annotated persons in our RefAVA dataset, shown on the right.

Dataset. We selected 17,946 video clips from 127 movies of the AVA dataset, preserving the most complex scenarios, and annotated each person on each center frame of every 90s video clip. In total, RefAVA has 36,630 labeled instances, and we split them into 22,658 train instances, 10,916 validation instances, and 3,056 test instances, where the samples from different sets are from different movie scenarios. The textual annotations cover information on approximate age, gender, appearance, relative position in the center frame, *etc.*, and without action description. The chord visualization in Fig. 2a reveals the quantitative causal relations between the most frequent notional words in different categories from the textual references. Further, Fig. 2b shows the statistics of instance amount towards the annotated person number inside the scenario where each instance belongs. We deliver the comparison of the statistics between representative existing referring scene understanding datasets and our RefAVA dataset in Tab. 1. The atomic actions involve 80 categories covering Object Manipulation (OM), Person Interactions (PI), and Person Movement (PM). The videos cover diverse scenarios. The test set is sourced from 26 movies different from those in the train set (67 different movies) and val set (34 different movies) to achieve the evaluation of generalizability.

3.2 The Baselines for the RAVAR Benchmark

We have adapted methodologies from similar fields as our baselines for RAVAR.
AAL Baselines. We first reformulate the existing methods from the general atomic video action localization field by integrating the textual reference embeddings into the visual branch. Some action recognition approaches, *e.g.*, I3D [4] and X3D [16], are adapted to AAL following [17]. BERT [14] is used for textual embedding extraction. The selected baselines from this domain can be grouped into CNN-based approaches [4,16] and transformer-based approaches [47,69,81]. All these approaches leverage the pre-trained weight on the Kinetics400 [4].

VQA Baselines. The second group of approaches comes from the video question answering domain, which contains the GPT-based model, *i.e.*, AskAnything [44], and conventional transformer-based model [38]. We input the reference sentence in a questioning manner and add a classification head and a bounding box regression head after the acquisition of the logits. We finetuned the two models for the RAVAR task based on their pre-trained weight on ActivityNetQA [92].
VTR Baselines. The last group of the foundation model baselines comes from the video-text retrieval task, where XCLIP [56], CLIP4CLIP [55], BLIPv2 [42], and MeVTR [95] are selected. These foundation models are pre-trained on a combination of numerous datasets, incorporating Conceptual Captions [74], SBU Captions [61], and COCO Captions [8], *etc.* We reformulate these approaches in the same way as the approaches in VQA.
SF Baselines. Some Single Frame (SF) baselines are adopted, where SAM [33], DETR [3], and REFCLIP [29] are utilized together with the CLIP [67] feature extraction backbone.
VOS Baseline. We adopt one Video Object Segmentation (VOS) baseline using the encoder of the approach from Su *et al.* [79] and our prediction head.

4 RefAtomNet: Proposed Method

4.1 Overview

We propose a new model, `RefAtomNet`, with an overview provided in Fig. 3. It leverages three token streams: visual, textual reference, and location-semantic streams. Visual tokens are extracted from the video using ViT [15], while textual reference tokens are obtained from the text using BERT [14]. Both streams are enhanced by QFormer [42]. Additionally, in the location-semantic stream, location-semantic tokens are extracted from the center frame by fusing predicted object coordinates and semantic embeddings of the object categories estimated by a frozen object detector and BERT [14]. To suppress irrelevant information for each stream, we propose the agent-based location-semantic aware attentional fusion to achieve better amplification of relevant information during the cross-stream exchange. The detailed structure is discussed in the following subsections.

4.2 RefAtomNet

Background of QFormer. We rely on BLIPv2 to extract the visual features, where QFormer is the most essential component. Multiple learnable queries are initialized in QFormer as trainable parameters and interact with input data via Transformer attention mechanisms. Each query selectively attends to input parts, capturing task-specific features and updating based on attention outputs. During training, these queries are optimized for specialized information extraction. Queries then directly contribute to generating outputs.
Extraction of Visual and Textual Reference Tokens. Similar to BLIPv2 [42], we use a pre-trained multimodal model for token extraction from both the video

and textual reference streams. This model leverages the QFormer [42] architecture to effectively extract and integrate multimodal embeddings from the visual and textual reference data. Specifically, the visual stream employs a ViT [15] backbone encoder to process visual inputs, while the textual branch utilizes BERT [14] to encode textual reference cues. This combined approach facilitates a holistic understanding of both visual content and textual descriptions. The model extracts visual tokens ($\mathbf{t}^{VT}$) and textual reference tokens ($\mathbf{t}^{RT}$) through Eq. 1:

$$\mathbf{t}^{VT}, \mathbf{t}^{RT} = \mathcal{V}_{VL}(\mathcal{V}_{VT}(\mathbf{x}^{VT}),\ \mathcal{V}_{RT}(\mathbf{x}^{RT})), \tag{1}$$

where $\mathcal{V}_{VL}$ represents the visual-textual integration model (*i.e.*, QFormer [42]). $\mathcal{V}_{VT}$, and $\mathcal{V}_{RT}$ denote the backbones for extracting visual tokens (*i.e.*, ViT [15]) and textual reference tokens (*i.e.*, BERT [14]), respectively. $\mathbf{x}^{VT}$ and $\mathbf{x}^{RT}$ represent the input video and textual reference caption, which are then fed into liner projection layers, *i.e.*, $\mathbf{P}_{VT}$ and $\mathbf{P}_{RT}$, respectively.

