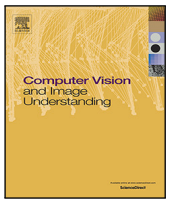
Computer Vision and Image Understanding 264 (2026) 104622

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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



Temporal video segmentation with natural language using text–video cross attention and Bayesian order-priors☆

Carlos Plou *,[1], Lorenzo Mur-Labadia[1], Jose J. Guerrero, Ruben Martinez-Cantin, Ana C. Murillo

DIIS - Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Spain

ARTICLE INFO

Keywords:
Video understanding
Egocentric vision
Temporal action segmentation

ABSTRACT

Video is a crucial perception component in both robotics and wearable devices, two key technologies to enable innovative assistive applications, such as navigation and procedure execution assistance tools. Video understanding tasks are essential to enable these systems to interpret and execute complex instructions in real-world environments. One such task is step grounding, which involves identifying the temporal boundaries of activities based on natural language descriptions in long, untrimmed videos. This paper introduces Bayesian-VSLNet, a probabilistic formulation of step grounding that predicts a likelihood distribution over segments and refines it through Bayesian inference with temporal-order priors. These priors disambiguate cyclic and repeated actions that frequently appear in procedural tasks, enabling precise step localization in long videos. Our evaluations demonstrate superior performance over existing methods, achieving state-of-the-art results in the Ego4D Goal-Step dataset, winning the *Goal Step* challenge at the EgoVis 2024 CVPR. Furthermore, experiments on additional benchmarks confirm the generality of our approach beyond Ego4D. In addition, we present qualitative results in a real-world robotics scenario, illustrating the potential of this task to improve human–robot interaction in practical applications. Code is released at https://github.com/cplou99/BayesianVSLNet.

1. Introduction

The proliferation of wearable and mobile devices and the growing availability of assistive robotic platforms presents plenty of opportunities to develop and integrate assistive technologies into users' daily lives. Many of these innovative applications are enabled by video perception systems, making understanding video content a crucial task for these domains. For example, wearable video devices are opening up new possibilities for health and safety among other fields, including applications such as assistance to impaired people with supermarket shopping (Mazzamuto et al., 2024), or detection of mistakes in procedural egocentric videos (Flaborea et al., 2024). These kinds of application require detailed video understanding. The temporal grounding of events of interest, often described by open language descriptions, is a challenging task toward this goal.

Understanding video demonstrations and aligning video content with natural language descriptions is also an essential perception task for robot learning (Mees et al., 2022). In particular, imitation learning and reinforcement learning using video demonstrations (Aytar et al., 2018; Baker et al., 2022; Du et al., 2024; Escontrela et al., 2024) have great potential in robotic applications, as videos can be easily reproduced multiple times with minimal cost and even transferred between robots. However, most related work is focused on learning using short videos and simple tasks, splitting complex behaviors into simple steps or single actions that can be trained individually (Schmeckpeper et al., 2021). This prior splitting limits the rich information that can be extracted from videos, and hinders the performance for complex plans. Autonomous robots, operating in real-world environments, gather vast amounts of visual data throughout the day (Liao et al., 2022; Fu et al., 2024) and numerous video demonstrations can be easily found on Internet (Aytar et al., 2018).

In this context of vast amounts of unlabeled video data, the task of *step grounding* (Song et al., 2024) is crucial. The objective of this task is to localize the temporal boundaries of activities, described in free-form natural language, within long and untrimmed videos. Assistive devices require strong episodic memory capabilities (Grauman et al., 2022), in order to identify the location of certain objects in a full video, discarding multiple irrelevant frames and focusing on certain actions.

☆ This article is part of a Special issue entitled: 'ACVR 2024' published in Computer Vision and Image Understanding.

* Corresponding author.
E-mail address: c.plou@unizar.es (C. Plou).
URL: https://cplou99.github.io/BayesianVSLNet/ (C. Plou).

[1] Carlos Plou and Lorenzo Mur-Labadia contributed equally in this work.

https://doi.org/10.1016/j.cviu.2025.104622
Received 15 April 2025; Received in revised form 2 September 2025; Accepted 17 December 2025
Available online 31 December 2025

Fig. 1. Step grounding task: localize the segment in a long untrimmed video that represents the free-form natural language description of the step. The example represents the step grounding task (step 7 and step 12) along a video captured by an autonomous robot performing household chores (Fu et al., 2024).

In robotics, it is usually needed to decompose complex tasks -such as object manipulation (Gao et al., 2024) or household chores (Cao et al., 2024)- into manageable steps, facilitating decision-making (Wang et al., 2022) and execution, or easing a posterior imitation learning (Karnan et al., 2022). Fig. 1 shows an example of our approach to analyze the video captured by a robot. To identify the particular video clips corresponding to the steps of *washing clothes*, we can just describe the task with simple steps in natural language: *"Prepare the soap"* or *"Put the washed clothes in the dryer"*.

Step grounding addresses two key challenges that are critical for real-world applications (Fig. 1). First, unlike traditional temporal action segmentation methods that rely on a fixed set of action labels (Zhang et al., 2022; Shi et al., 2023; Yang et al., 2023; Vahdani and Tian, 2022; Caba Heilbron et al., 2015; Kuehne et al., 2014), *step grounding* introduces flexibility by using natural language descriptions to identify actions. Second, the *step grounding* handles long, untrimmed videos, enabling assistive vision devices and autonomous robots to manage large-scale visual data and prolonged sequences. In contrast, previous works typically focus on short clips of only a few minutes (Kuehne et al., 2014; Bansal et al., 2022; Yi et al., 2022; Lu and Elhamifar, 2024). The extended duration of videos intensifies the "needle in a haystack" problem, where irrelevant frames interfere with the precise alignment between the video content and the textual query, leading to a loss of contextual detail.

In this paper, we propose Bayesian-VSLNet, a method that addresses the challenges presented in *step grounding*. Our approach first extracts a video–text feature representation for each processed video and the text query using Video Language Pre-trained (VLP) models, which are specially trained via contrastive learning to align both modalities in a common feature space. Then, we extend VSLNet (Zhang et al., 2020), a video span localizing network, with a novel training strategy that groups all the identical text queries of a video and a head that predicts the binary probability of each video segment representing the text query. This vector represents the likelihood of each segment being aligned with the queried step, providing a fully probabilistic formulation that naturally handles multiple occurrences of the same action.

Additionally, we introduce a test-time refinement strategy that integrates temporal-order priors into the predictions. This Bayesian inference step combines the network likelihood with a temporal prior, yielding a posterior distribution that disambiguates which specific instance of a repeated step is being queried. Our enhanced formulation emphasizes robustness to cyclic actions and scalability to very long videos while reducing the computational cost by replacing VSLNet's dual-head architecture with a single probabilistic prediction head.

We evaluate our approach on the Ego4D Goal Step dataset (Song et al., 2024), which comprises videos of procedural activities. This dataset presents additional challenges, such as cyclic actions, a long-tail distribution of step durations, and very long videos lasting up to 5 h. While Ego4D is our main benchmark, we also report experiments on additional datasets, confirming that our contributions generalize beyond Ego4D.