Extraction of Location-Semantic Aware Tokens. To integrate more location and semantic information into the token representations, we leverage a well-established object detector DETR [3] to deliver N_o detection results based on the input of the center frame of a video clip, noted as the keyframe. Note that, the detections provided by the object detector contain either the human or other objects with a high confidence score according to Eq. 2.

$$\mathbf{r}_{boxes}, \mathbf{r}_{cats} = \mathcal{V}_{dets}(\mathbf{x}_k), \tag{2}$$

where $\mathbf{x}_k$, $\mathcal{V}_{dets}$ indicates the keyframe and the detection network. $\mathbf{r}_{boxes} \in \mathbb{R}^{N_o \times 4}$ are detected 2D corner coordinates (top left and bottom right corners) of the bounding boxes. $\mathbf{r}_{cats}$ is the predicted category for each bounding box represented in textual format. These category labels of the detected objects are passed to a text encoder (BERT [14]) to extract semantic embeddings. We concatenate these two types of tokens together along the channel dimension and use a single linear projection layer to obtain the aggregated location-semantic aware tokens ($\mathbf{t}^{LS}$) as Eq. 3.

$$\mathbf{t}^{LS} = \mathbf{P}_{LS}(Concat\,[\mathcal{V}_{RT}(\mathbf{r}_{cats}), \mathbf{r}_{boxes}]), \tag{3}$$

where $\mathbf{P}_{LS}$ indicates a fully connected layer, $\mathcal{V}_{RT}$ indicates the language feature extraction backbone, *i.e.*, BERT [14], and *Concat* indicates the concatenation.

Agent-Based Location Semantic Aware Attentional Fusion.

Agent Tokens and Sequential Processing. We use *agent tokens*, inspired by agent attention [26]. Derived through fully connected layers from input tokens, similar to Query, Key, and Value extraction in transformers, agent tokens aggregate essential information via agent-key and agent-query pairs. To enable agent attention for 1D sequential tokens from our streams, we redefine the 2D agent spatial tokens proposed in [26] into a 1D sequential format for multi-stream fusion. Specifically, we replace 2D pooling with linear projection and eliminate the depthwise convolutional branch and biased position encoding to better suit agent acquisition in our model. These 1D sequential agent tokens serve dual functions:

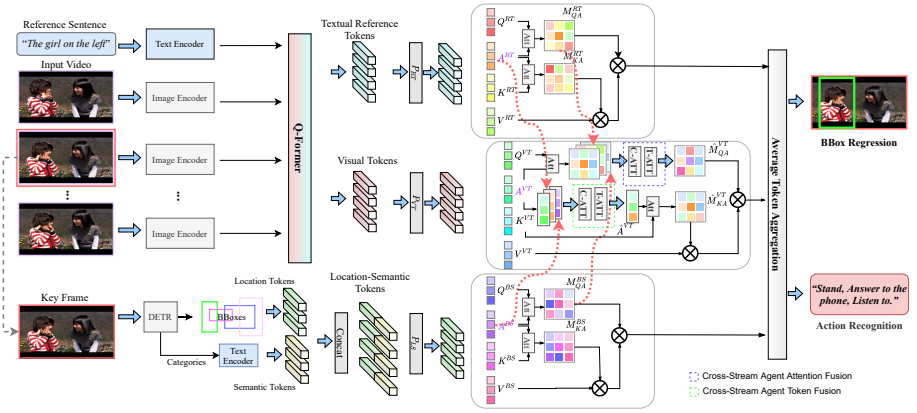


Fig. 3: An overview of the `RefAtomNet` architecture.

aggregating crucial information from agent-key and agent-query pairs and efficiently redistributing this information to the original value tokens, therefore, enhancing attention focus and reducing irrelevant cues.

Agent-Based Attention Mechanism. The agent-based attention mechanism for visual, textual reference, and location-semantic streams is computed as:

$$\mathbf{Q}^{\phi}, \mathbf{K}^{\phi}, \mathbf{V}^{\phi}, \mathbf{A}^{\phi} = \mathbf{W}_{Q}^{\phi}(\mathbf{t}^{\phi}), \mathbf{W}_{K}^{\phi}(\mathbf{t}^{\phi}), \mathbf{W}_{V}^{\phi}(\mathbf{t}^{\phi}), \mathbf{W}_{A}^{\phi}(\mathbf{t}^{\phi}), \tag{4}$$

where for better readability, we use ϕ to represent $[RT, VT, LS]$, where RT indicates the reference tokens, VT indicates the visual tokens, and LS indicates the location-semantic tokens. $\mathbf{Q}^{\phi}$, $\mathbf{K}^{\phi}$, $\mathbf{V}^{\phi}$, and $\mathbf{A}^{\phi}$ are the query, key, value, and agent tokens. t^{ϕ} depicts the input. $\mathbf{W}_{Q}^{\phi}$, $\mathbf{W}_{K}^{\phi}$, $\mathbf{W}_{V}^{\phi}$, and $\mathbf{W}_{A}^{\phi}$ are constructed by linear projection layers. We then project the agent tokens using fully connected layers $\mathbf{P}_{A}^{\phi}$ through $\mathbf{A}_{*}^{\phi} = \mathbf{P}_{A}^{\phi}(\mathbf{A}^{\phi})$ as aforementioned. To mitigate redundancy cues of the textual reference and location-semantic streams, we leverage agent attention for both of these two streams. The agent query attention mask $\mathbf{M}_{QA}^{\pi}$ and the agent key attention mask $\mathbf{M}_{KA}^{\pi}$ are obtained by matrix multiplication ($MatMul$) and SoftMax operations along the channel dimension (indicated by σ_c) as shown in Eq. 5, where $\pi \in [RT, LS]$ and α indicates a fixed scale factor. The agent attention masks and tokens are computed and refined to ensure that only pertinent information influences the attention mechanism, as shown in Eq. 5 and Eq. 6:

$$\mathbf{M}_{QA}^{\pi}, \mathbf{M}_{KA}^{\pi} = \sigma_c(MatMul[\alpha * \mathbf{A}_{*}^{\pi}, \mathbf{Q}^{\pi}]), \sigma_c(MatMul[\alpha * \mathbf{A}_{*}^{\pi}, \mathbf{K}^{\pi}]), \tag{5}$$

$$\mathbf{t}_{*}^{\pi} = FFN(MatMul[\mathbf{M}_{KA}^{\pi}, MatMul[\mathbf{M}_{QA}^{\pi}, \mathbf{V}^{\pi}]]), \tag{6}$$

where FFN indicates the Feed Forward Network. These steps ensure precise model attention, enhancing the contextual relevance of the resulting tokens.

Cross-Stream Agent Attention and Agent Token Fusions. Finally, cross-stream agent attention and agent token fusion are proposed, which are explicitly tailored for the visual stream by applying the agent query attention maps and the

agent tokens from the other two streams. This process involves recalculating the attention maps, thereby ensuring that the final token representation is highly relevant and contextually enriched. We compute the agent query attention map for the visual token stream, where $\gamma = VT$, as shown in Eq. 7,

$$\mathbf{M}_{QA}^{\gamma} = \sigma_c(MatMul\,[\alpha * \mathbf{A}_*^{\gamma}, \mathbf{Q}^{\gamma}]). \tag{7}$$

Then, we calculate the cross-stream irrelevance-suppressed attention for the agent query attention of the visual stream through Eq. 8 to achieve cross-stream agent attention fusion to suppress irrelevant information in the agent query attentions of the visual stream, the operations used in the second and third terms are abbreviated as C-ATT and T-ATT in Fig. 3,

$$\hat{\mathbf{M}}_{QA}^{\gamma} = AVG\left[\mathbf{M}_{QA}^{\gamma}, \sigma_c(\sum_{\pi}\mathbf{M}_{QA}^{\pi}) * \mathbf{M}_{QA}^{\gamma}, \sigma_t(\sum_{\pi}\mathbf{M}_{QA}^{\pi}) * \mathbf{M}_{QA}^{\gamma}\right], \tag{8}$$

where σ_t denotes the SoftMax operation along the token dimension. AVG indicates the mean operation. We follow the same procedure to calculate the cross-stream irrelevance-suppressed agent tokens before the calculation of the agent key attention, as in Eq. 9 to achieve cross-stream agent token fusion.

$$\hat{\mathbf{A}}_*^{\gamma} = AVG\left[\mathbf{A}_*^{\gamma} + \sigma_c(\sum_{\pi}\mathbf{A}_*^{\pi}) * \mathbf{A}_*^{\gamma} + \sigma_t(\sum_{\pi}\mathbf{A}_*^{\pi}) * \mathbf{A}_*^{\gamma}\right], \tag{9}$$

then, we calculate the agent key attention as shown in Eq. 10,

$$\hat{\mathbf{M}}_{KA}^{\gamma} = \sigma_c(MatMul\left[\alpha * \hat{\mathbf{A}}_*^{\gamma}, \mathbf{K}^{\gamma}\right]). \tag{10}$$

The final aggregated visual tokens can be obtained through the following equation as demonstrated in Eq. 11,

$$\mathbf{t}_*^{\gamma} = FFN(MatMul\left[\hat{\mathbf{M}}_{KA}^{\gamma}, MatMul\left[\hat{\mathbf{M}}_{QA}^{\gamma}, \mathbf{V}^{\gamma}\right]\right]). \tag{11}$$

Finally, we aggregate all the tokens from three branches by using the mean operation, where N_s indicates the number of the stream as shown in Eq. 12,

$$\mathbf{t}_{agg} = \sum_{\phi}\left[\mathbf{t}_*^{\phi}\right]/N_s. \tag{12}$$

We construct MLP-based classification and regression heads atop aggregated tokens for center frame bounding box prediction and atomic action recognition.

4.3 Loss Functions

We use Binary Cross Entropy (BCE) loss and Mean Squared Error (MSE) loss for the multi-label supervision and the bounding box regression supervision in

the same way as all the baselines. The equation of the BCE loss is as Eq. 13,

$$L_{BCE}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N_c}\sum_{i=1}^{N_c}[\mathbf{y}_i \log(\hat{\mathbf{y}}_i) + (1-\mathbf{y}_i)\log(1-\hat{\mathbf{y}}_i)], \tag{13}$$

where $\mathbf{y}$, $\hat{\mathbf{y}}$ and N_c indicate the one-hot ground truth, the prediction, and the category number. The bounding boxes regression loss is expressed as Eq. 14,

$$L_{MSE}(\mathbf{b}, \hat{\mathbf{b}}) = \sum_{j=1}^{4}(\mathbf{b}_j - \hat{\mathbf{b}}_j)^2, \tag{14}$$

where $\mathbf{b}$ indicates the coordinates of the left top and right bottom corners, and $\hat{\mathbf{b}}$ represents the predicted boxes. j indicates the coordinate index.