In summary, our contributions are as follows:

- A fully probabilistic formulation of step grounding, where the model predicts a likelihood distribution over segments and combines it with temporal priors via Bayesian inference. This resolves the challenges posed by cyclic and repeated actions in long videos. Our approach achieves state-of-the-art results on the Ego4D 2024 challenge and strong performance on additional benchmarks.
- Two novel temporal grounding metrics that measure different aspects related to the step grounding task.
- Qualitative results in household assistive scenarios demonstrating its applicability to real-world robotic use cases.

2. Related works

2.1. Egocentric video understanding.

Egocentric (first-person) vision offers a unique perspective for understanding human behavior, as it closely captures fine-grained hand–object interaction details. The arrival of large-scale datasets such as Ego4D (Grauman et al., 2022) and Epic-Kitchens (Damen et al., 2018) has driven progress in action recognition (Bansal et al., 2022; Radevski et al., 2023), action anticipation (Mur-Labadia et al., 2024; Furnari and Farinella, 2020), affordance segmentation (Mur-Labadia et al., 2023; Nagarajan et al., 2020), and episodic memory (Mai et al., 2023; Bärmann and Waibel, 2022). These advancements in perception capabilities have led to various applications in assistive technologies. For example, (Mazzamuto et al., 2024) present an augmented reality system to guide individuals with impairments during supermarket shopping; (Bonanno et al., 2023) assist industrial operators in retrieving information on tools, equipment, and safety procedures; (Flaborea et al., 2024) detect mistakes in procedural egocentric videos; (Capi et al., 2014) combine GPS with visual information to guide users in urban environments; and Wong et al. (2022) provide step-by-step assistance in instructional videos through visual affordances.

The unique perspective of egocentric vision for capturing interactions has also converted it into promising way to scale up learning in robotics (Meltzoff, 1993; Bahl et al., 2022). For example, Goyal et al. (2022) learn interaction regions and afforded grasps from attending the hands movements, (Chang et al., 2020) leverage egocentric YouTube videos to learn navigation policies, while (Sermanet et al., 2024) fine-tune video captioning models on egocentric data to enable high-level reasoning over long-horizon tasks. However, the egocentric video modality introduces additional challenges, such as the sensor gap and the need for models to process extremely long recordings with a low density of informative frames. In this work, we introduce a test-time refinement strategy that incorporates the expected step order, improving the model's ability to handle long videos involving multiple action steps.

2.2. Video understanding for autonomous agents and robots

Automatically understanding video content is a fundamental perception task in robotics, enabling a wide range of applications such as aerial drone action recognition (Kothandaraman et al., 2023; Wang et al., 2023), forecasting in autonomous driving (Li et al., 2023; Kung et al., 2024), and real-world surveillance (Seo et al., 2021). By leveraging video data, robotic systems can track objects with higher spatial and temporal resolution, facilitating more accurate navigation in complex environments (Yang et al., 2022; Pieroni et al., 2024; Yao et al., 2023; Dionigi et al., 2022; Ballester et al., 2021). Moreover, video provides fine-grained features and temporally consistent representations, helping to compensate for the lack of visual detail in texture-less modalities such as point clouds (Deng et al., 2024; Liu et al., 2024). For instance, Deng et al. (2024) employed a video-language model to address the texture deficiency in 4D point clouds, effectively aligning both modalities to improve action recognition. Moreover, video is frequently employed in reinforcement learning for robotic manipulation (Aytar et al., 2018; Baker et al., 2022; Schmeckpeper et al., 2021), as it provides rich cues about interactions and the sequential structure of complex tasks. Schmeckpeper et al. (2021) combined data collected through robot interactions with observed videos of the same tasks to learn more effective control policies.

2.3. Bayesian statistics for video understanding

Bayesian statistics has also been applied to video understanding. Subedar et al. (2019) incorporated a Bayesian dropout-based variational layer into an audio–visual activity classifier to capture epistemic uncertainty in predictions and Guo et al. (2024) introduced a teacher–student Bayesian evidential deep learning model for uncertainty quantification in online action detection. Plou et al. (2024) incorporated Laplace ensembles into the final layer of a ViT to reduce overconfidence in action recognition. Rather than explicitly computing uncertainty, other approaches leverage Bayes' theorem to encode prior knowledge of step sequences, thereby enhancing robustness. Goel and Brunskill (2019) introduced a hierarchical Bayesian model that captures the realization of repeated procedural actions. Similarly, Fernando et al. (2015), Lee et al. (2017) exploited the temporal ordering of action steps to learn more robust video representations. In this work, we propose a Bayesian temporal order prior that adjusts for cyclic and repetitive actions, widely present in long, untrimmed videos.

2.4. Video temporal segmentation

From all the tasks related to video understanding, our work is focused on temporal segmentation, which aims to predict the start and end of actions. Initially, the *Temporal Action Localization* (TAL) task assumed a fixed and closed set of action labels (Zhang et al., 2022; Shi et al., 2023; Vahdani and Tian, 2022; Caba Heilbron et al., 2015; Yang et al., 2023; Kuehne et al., 2014). ActionFormer (Zhang et al., 2022) predicts actions boundaries using a multi-scale feature representation processed by a self-attention based transformer. Yang et al. (2023) introduced a weakly supervised technique that achieved comparable state of the art results using less than 1% of the fully supervised labels. The precise temporal segmentation of the actions has facilitated human–robot interaction in various scenarios, such as surgical cooperation with robots (Fard et al., 2016; De Rossi et al., 2021; Meli and Fiorini, 2021), assistance in daily living tasks (Zhu and Sheng, 2011) and cooking (Fukuda et al., 2005). Kukleva et al. (2019) leverage the sequential nature of activities to guide a clustering algorithm in an unsupervised approach. However, TAL approaches are constrained by the set of action labels used during training, which limits its development in real-world scenarios with a higher variety of actions and object semantics.

The arrival of Video-Language Pre-training (VLP) methods (Bain et al., 2021; Miech et al., 2019; Lin et al., 2022; Pramanick et al., 2023; Pei et al., 2024) allows the opportunity of more complicated challenges, like Natural Language Queries (NLQ) task (Grauman et al., 2022). It consists of identifying the moment in a video that answers a text query. VLP methods (Bain et al., 2021; Miech et al., 2019; Lin et al., 2022) learn transferable representations from a large-scale training on pairs of video and the respective text narration, using a contrastive learning objective that aligns both modalities. EgoVLPv2 (Pramanick et al., 2023) incorporates a cross-modal fusion mechanism inside the video and text backbones, learning stronger video–text representation. EgoVideo (Pei et al., 2024) collects a massive egocentric and exocentric video–text dataset, obtaining the SOTA in a wide variety of tasks. NLQ approaches (Hou et al., 2023, 2022; Lei et al., 2021; Nagarajan et al., 2024; Ramakrishnan et al., 2023; Wang et al., 2022) leverage VLP models to propose multiple solutions. GroundNLQ (Hou et al., 2023) incorporated a text-aware temporal pyramid to capture temporal intervals of varying lengths. Hou et al. (2022) proposed a sliding window technique to pre-filter candidate windows, preserving temporal resolution. EgoEnv (Nagarajan et al., 2024) contextualizes videos within their 3D physical environment, rather than relying on naive temporal feature aggregation. Ramakrishnan et al. (2023) transforms the common video–text narrations into training data for NLQ, substantially boosting the performance across top models. Lastly, Fang et al. (2024) introduces a novel clip selection approach to search the core clip iteratively. Specifically, we build upon VSLNet (Zhang et al., 2021), the gold-standard neural network for open-vocabulary video localization which proposes a video span localization network that employs a shared feature encoder followed by context-query attention to learn cross-modal features.

3. Background

In this work, we focus on the *step grounding* task (Song et al., 2024), which aims to identify the temporal clip *(start_time, end_time)* corresponding to a given free-form language description, representing a step within a procedural task. Our proposed method, Bayesian-VSLNet, extends the VSLNet architecture (Zhang et al., 2020; Zhang et al., 2021), the gold-standard neural network for natural language video localization, designed to localize a single query within a video. We adopt the VSLNet baseline due to its suitability for our target applications (robotics and assistive vision devices), offering a lightweight architecture compatible with embedded devices and enabling real-time inference with an average processing time of 125 ms per text query.

3.1. Problem definition

Given a set of long videos of procedural tasks, let $\mathcal{V}$ represent one of these videos. Assume that $\mathcal{V}$ has a duration of D seconds and contains a process of n steps. Each video $\mathcal{V}$ is linked with a text description of the process in the form of a set of natural language descriptions $\mathcal{T} = \{t_j\}_{j=1}^{n}$, where each t_j describes a specific step of the procedural task. Thus, our goal is, given a video $\mathcal{V}$ and a text query -i.e. natural language description of a step- t_j, to predict its starting and ending times (s_j, e_j) inside the video. We will refer to this time interval as a *step clip*.

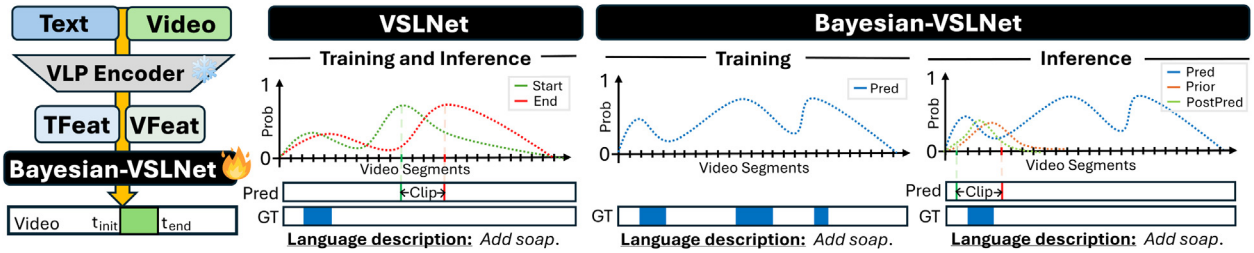


Fig. 2. Bayesian-VSLNet. (Left) Our architecture is an extension of VSLNet (Zhang et al., 2020) with two novel components: a novel head predicts the probability of the text query in each video segment and a Bayesian temporal-order prior refines the predictions during the inference stage. (Center) VSLNet predicts each step independently, producing a prediction probability for each segment of the video, resulting in inconsistent results for long videos or descriptions with multiple steps. (Right) Bayesian-VSL use a step prior based on the order of the sequence of steps which improves the accuracy for long videos and guarantees consistency in the process description. During training, a step description might be repeated multiple times (see training GT) which confuses VSLNet. However, at inference time we want to segment the video to the exact occurrence of that step where the ordering prior plays a fundamental role.

3.2. VSLNet

Given a video $\mathcal{V}$ and a text query t_j, VSLNet method starts pre-extracting the video and text representations through a video encoder (Omnivore-L (Girdhar et al., 2022)) and a text encoder (BERT (Devlin et al., 2018)), respectively.

Formally, let F_i denote the ith sampled frame from $\mathcal{V}$. The video encoder produces a d_v-dimensional feature vector for each frame,

$$\mathbf{f}_i = \text{VideoEncoder}(F_i) \in \mathbb{R}^{d_v}. \tag{1}$$

For long videos, it adopts a sparse sampling technique for compressing the video features. This technique involves dividing the video $\mathcal{V}$ into K uniform segments and averaging the video features within each segment.

$$\mathbf{v}_k = \frac{1}{|S_k|} \sum_{F_i \in S_k} \mathbf{f}_i, \quad k = 1, \dots, K, \tag{2}$$

where S_k denotes the set of frames in the kth segment. This transformation makes the problem more efficient by converting each step clips (s_j, e_j) into a discrete representation $(k_j^s, k_j^e) \in [0, K]$,

$$k_j^s = \left\lfloor \frac{s_j \cdot K}{D} \right\rfloor, \qquad k_j^e = \left\lfloor \frac{e_j \cdot K}{D} \right\rfloor. \tag{3}$$

Similarly, the text encoder processes the query t_j to obtain a d_t-dimensional feature vector:

$$\mathbf{q}_j = \text{TextEncoder}(t_j) \in \mathbb{R}^{d_t}. \tag{4}$$

After feature extraction, the K segment features $\mathbf{v}_{1:K}$ and the text feature $\mathbf{q}_j$ are processed jointly by VSLNet to produce two probability vectors, $\hat{\mathbf{p}}_j^s, \hat{\mathbf{p}}_j^e \in [0,1]^K$, estimating for each segment the likelihood of being the start or the end of the step described by t_j (Fig. 2). This can be expressed as:

$$\hat{\mathbf{p}}_j^s, \hat{\mathbf{p}}_j^e = \text{VSLNet}(\mathbf{v}_{1:K}, \mathbf{q}_j). \tag{5}$$

The model employs a dual-head prediction layer, one for start probabilities and another for end probabilities, applied over the fused video–text features obtained after context–query attention. To predict the final *step clip*, VSLNet computes an outer product of the start and end probability vectors:

$$\mathbf{M}_j = \hat{\mathbf{p}}_j^s \cdot (\hat{\mathbf{p}}_j^e)^\top \in \mathbb{R}^{K \times K}. \tag{6}$$

Only the upper-triangular part of $\mathbf{M}_j$ (where the end index is greater than or equal to the start index) is considered, as it corresponds to valid temporal intervals. The final predicted segment $(\hat{k}_j^s, \hat{k}_j^e)$ is obtained by selecting the (k_s, k_e) pair with the maximum value in this upper-triangular region. These indices are then mapped back to the continuous time domain to produce the final predicted *step clip* $(\hat{s}_j, \hat{e}_j)$.