5 Experiments

5.1 Implementation Details

We conduct experiments on four NVIDIA A100 GPUs. We use BertAdam [32] optimizer with learning rate $lr = 1e^{-4}$, batch size 128, learning rate decay 0.9, and warmup ratio 0.1. The agent number and head number in our `RefAtomNet` are 4 and 1. Our model has $214M$ trainable parameters. The weight for L_{MSE} and α are 5 and 0.125. We trained our model for 40 epochs on our dataset. The text encoder is frozen during training. We use multi-label mean Average Precision (mAP), Area Under the Receiver Operating Characteristic curve (AUROC), and the mean Intersection over Union (mIOU) as metrics, of which mIOU has less priority since we prioritize the atomic video action recognition task.

5.2 Experimental Results

Main Results. Results of the main experiments on the proposed RefAVA benchmark are summarized in Tab. 2. We compare our RefAtomNet model with a multitude of published approaches adopted to suit our task (see Section 3.2) stemming from the fields of (1) Atomic Action Localization (AAL); (2) Video Question Answering (VQA); (3) Video-Text Retrieval (VTR); (4) Single Frame (SF) baselines; and (5) Video Object Segmentation (VOS). Baseline methods from the AAL group consistently underperform in spatially localizing the action according to the textual reference (measured as mIOU). Even though the ability of the human spatial localization is not the major expected output, it can still generally illustrate if the model can grasp the correct referring person or not. Among the AAL methods, X3D [16] achieves the highest AUROC score of 59.09% and 64.51% on the val and test sets, respectively. The VQA and VTR baselines work better than the AAL baselines since they can reasonably capture the referring person while delivering acceptable atomic human action recognition performances, which benefits from the text-aware pretraining of these two

Table 2: Experimental results on our RAVAR benchmark.

	Method	mIOU	mAP	AUROC	mIOU	mAP	AUROC
		Val			Test		
AAL	I3D [4]	0.00	44.04	57.77	0.00	44.64	62.71
	X3D [16]	0.26	44.45	59.09	0.27	46.34	64.51
	MViTv2-B [47]	0.76	42.32	56.43	0.66	42.60	59.30
	VideoMAE2-B [81]	0.16	42.02	55.12	0.29	41.87	59.10
	Hiera-B [69]	0.49	42.74	56.72	0.62	41.14	58.70
VQA	Singularity [38]	26.34	42.43	59.52	29.78	42.18	56.12
	AskAnything13B [44]	20.09	51.42	66.12	22.35	52.25	69.35
VTR	MeVTR [95]	30.78	38.42	51.01	29.79	36.27	52.45
	CLIP4CLIP [55]	34.75	39.48	52.57	32.33	37.17	55.05
	XClIP [56]	35.54	42.46	56.30	31.79	40.82	58.71
	BLIPv2 [42]	32.99	52.13	66.56	32.75	53.19	69.92
SF	CLIP [67]+SAM [33]	32.95	49.74	64.55	29.80	51.34	69.11
	CLIP [67]+DETR [3]	33.96	47.57	61.83	33.84	50.92	67.29
	CLIP [67]+REFCLIP [29]	34.43	47.79	62.50	32.28	49.90	67.44
VOS	Su *et al.* [79]	23.71	52.17	66.67	26.02	53.20	70.19
	RefAtomNet (Ours)	**38.22**	**55.98**	**69.73**	**36.42**	**57.52**	**73.95**

Table 3: Experiments for module ablation and comparison with multi-modal fusion.

(a) Module ablation of the RefAtomNet.

Method	mIOU	mAP	AUROC	mIOU	mAP	AUROC
	Val			Test		
w/o ALSAF	27.30	50.70	65.31	29.09	51.26	68.20
w/o CAAF	36.21	55.43	69.66	35.38	55.90	72.60
w/o CATF	35.01	53.83	67.71	34.55	56.96	73.40
w/o LSAS	31.90	55.21	69.47	31.25	55.73	72.38
Ours	**38.22**	**55.98**	**69.73**	**36.42**	**57.52**	**73.95**

(b) Comparison with other fusion approaches.

Fusion	mIOU	mAP	AUROC	mIOU	mAP	AUROC
	Val			Test		
Addition	27.30	50.70	65.31	29.09	51.26	68.20
Concatenation	18.64	52.23	66.45	20.65	53.44	70.70
Multiplication	23.90	51.55	65.66	25.05	53.33	70.48
AttentionBottleneck [59]	33.47	50.97	65.07	33.02	54.02	71.08
McOmet [98]	23.88	51.58	65.65	25.02	53.21	70.42
Ours	**38.22**	**55.98**	**69.73**	**36.42**	**57.52**	**73.95**

tasks. AskAnything [44] from the VQA group achieves 20.09% and 22.35% of mIOU, 51.42% and 52.25% of mAP, and 66.12% and 69.35% of AUROC on the val and test sets. BLIPv2 [43] from the VTR group delivers 32.99% and 32.75% of mIOU, 52.13% and 53.19% of mAP, and 66.56% and 69.92% of AUROC, on the val and test sets, respectively. Our `RefAtomNet` achieves state-of-the-art performances, outperforming the best baseline BLIPv2 [42] by 5.23%, 3.85%, 3.17% and 3.67%, 4.33%, 4.03% of mIOU, mAP, and AUROC, on val and test sets.