In this way, VSLNet (Zhang et al., 2020) handles each step t_j of the video independently, making it challenging to model scenarios where different steps share the same text description, such as repeated actions within a procedural task. This is particularly problematic during training: when identical text queries appear, the model treats each instance separately, potentially leading to contradictions between iterations. Besides, during inference, all step clips within a video that share the same natural language description are assigned identical predictions, forgetting about the temporal order of the steps.

4. Method

Our method, Bayesian-VSLNet, reformulates the step grounding task by directly predicting, for a given video V and text query t_j, a probability vector $\hat{\mathbf{p}}_j = \{\hat{p}_j^k\}_{k=1}^K$, where each $\hat{p}_j^k$ indicates how likely it is that the kth segment contains the queried step. This single probability distribution over time replaces VSLNet's separate start/end predictions and their cross-product (Section 3.2), making the output more interpretable and enabling our approach to tackle two key aspects:

1. **Predicting the likelihood for grouped training.** Our network has a single head oriented to predict the *likelihood* $p_{\text{lik}}(\mathcal{V}, t_j \mid k)$, i.e., the probability of each segment k containing the queried step. This allows us to aggregate all occurrences of the same description into a single ground-truth vector marking every segment belonging to any instance of that step. For example, if "add salt" appears three times in a video, we merge these three occurrences into one binary target vector and train the model to assign high likelihood to all of them simultaneously. This grouping improves robustness in learning the visual–text pattern for the step and overcomes VSLNet's limitation of treating each occurrence independently.
2. **Bayesian inference to pinpoint the exact occurrence.** At inference, we may want to localize a specific occurrence—e.g., the second "add salt". The raw likelihood p_{lik} will typically exhibit peaks for all visually similar repetitions. To resolve this ambiguity, we apply Bayesian inference by combining the network's likelihood with a temporal prior $p_{\text{prior}}(k \mid j)$ encoding the jth order of the step t_j. The posterior distribution $p_{\text{post}} \propto p_{\text{lik}} \cdot p_{\text{prior}}$ from Eq. (12) becomes the final prediction of our Bayesian-VSLNet, reweighting the likelihood to suppress unrelated matches and concentrate probability mass around the intended occurrence, enabling precise localization even in cyclic or repetitive scenarios.

In summary, Bayesian-VSLNet reduces the computational cost of the baseline by removing one of VSLNet's two prediction heads—predicting a single probability vector instead of separate start/end distributions—and, at the same time, solves its two main limitations: ambiguity in repeated actions and loss of precision in long videos.

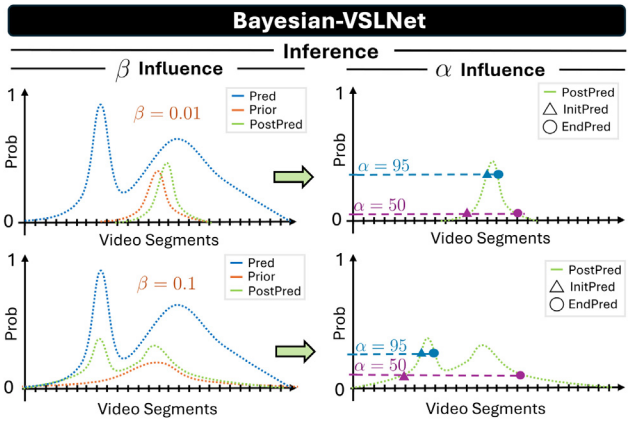


Fig. 3. Influence of the α and β hyper-parameters at the inference stage. β determines the variance of the prior that controls the smoothness of the posterior. It can be seen as the weight that we give to the step ordering. Once we have the posterior, α sets the threshold (α-percentile of the posterior probability value p_j^k) that controls the length of the predicted clip and can be used to control the ratio of true positives and false positives.

4.1. Text and video representations

Following previous work (Hou et al., 2022), we enhance our video representations by aggregating features from various video encoders (Girdhar et al., 2022; Pei et al., 2024; Pramanick et al., 2023). Omnivore-L (Girdhar et al., 2022) is a multi-modal vision model trained on images, videos, depth maps and 3D data using supervised learning, enabling it to generalize across different visual input modalities and producing modality-invariant embeddings. Additionally, we employ EgoVideo (Pei et al., 2024) and the dual-encoder version of EgoVLP-v2 (Pramanick et al., 2023), which are egocentric VLP models. Both models are trained via contrastive learning to align video and text embeddings, obtaining notable performance on episodic memory or action recognition tasks.

For each video encoder $m \in \{1, \dots, M\}$, we compute frame-level features $\mathbf{f}_i^{(m)}$ using Eq. (1), and then aggregate them into K uniform temporal segments $\mathbf{v}_k^{(m)}$ following Eq. (2).

Finally, the segment-level features from all encoders are concatenated to form a joint multi-encoder representation:

$$\mathbf{v}_k = \text{concat}\left(\mathbf{v}_k^{(1)}, \dots, \mathbf{v}_k^{(M)}\right) \in \mathbb{R}^{d_v}, \tag{7}$$

where $d_v = \sum_{m=1}^{M} d_m$. For the text modality, each query t_j is processed with the text encoder $\mathcal{E}_{\text{text}}$ using Eq. (4), and, when multiple VLP models are used, their text embeddings are concatenated analogously to the video features (Eq. (7)). Finally, in the same way as Eq. (3), we transform the ground truth step clips from time reference (s_j, e_j) into segments reference $(k_j^s, k_j^e) \in [0, K]$.

4.2. Bayesian-VSLNet

Architecture. Inspired by VSLNet (Zhang et al., 2020), our design reduces its two-head prediction scheme – where start and end probabilities are estimated independently – to a single prediction head (an LSTM layer followed by a sigmoid activation). This simplification lowers computational cost and produces a single probability score per segment. Beyond this architectural change, the output can now be interpreted in explicitly probabilistic terms: for a given video $\mathcal{V}$ and text query t_j, the network predicts a likelihood distribution over the K segments,

$$p_{\text{lik}}(\mathcal{V}, t_j \mid k) = \text{BayesianVSLNet}(\mathbf{v}_{1:K}, \mathbf{q}_j), \tag{8}$$

where $\mathbf{v}_{1:K}$ are the aggregated video features and $\mathbf{q}_j$ is the encoded query. This single probability distribution replaces VSLNet's cross-product of start and end vectors, yielding a more direct and interpretable formulation that is seamlessly extended into Bayesian inference (Fig. 2).