Ablations of the Individual Modules. In Tab. 3a, we show the ablation for the components of `RefAtomNet` by removing each of the proposed designs, where LSAS indicates the location-semantic stream, CAAF indicates the cross-stream agent attention fusion, CATF indicates the cross-stream agent token fusion, and *w/o* ALSAF indicates by simply using addition to fuse these three streams and without the agent-based location-semantic aware attentional fusion mechanism. Compared with the ablation *w/o* ALSAF, `RefAtomNet` achieves promising improvements of 10.92%, 5.28%, 4.42% and 7.33%, 6.26%, 5.75% in terms of mIOU, mAP, and AUROC for the val and test sets, respectively. It indicates that using simple aggregation of the three streams makes the model distract from the important cues for the referring person, illustrated by the large decay of mIOU metric.

Table 4: Generalizability to different referring styles and encoder architectures.

(a) Generalizability to test time rephrasing

Fusion	mIOU	mAP	AUROC	mIOU	mAP	AUROC
	Val			Test		
Singularity [38]	18.45	41.27	58.47	20.54	41.39	55.32
XCLIP [56]	31.95	41.35	54.35	29.84	40.74	58.45
AskAnything [44]	19.74	51.11	65.83	21.60	51.96	69.04
BLIPv2 [42]	31.00	51.45	65.88	31.34	52.35	68.87
Ours	**34.65**	**55.75**	**69.52**	**33.14**	**57.23**	**73.76**

(b) Generalizability to different visual-textual encoder architectures.

Fusion	mIOU	mAP	AUROC	mIOU	mAP	AUROC
	Val			Test		
XCLIP [56]	35.54	42.46	56.30	31.79	40.82	58.71
RefAtomNet (XCLIP)	**38.59**	**47.40**	**61.20**	**36.61**	**48.59**	**66.47**
BLIPv2 [42]	32.99	52.13	66.56	32.75	53.19	69.92
RefAtomNet (BLIPv2)	**38.22**	**55.98**	**69.73**	**36.42**	**57.52**	**73.95**

Compared with the ablation *w/o* LSAS, we find that the location-semantic tokens benefit more for the localization of the correct person in the center frame. The advantages of the LSAS on atomic video action recognition metrics are highlighted in the test set which has higher scenario diversity compared with the val set. Using CAAF and CATF, each of them brings promising benefits by suppressing the irrelevant information in the visual tokens.

Comparison with Other Fusion Mechanisms. We compare our proposed agent-based location-semantic aware attentional fusion mechanism with late fusion addition, multiplication, concatenation, AttentionBottleNeck [59], and McOmet [98] in Tab. 3b. Our approach outperforms all by suppressing those irrelevant visual tokens on the agent tokens and attention masks.

Generalizability to Test-time Reference Rephrasing. The reference sentences given by different users may differ in daily life scenarios. To test the generalizability of the model towards different referring styles, we invoke API (gpt-3.5-turbo) of ChatGPT [60] to rephrase the test set description two times and then deliver the averaged performance for RAVAR on the original val and test sets, and the two rephrased val and test sets, where we select 4 most outperforming baselines from our benchmark to construct this ablation study, *i.e.*, Singularity [38], AskAnything [44], XCLIP [56], and BLIPv2 [42], as shown in Tab. 4a. `RefAtomNet` delivers the best performances. We find out that the test time reference rephrasing will cause performance decay for localization and less decay for the atomic video-based action recognition delivered by our model.

Generalizability to Different Visual Textual Backbones. Since we use the best-performing baseline BLIPv2 [42] as the visual and textual encoder in our model, it would be interesting to see if the proposed architecture can generalize to different encoder backbones. We thereby demonstrate another ablation study in Tab. 4b, where we equip our `RefAtomNet` with the XCLIP [56] backbone. Compared with the XCLIP baseline, the `RefAtomNet` (XCLIP) achieves 3.05%, 4.94%, 4.90% and 4.82%, 7.77%, 7.76% performance improvements in terms of mIOU, mAP, and AUROC on the val and test sets, respectively, showing the great generalizability of our `RefAtomNet` of the visual and textual encoder.

Analysis of the Qualitative Results. The qualitative results are delivered in Fig. 4. These examples demonstrate the effectiveness of `RefAtomNet` compared with the best-performing baseline BLIPv2 [42]. Samples ❶ and ❽ validate the recognition quality for references without any location indications.

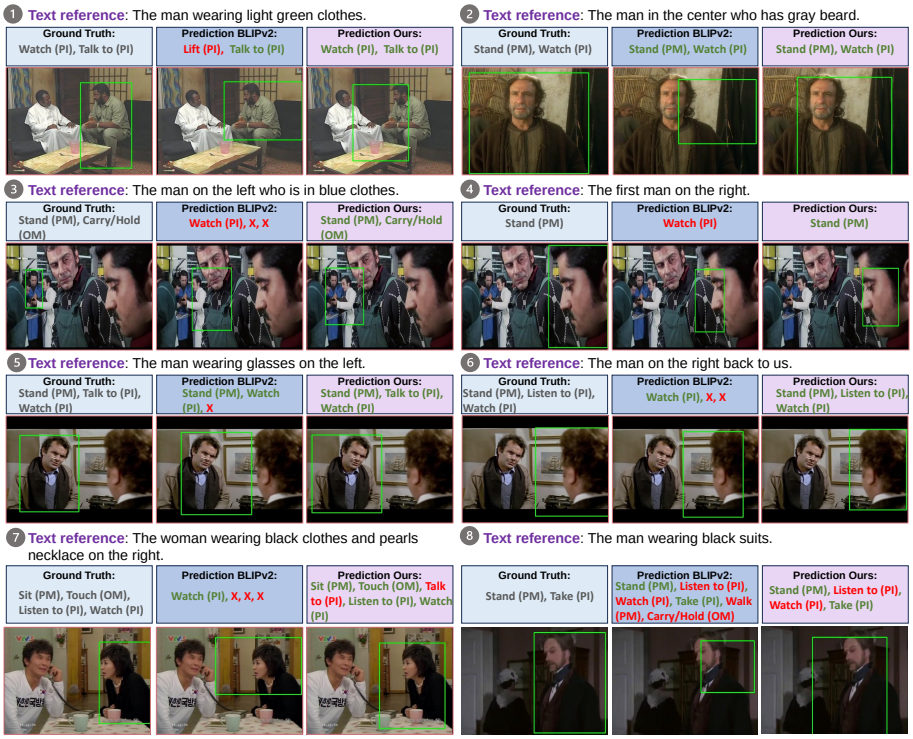


Fig. 4: An overview of qualitative results. Missed predictions are marked with a red cross, while true positive and false positive predictions are shown in green and red.

Samples ❸, ❹, ❺, and ❻ showcase the results for different humans in a shared scene. Samples ❷ and ❼ show the RAVAR performance for references containing finer details, *e.g.*, the necklace type and the beard color. Our `RefAtomNet` outperforms the BLIPv2 baseline on selected samples obviously, attributable to its superior capability in eliminating extraneous visual cues and its advanced location-semantic reasoning prowess.

6 Conclusions

In this work, we introduce a novel task called Referring Atomic Video Action Recognition (RAVAR). We establish the RefAVA dataset to address the challenge of identifying atomic actions for individuals of interest in videos based on textual descriptions. Existing methods exhibit poor performance on this new task, prompting the development of `RefAtomNet`, a vision-language architecture that effectively integrates cross-stream tokens for precise referring atomic video action recognition. Through reference-relevant token enhancement and comprehensive token integration, `RefAtomNet` achieves impressive results, highlighting its effectiveness in tackling the complexities of the RAVAR task.

Acknowledgements

The project served to prepare the SFB 1574 Circular Factory for the Perpetual Product (project ID: 471687386), approved by the German Research Foundation (DFG, German Research Foundation) with a start date of April 1, 2024. This work was also partially supported in part by the SmartAge project sponsored by the Carl Zeiss Stiftung (P2019-01-003; 2021-2026). This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. The authors also acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. This project is also supported by the National Key RD Program under Grant 2022YFB4701400. Lastly, the authors thank for the support of Dr. Sepideh Pashami, the Swedish Innovation Agency VINNOVA, the Digital Futures.

References

1. Bagad, P., Tapaswi, M., Snoek, C.G.M.: Test of time: Instilling video-language models with a sense of time. In: CVPR (2023)
2. Bu, Y., Li, L., Xie, J., Liu, Q., Cai, Y., Huang, Q., Li, Q.: Scene-text oriented referring expression comprehension. TMM (2022)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: CVPR (2017)
5. Castro, S., Deng, N., Huang, P., Burzo, M., Mihalcea, R.: In-the-wild video question answering. In: COLING (2022)
6. Chai, W., Guo, X., Wang, G., Lu, Y.: StableVideo: Text-driven consistency-aware diffusion video editing. In: ICCV (2023)
7. Chen, J., Zhu, D., Haydarov, K., Li, X., Elhoseiny, M.: Video ChatCaptioner: Towards enriched spatiotemporal descriptions. arXiv preprint arXiv:2304.04227 (2023)
8. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
9. Chen, Y., Wang, J., Lin, L., Qi, Z., Ma, J., Shan, Y.: Tagging before alignment: Integrating multi-modal tags for video-text retrieval. arXiv preprint arXiv:2301.12644 (2023)
10. Chen, Z., Ma, L., Luo, W., Wong, K.Y.K.: Weakly-supervised spatio-temporally grounding natural sentence in video. arXiv preprint arXiv:1906.02549 (2019)
11. Chung, J., Wuu, C.h., Yang, H.r., Tai, Y.W., Tang, C.K.: HAA500: Human-centric atomic action dataset with curated videos. In: ICCV (2021)
12. Dang, R., Feng, J., Zhang, H., Ge, C., Song, L., Gong, L., Liu, C., Chen, Q., Zhu, F., Zhao, R., Song, Y.: InstructDET: Diversifying referring object detection with generalized instructions. arXiv preprint arXiv:2310.05136 (2023)
13. Deruyttere, T., Vandenhende, S., Grujicic, D., Van Gool, L., Moens, M.F.: Talk2Car: Taking control of your self-driving car. In: EMNLP (2019)