Training. During training, we aggregate all occurrences of the same step description into one ground-truth event vector $\mathbf{p}_j = \{p_j^k\}_{k=1}^K$. Formally:

$$p_j^k = \begin{cases} 1, \text{if } \exists t_q \in \mathcal{T} \text{ s.t. } \begin{array}{l} \textcircled{1}\ t_q = t_j \\ \textcircled{2}\ k_q^s \le k \le k_q^e \end{array} \\ 0, \text{otherwise} \end{cases} \tag{9}$$

This forces the network to learn the likelihood $p_{\text{lik}}(\mathcal{V}, t_j \mid k)$ by assigning high probability to every segment where the step occurs. For example, if "add salt" appears in steps 2, 5, and 9, all three are marked in the ground-truth vector, and the network is trained to produce peaks at each. The loss function is Binary Cross-Entropy:

$$\mathcal{L}_{\text{BCE}} = \text{BCE}(p_{\text{lik}}(\mathcal{V}, t_j \mid k), \mathbf{p}_j). \tag{10}$$

Inference. At test time, repeated or cyclic steps t_j within a video $\mathcal{V}$ may cause the likelihood $p_{\text{lik}}(\mathcal{V}, t_j \mid k)$ to peak at multiple positions. In our example, if we aim to predict the clip corresponding to step 5 "add salt", we must determine around which of the three distribution peaks the prediction should be centered. To disambiguate and select the intended occurrence, we apply Bayesian inference. A temporal prior $p_{\text{prior}}(k \mid j)$ encodes the jth order of the queried step t_j:

$$p_{\text{prior}}(k \mid j) = \mathcal{N}\left(k;\ \tfrac{j \cdot K}{n},\ K \cdot \beta\right), \tag{11}$$

where β controls the prior shape distribution and therefore its influence on the final posterior, which is then given by Bayes' rule:

$$p_{\text{post}}(k \mid \mathcal{V}, t_j) \propto p_{\text{lik}}(\mathcal{V}, t_j \mid k) \cdot p_{\text{prior}}(k \mid j). \tag{12}$$

In this way, we obtain a fully Bayesian posterior distribution over video segments, which provides a principled probabilistic view of the step t_j location inside a video $\mathcal{V}$. However, this is not yet our final goal, as the model must translate this distribution into a concrete predicted step clip $(\hat{k}_j^s, \hat{k}_j^e)$. To achieve this, the most likely segment is obtained with a process inspired in slice sampling algorithm in the MCMC literature. First, we find a point guaranteed to belong to the clip through MAP estimation:

$$k_j^* = \arg\max_k p_{\text{post}}(k \mid \mathcal{V}, t_j), \tag{13}$$

and, from it, we extend the clip forward $(k_j^* \rightarrow k)$ and backward $(k \leftarrow k_j^*)$ until $\hat{p}_j^k$ is under the α-percentile $\hat{p}_j^\alpha$, where α controls the segment amplitude. In other terms,

$$\begin{aligned} &\hat{k}_j^s : k \le k_j^* \text{ s.t.} \quad \hat{p}_j^{k-1} < \hat{p}_j^\alpha \le \hat{p}_j^k, \\ &\hat{k}_j^e : k \ge k_j^* \text{ s.t.} \quad \hat{p}_j^k \ge \hat{p}_j^\alpha > \hat{p}_j^{k+1}. \end{aligned} \tag{14}$$

Thus, the network produces a likelihood distribution (Eq. (8)), while our model combines it with a temporal prior (Eq. (11)) to get a posterior distribution (Eq. (12)) and, then, output the final *predicted step clip* $(\hat{k}_j^s, \hat{k}_j^e)$. The roles of both hyper-parameters α and β are visualized in Fig. 3 and their selection process is explained in Section 6.2.

5. Experimental setup

5.1. Dataset

We conduct our experiments on the Ego4D Goal-Step dataset (Song et al., 2024), which comprises egocentric videos of procedural activities, allowing us to evaluate our model in real-world scenarios that require processing long video sequences (Fu et al., 2024). The dataset consists of 368 h of egocentric video footage, spanning 851 videos

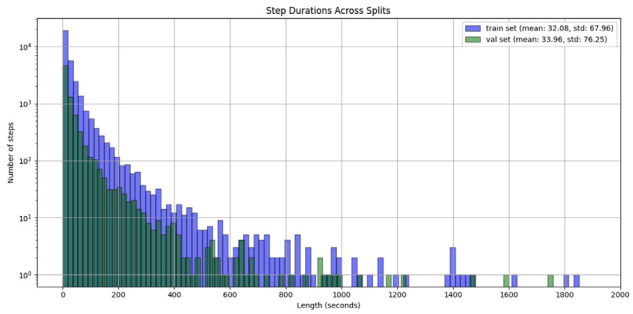


Fig. 4. Distribution of the step durations in the Ego4D Goal-Step dataset (Song et al., 2024).

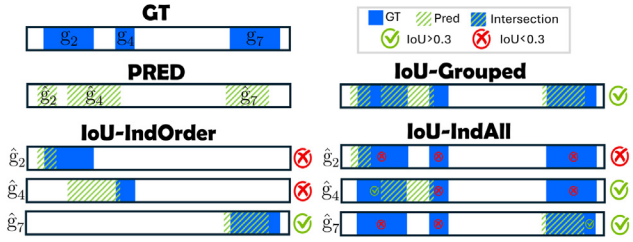


Fig. 5. Visualization of the metrics (`IoU-IndOrder`, `IoU-IndAll`, `IoU-Grouped`) for a video and three step clips g_2, g_4, g_7 that share the same natural language description. Here `R@1mIoU=0.3` values would be 33.3, 66.6, and 100, respectively.

with durations ranging from 15 s to 5 h, with an average length of 26 min. These videos exhibit a long-tail distribution of segment lengths, capturing procedural activities that vary from brief atomic actions lasting a few seconds to extended activities spanning several minutes. As shown in Fig. 4, this variability necessitates both fine-grained and global video understanding. In total, the dataset contains 48 K step annotations, densely labeled across the videos, averaging 56 annotations per video. Each annotation includes a time interval and a free-form natural language description of the ongoing action. Following previous work (Song et al., 2024), we extract video features with a stride of 16 frames – equivalent to 1.875 features per second – using a sliding window of 32 frames.

5.2. Metrics

Temporal grounding tasks commonly use the Intersection over Union (`IoU`) to measure the temporal similarity between the predicted step clip $\hat{g}_j = (\hat{s}_j, \hat{e}_j)$ and the ground truth step clip $g_j = (s_j, e_j)$ for each text query t_j. We will refer to this metric as the `IoU-IndOrder` metric.

However, `IoU-IndOrder` does not account for cases where some step clips from the procedural video $\mathcal{V}$ share identical natural language descriptions. As a result, the predicted step clip $\hat{g}_j$ for a text query t_j could perfectly overlap with another step clip g_k that shares the same description ($t_k = t_j$), yet the `IoU-IndOrder` value would be zero. To address this limitation and provide more flexibility, we propose two alternative temporal grounding metrics: `IoU-IndAll` and `IoU-Grouped`. Next, we will provide a detailed explanation of these metrics.

• `IoU-IndOrder`: it computes the `IoU` between a predicted segment $\hat{g}_j$ and the ground truth segment g_j located in the same order position j.

$$\texttt{IoU} - \texttt{IndOrder}\left(\hat{g}_j, g_j\right) = \texttt{IoU}\left(\hat{g}_j, g_j\right). \tag{15}$$

Table 1
Results in the Ego4D Goal-Step validation set (Song et al., 2024). We measure `R@1,mIoU=0.3` and `R@1,mIoU=0.5` for each temporal grounding metric.

	Validation-R@1 mIoU					
	Grouped		IndAll		IndOrder	
Model name	0.3	0.5	0.3	0.5	0.3	0.5
VSLNet (Baseline)	19.17	12.98	19.62	12.27	12.15	7.67
Bayesian-VSLNet-v0	24.28	12.70	24.83	12.01	18.15	8.97
Bayesian-VSLNet++	**30.43**	**18.71**	**32.48**	**19.62**	**23.75**	**14.41**

Table 2
Results for the different methods evaluated in the test set of the Ego4D Goal-Step dataset by the evaluator server as part of the Step Grounding challenge. It measures `R@1,mIoU=0.3` (primary metric) and `R@1,mIoU=0.5` for the `IoU-IndOrder` metric.

	Test-R@1mIoU-IndOrder	
Model Name	0.3	0.5
VSLNet (Baseline)	19.04	12.04
FlyFishing*	29.69	18.99
iLearn*	33.00	26.37
EgoVideo (Pei et al., 2024)	34.06	**26.97**
Bayesian-VSLNet-v0 (Ours)	**35.18**	20.48

• `IoU-IndAll`: For every predicted clip $\hat{g}_j$, this metric computes the `IoU` with each ground truth clip of the video $\mathcal{V}$ that has an identical text query, retaining the maximum `IoU` value obtained. In mathematical terms,

$$\texttt{IoU} - \texttt{IndAll}\left(\hat{g}_j, g_j\right) = \max_{q|t_q=t_j} \texttt{IoU}\left(\hat{g}_j, g_q\right). \tag{16}$$

• `IoU-Grouped`: This score calculates the `IoU` between the set of predicted step clips $\hat{G}_j = \{\hat{g}_q | t_q = t_j\}$ and the set of ground truth step clips $G_j = \{g_q | t_q = t_j\}$ whose text queries share the same natural language description.

$$\texttt{IoU} - \texttt{Grouped}\left(\hat{g}_j, g_j\right) = \texttt{IoU}\left(\hat{G}_j, G_j\right). \tag{17}$$

Specifically, Ego4D Goal-Step dataset leverages the `IoU-IndOrder` to report the recall-at-one (`R@1`) for `mIoU=0.3` and `mIoU=0.5`[2] (Grauman et al., 2022; Song et al., 2024), which measure the percentage of predicted clips that get an IoU value greater than 0.3 and 0.5, respectively. We show in Fig. 5 the main differences among the three temporal grounding metrics when leveraging them to report the `R@1,mIoU=0.3`.

6. Results

In this section, we analyze the impact of the improvements introduced by Bayesian-VSLNet for the *step grounding* task. Specifically, we evaluate two configurations of Bayesian-VSLNet, both of which share the same architecture, training, and inference procedures described in Section 4.2. The key difference between these configurations lies in their video/text feature representations and hyperparameter settings.

The first configuration of our model, Bayesian-VSLNet-v0, utilizes video features from Omnivore-L (Girdhar et al., 2022) and EgoVLPv2 (Pramanick et al., 2023), with test-time refinement hyperparameters set to $\beta = 0.1$ and $\alpha = 90$. As Table 2 shows, this model won the Ego4D Step Grounding challenge at the CVPR 2024 EgoVis workshop. Next, we developed an improved version, called Bayesian-VSLNet++, which employs a more robust video representation composed of features from Omnivore-L (Girdhar et al., 2022), EgoVLPv2 (Pramanick et al., 2023), and EgoVideo (Pei et al., 2024), along with refined test-time hyperparameters. However, since the evaluation server was promptly closed,

[2] `R@1,mIoU=0.3` was the metric used to rank the Ego4D 2024 challenge

Table 3
Ablation study of α and β hyper-parameters for the Bayesian-VSLNet model in the Ego4D Goal-Step validation set. We measure R@1,mIoU=0.3 and R@1,mIoU=0.5 for each of the temporal grounding metrics. Gray row shows the α and β values used in the Bayesian-VSLNet++ configuration.

Hyper.		Validation-R@1mIoU					
		Grouped		IndAll		IndOrd	
α	β	0.3	0.5	0.3	0.5	0.3	0.5
85	0.10	26.87	15.17	26.57	14.15	20.13	10.91
90	0.10	29.17	16.99	29.81	16.71	22.45	12.82
95	0.10	**30.43**	18.71	**32.48**	19.62	**23.75**	**14.41**
98	0.10	27.79	16.97	29.63	18.49	21.86	13.50
100	0.10	2.28	0.46	2.77	0.77	1.81	0.45
95	1e-4	0.05	0.01	0.74	0.49	0.05	0.01
95	0.01	17.11	8.15	17.83	7.33	13.84	5.99
95	0.05	28.39	16.51	29.89	16.93	22.82	13.03
95	0.10	**30.43**	18.71	**32.48**	19.62	**23.75**	**14.41**
95	0.15	29.89	**18.81**	32.41	**20.11**	23.12	14.37
95	0.20	29.17	18.75	31.55	19.97	22.27	14.10

we report results for the Bayesian-VSLNet++ only on the validation set, as Table 1 shows. All experiments were conducted using two NVIDIA GeForce RTX 4090 GPUs. We employed the AdamW optimizer with a linear learning rate scheduler for training.

6.1. Quantitative results

Table 1 shows that our proposed training and inference strategy (Section 4.2) enables Bayesian-VSLNet to significantly surpass the baseline VSLNet across all evaluated metrics. Notably, in the primary metric R@1mIoU-IndOrder=0.3, Bayesian-VSLNet-v0 achieves a 49.5% improvement over the baseline (18.15 vs. 12.15 R@1mIoU-IndOrder=0.3), while Bayesian-VSLNet++ exhibits an even more substantial gain, with a 95.5% increase (23.75 vs. 12.15 R@1mIoU-IndOrder=0.3). These results highlight the effectiveness of incorporating Bayesian priors in tasks where specifying the exact step instance is crucial, particularly when steps are duplicated in the description. Additionally, the improvements remain consistent across other metrics (mIoU-Grouped and mIoU-IndAll), indicating that our approach achieves more robust video understanding while effectively leveraging the temporal priors of Bayesian-VSLNet.

Next, we report the results of the Ego4D Step Grounding CVPR 2024 EgoVis workshop in Table 2. In order to ensure fair competition, it featured an evaluation server with a withheld ground truth test set. Our best configuration up to that point, Bayesian-VSLNet-v0, achieved the state-of-the-art by surpassing all other teams in the primary metric (35.18 R@1, mIoU-IndOrd=0.3) and winning the challenge. As the primary metric (R@1 mIoU-IndOrd=0.3) considers relevant the step order in repeated actions, the impact of our temporal priors results crucial, as it discards false positives outside the relevant video region.

6.2. Hyper-parameter selection and configuration studies

In this section, we describe the procedure used to select the key hyper-parameters of BayesianVSLNet that result in its top-performing configuration (BayesianVSLNet++) and several studies that compare its performance with respect to other configurations.