14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: ACL (2019)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
16. Feichtenhofer, C.: X3D: Expanding architectures for efficient video recognition. In: CVPR (2020)
17. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: ICCV (2019)
18. Gandhi, M., Gul, M.O., Prakash, E., Grunde-McLaughlin, M., Krishna, R., Agrawala, M.: Measuring compositional consistency for video question answering. In: CVPR (2022)
19. Gao, D., Zhou, L., Ji, L., Zhu, L., Yang, Y., Shou, M.Z.: MIST: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In: CVPR (2023)
20. Garcia, N., Otani, M., Chu, C., Nakashima, Y.: KnowIT VQA: Answering knowledge-based questions about videos. In: AAAI (2020)
21. Gavrilyuk, K., Ghodrati, A., Li, Z., Snoek, C.G.: Actor and action video segmentation from a sentence. In: CVPR (2018)
22. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fründ, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense. In: ICCV (2017)
23. Gritsenko, A., Xiong, X., Djolonga, J., Dehghani, M., Sun, C., Lučić, M., Schmid, C., Arnab, A.: End-to-end spatio-temporal action localisation with video transformers. arXiv preprint arXiv:2304.12160 (2023)
24. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: AVA: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR (2018)
25. Guo, W., Zhang, Y., Yang, J., Yuan, X.: Re-attention for visual question answering. TIP (2021)
26. Han, D., Ye, T., Han, Y., Xia, Z., Song, S., Huang, G.: Agent attention: On the integration of softmax and linear attention. arXiv preprint arXiv:2312.08874 (2023)
27. Ji, Y., Zhan, Y., Yang, Y., Xu, X., Shen, F., Shen, H.T.: A context knowledge map guided coarse-to-fine action recognition. TIP (2020)
28. Jiang, J., Chen, Z., Lin, H., Zhao, X., Gao, Y.: Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In: AAAI (2020)
29. Jin, L., Luo, G., Zhou, Y., Sun, X., Jiang, G., Shu, A., Ji, R.: Refclip: A universal teacher for weakly supervised referring expression comprehension. In: CVPR (2023)
30. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: ACCV (2019)
31. Kim, M., Spinola, F., Benz, P., Kim, T.h.: A*: Atrous spatial temporal action recognition for real time applications. In: WACV (2024)
32. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
33. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: CVPR (2023)
34. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A large video database for human motion recognition. In: ICCV (2011)

35. Laput, G., Harrison, C.: Sensing fine-grained hand activity with smartwatches. In: CHI (2019)
36. Le, T.M., Le, V., Venkatesh, S., Tran, T.: Hierarchical conditional relation networks for video question answering. In: CVPR (2020)
37. Lea, C., Vidal, R., Hager, G.D.: Learning convolutional action primitives for fine-grained action recognition. In: ICRA (2016)
38. Lei, J., Berg, T.L., Bansal, M.: Revealing single frame bias for video-and-language learning. arXiv preprint arXiv:2206.03428 (2022)
39. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: ClipBERT for video-and-language learning via sparse sampling. In: CVPR (2021)
40. Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.R., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios. In: CVPR (2022)
41. Li, J., Niu, L., Zhang, L.: From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In: CVPR (2022)
42. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
43. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
44. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: VideoChat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
45. Li, L., Gan, Z., Lin, K., Lin, C.C., Liu, Z., Liu, C., Wang, L.: LAVENDER: Unifying video-language understanding as masked language modeling. In: CVPR (2023)
46. Li, R., Li, K., Kuo, Y.C., Shu, M., Qi, X., Shen, X., Jia, J.: Referring image segmentation via recurrent refinement networks. In: CVPR (2018)
47. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: MViTv2: Improved multiscale vision transformers for classification and detection. In: CVPR (2022)
48. Lin, J., Chen, J., Peng, K., He, X., Li, Z., Stiefelhagen, R., Yang, K.: EchoTrack: Auditory referring multi-object tracking for autonomous driving. arXiv preprint arXiv:2402.18302 (2024)
49. Lin, X., Tiwari, S., Huang, S., Li, M., Shou, M.Z., Ji, H., Chang, S.F.: Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval. In: CVPR (2023)
50. Liu, J., Wang, L., Yang, M.H.: Referring expression generation and comprehension via attributes. In: ICCV (2017)
51. Liu, R., Zhang, J., Peng, K., Zheng, J., Cao, K., Chen, Y., Yang, K., Stiefelhagen, R.: Open scene understanding: Grounded situation recognition meets segment anything for helping people with visual impairments. In: ICCVW (2023)
52. Liu, R., Liu, C., Bai, Y., Yuille, A.L.: CLEVR-Ref+: Diagnosing visual reasoning with referring expressions. In: CVPR (2019)
53. Liu, S., Hui, T., Huang, S., Wei, Y., Li, B., Li, G.: Cross-modal progressive comprehension for referring segmentation. TPAMI (2021)
54. Liu, Y., Li, G., Lin, L.: Cross-modal causal relational reasoning for event-level visual question answering. TPAMI (2023)
55. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. Neurocomputing (2022)
56. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In: MM (2022)