Hyper-parameter selection. For parameters affecting the *training* stage – the number of temporal segments each video is divided into (K) and the text and video encoders – we performed independent training runs, fully training and evaluating each configuration on the validation set. The best-performing configuration for each model is reported in Table 4, ensuring fair comparisons between VSLNet and Bayesian-VSLNet at their optimal temporal resolutions.

For parameters affecting only the *inference* stage – specifically α and β – we conducted a systematic grid search over the values reported in Table 3. We used the validation set to identify the combination that maximized performance in the primary metric (R@1, mIoU-IndOrder=0.3) while maintaining stability across the other metrics.

We note that the optimal values of these hyper-parameters may vary depending on the dataset characteristics (e.g., average video duration, number of text queries per video) and on the available hardware resources, since larger K values generally improve temporal precision at the cost of increased computational and memory demands. Therefore, to achieve optimal performance in a new setting, this selection process should be repeated.

Configuration studies. First, we evaluated the impact of encoding video and the step description across different models in Table 4. The combination of Omnivore-L with BERT features yields the lowest performance (15.20 R@1mIoU-IndOrder=0.3, and 7.32 mIoU=0.5) since both encoders are trained with data of their respective modality and consequently, they lack the fine-grained alignment necessary for effective step grounding. In contrast, using specialized video-language pre-trained models significantly improves performance. For instance, EgoVideo from (Pei et al., 2024) achieves 19.90 R@1mIoU-IndOrder=0.3 and 11.29 mIoU=0.5, as it is trained via contrastive learning to align video and text embeddings, producing more representative features that enhance the subsequent temporal segmentation by our Bayesian-VSLNet head. Next, we combined video features from different models to obtain a more comprehensive video representation, leading to the best performance for both VSLNet (18.19 R@1mIoU-IndOrder=0.3 and 13.55 mIoU=0.5) and Bayesian-VSLNet (21.05 R@1mIoU-IndOrder=0.3 and 13.50 mIoU=0.5). These results further highlight the advantages of our Bayesian-VSLNet approach, which consistently outperforms its VSLNet counterpart, demonstrating its effectiveness in improving step grounding accuracy.

Further, we also present a comparison of the number of sampled segments in Table 4. As the results show, the optimal configuration, Bayesian-VSLNet++, is achieved with 1024 segments. This configuration reports 23.75 mIoU=0.3 and 14.41 mIoU=0.5. This is a good balance between the model complexity and the information loss due to the diffusion of individual video features in the segment. We note that our Bayesian strategy primarily improves recall by enforcing step consistency, which is better captured at the more tolerant 0.3 mIoU threshold. At 0.5 mIoU, however, the stricter boundary requirement makes the discretization imposed by K a limiting factor, since even small temporal misalignments can drop the IoU below the threshold.

Lastly, we provide another configuration study of both α and β hyper-parameters from Bayesian-VSLNet++ configuration in Table 3. The best results were obtained with a percentile threshold of $\alpha = 95$ and $\beta = 0.1$. The high α value ensures the selection of segments that are not excessively long, even after the smoothing effect introduced by an intermediate β covariance value in the prior, which reduces probability differences between consecutive segments.

6.3. Qualitative results

Fig. 6 presents qualitative results from our best-performing configuration, Bayesian-VSLNet++, across five egocentric videos covering different activities, illustrating the impact of our novel temporal priors. The figure demonstrates how Bayesian-VSLNet predicts the probability of each step description corresponding to a given video segment. Notably, similar descriptions (e.g., *"Adds minced cocoa into milk?"* vs. *"Mix minced cocoa and milk together?"* in the first example) yield nearly identical predictions, highlighting the challenge of achieving precise segmentations. As shown in the fourth example, when a video contains multiple repeated procedural actions, an effective temporal prior becomes crucial for accurately segmenting steps. Our temporal ordering