57. Madasu, A., Aflalo, E., Ben Melech Stan, G., Tseng, S.Y., Bertasius, G., Lal, V.: Improving video retrieval using multilingual knowledge transfer. In: ECIR (2023)
58. McIntosh, B., Duarte, K., Rawat, Y.S., Shah, M.: Visual-textual capsule routing for text-based video segmentation. In: CVPR (2020)
59. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. In: NeuIPS (2021)
60. OpenAI: ChatGPT: Optimizing language models for dialogue. `https://openai.com/` (2022)
61. Ordonez, V., Kulkarni, G., Berg, T.: Im2Text: Describing images using 1 million captioned photographs. In: NeurIPS (2011)
62. Ou, W., Zhang, J., Peng, K., Yang, K., Jaworek, G., Müller, K., Stiefelhagen, R.: Indoor navigation assistance for visually impaired people via dynamic SLAM and panoptic segmentation with an RGB-D sensor. In: ICCHP (2022)
63. Peng, K., Roitberg, A., Yang, K., Zhang, J., Stiefelhagen, R.: TransDARC: Transformer-based driver activity recognition with latent space feature calibration. In: IROS (2022)
64. Pramanick, P., Sarkar, C., Paul, S., dev Roychoudhury, R., Bhowmick, B.: DoRO: Disambiguation of referred object for embodied agents. RA-L (2022)
65. Pramono, R.R.A., Chen, Y.T., Fang, W.H.: Spatial-temporal action localization with hierarchical self-attention. TMM (2021)
66. Qiu, H., Li, H., Wu, Q., Meng, F., Shi, H., Zhao, T., Ngan, K.N.: Language-aware fine-grained object representation for referring expression comprehension. In: MM (2020)
67. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
68. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Feichtenhofer, C., Malik, J.: On the benefits of 3D pose and tracking for human action recognition. In: CVPR (2023)
69. Ryali, C., Hu, Y., Bolya, D., Wei, C., Fan, H., Huang, P., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., Malik, J., Li, Y., Feichtenhofer, C.: Hiera: A hierarchical vision transformer without the bells-and-whistles. In: ICML (2023)
70. Saha, J., Chowdhury, C., Chowdury, I.R., Roy, P.: Fine grained activity recognition using smart handheld. In: ICDCN (2018)
71. Seibold, C.M., Reiß, S., Kleesiek, J., Stiefelhagen, R.: Reference-guided pseudo-label generation for medical semantic segmentation. In: AAAI (2022)
72. Seo, S., Lee, J.Y., Han, B.: URVOS: Unified referring video object segmentation network with a large-scale benchmark. In: ECCV (2020)
73. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: CVPR (2020)
74. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
75. Shi, H., Pan, W., Zhao, Z., Zhang, M., Wu, F.: Unsupervised domain adaptation for referring semantic segmentation. In: MM (2023)
76. Shi, H., Li, H., Meng, F., Wu, Q.: Key-word-aware network for referring expression image segmentation. In: ECCV (2018)
77. Shi, Y., Xu, H., Yuan, C., Li, B., Hu, W., Zha, Z.J.: Learning video-text aligned representations for video captioning. TOMM (2023)
78. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)

79. Su, Y., Wang, W., Liu, J., Ma, S., Yang, X.: Sequence as a whole: A unified framework for video action localization with long-range text query. TIP (2023)
80. Vasudevan, A.B., Dai, D., Van Gool, L.: Object referring in videos with language and human gaze. In: CVPR (2018)
81. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: VideoMAE V2: Scaling video masked autoencoders with dual masking. In: CVPR (2023)
82. Wang, M., Xing, J., Mei, J., Liu, Y., Jiang, Y.: ActionCLIP: Adapting language-image pretrained models for video action recognition. TNNLS (2023)
83. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Yuan, L., Jiang, Y.G.: Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In: CVPR (2023)
84. Wang, S., Yan, R., Huang, P., Dai, G., Song, Y., Shu, X.: Com-STAL: Compositional spatio-temporal action localization. TCSVT (2023)
85. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., Qiao, Y.: InternVideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
86. Wu, D., Han, W., Wang, T., Dong, X., Zhang, X., Shen, J.: Referring multi-object tracking. In: CVPR (2023)
87. Wu, W., Luo, H., Fang, B., Wang, J., Ouyang, W.: Cap4Video: What can auxiliary captions do for text-video retrieval? In: CVPR (2023)
88. Xiao, J., Shang, X., Yao, A., Chua, T.S.: NExT-QA: Next phase of question-answering to explaining temporal actions. In: CVPR (2021)
89. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Just ask: Learning to answer questions from millions of narrated videos. In: ICCV (2021)
90. Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., Zhu, W.: AVQA: A dataset for audio-visual question answering on videos. In: MM (2022)
91. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV (2016)
92. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: ActivityNet-QA: A dataset for understanding complex web videos via question answering. In: AAAI (2019)
93. Yuan, Z., Yan, X., Liao, Y., Zhang, R., Wang, S., Li, Z., Cui, S.: InstanceRefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In: ICCV (2021)
94. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: MOTR: End-to-end multiple-object tracking with transformer. In: ECCV (2022)
95. Zhang, G., Ren, J., Gu, J., Tresp, V.: Multi-event video-text retrieval. In: CVPR (2023)
96. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In: EMNLP (2023)
97. Zheng, J., Zhang, J., Yang, K., Peng, K., Stiefelhagen, R.: MateRobot: Material recognition in wearable robotics for people with visual impairments. In: ICRA (2024)
98. Zong, D., Sun, S.: McOmet: Multimodal fusion transformer for physical audiovisual commonsense reasoning. In: AAAI (2023)