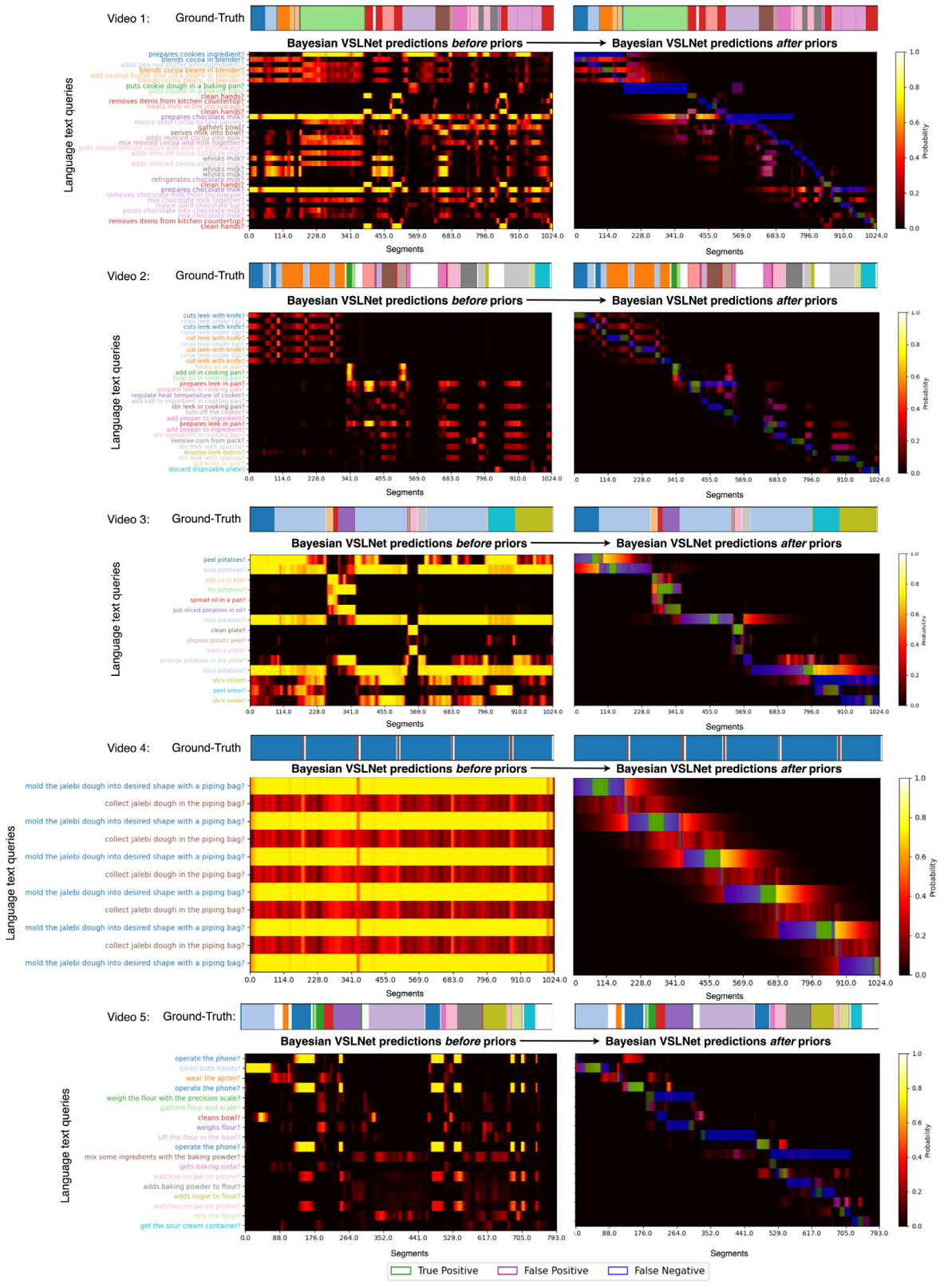


Fig. 6. Bayesian-VSLNet++ qualitative results on the Ego4D Goal-Step dataset. First, we present the Ground-Truth *step clip* for each language description, where identical descriptions share the same color. The plots in the left column show the predicted probabilities by our Bayesian-VSLNet per each step description, while the plots in the right column display the refined probabilities after applying our temporal-order prior. The final *predicted step clips* is extracted from the refined probabilities. We report the true positive segments in green, the false positives in purple and the false negatives in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Ablation study about number of segments K and feature extractors in the Ego4D Goal-Step validation set for both VSLNet and Bayesian-VSLNet architectures. We leverage Omnivore-L (Girdhar et al., 2022), BERT (Devlin et al., 2018), Ego-VLP (Pramanick et al., 2023) and EgoVIDEO (Pei et al., 2024) features. We take `R@1,mIoU-IndOrd` at 0.3 and 0.5 as reference metric.

Model		Video features			Text features			K	mIoU-IndOrd	
Name	Config	Omnivore-L	Ego-VLP	Ego-Video	BERT	Ego-VLP	Ego-Video		0.3	0.5
VSLNet		✓	–	–	✓	–	–	128	11.77	7.77
VSLNet		✓	✓	–	–	✓	–	512	16.26	11.81
VSLNet		–	–	✓	–	–	✓	512	17.74	12.32
VSLNet		✓	✓	✓	–	–	✓	512	18.19	13.55
VSLNet		✓	✓	✓	-	-	✓	1024	16.94	12.68
Bayesian-VSLNet		✓	–	–	✓	–	–	128	15.20	7.32
Bayesian-VSLNet	v0	✓	✓	–	–	✓	–	512	18.15	8.97
Bayesian-VSLNet		–	–	✓	–	–	✓	512	19.90	11.29
Bayesian-VSLNet		✓	✓	✓	–	–	✓	512	21.05	11.53
Bayesian-VSLNet	++	✓	✓	✓	–	–	✓	1024	**23.75**	**14.41**
Bayesian-VSLNet		✓	✓	✓	–	–	✓	2048	22.27	13.70

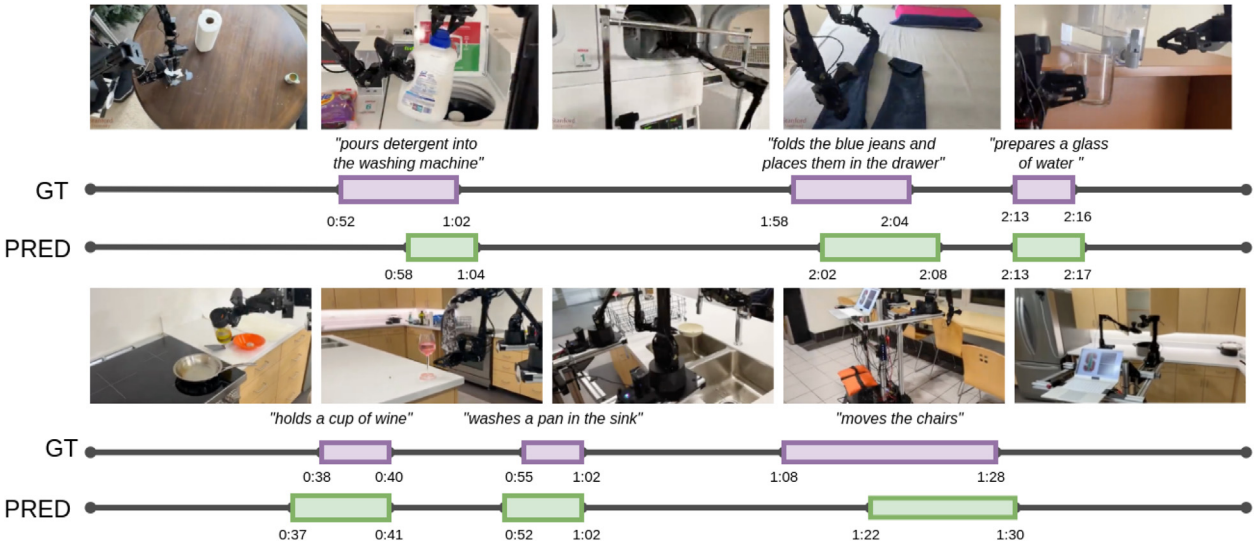


Fig. 7. Qualitative examples in real-world robotics scenarios. Bayesian-VSLNet predicts with high precision the moment associated to the provided step description without fine-tuning on this type of data. Video sourced from the Mobile Aloha project (Fu et al., 2024).

mechanism refines the network's probability predictions, aligning them with the correct temporal sequence to enhance segmentation accuracy. However, the sparse-sampling technique poses a limitation, as it causes the model to struggle with short atomic actions due to the loss of fine-grained temporal information during sampling.

6.4. Results in other datasets

To further assess the generality of our approach, we evaluate Bayesian-VSLNet on other widely used temporal video grounding benchmarks: TACoS (Regneri et al., 2013), a cooking-centric dataset with long untrimmed videos and fine-grained textual queries; and ActivityNet Captions (Krishna et al., 2017), a large-scale dataset of open-domain videos paired with temporally localized captions describing diverse and complex events. These benchmarks complement Ego4D by covering shorter videos with structured procedural tasks (TACoS), and open-domain scenarios (ActivityNet Captions).

As shown in Table 5, our method consistently outperforms the VSLNet baseline across all datasets. While the model shows solid gains in all the datasets, we remark that its design is particularly suited for long untrimmed videos, where step grounding becomes most challenging—akin to finding a needle in a haystack when the temporal window of the text query is very short compared to the full video length.

Table 5
Comparison between VSLNet and BayesianVSLNet across two other datasets using mIoU-IndOrder metric with R1 at 0.3, 0.5, and mIoU.

Model	TACoS			ActivityNet caption		
	IoU=0.3	IoU=0.5	mIoU	IoU=0.3	IoU=0.5	mIoU
VSLNet	29.61	24.27	20.03	63.16	43.22	26.16
BayesianVSLNet++	**40.29**	**25.52**	**27.56**	**67.66**	**44.86**	**29.07**

6.5. Qualitative results on assistive robotics data

We present qualitative results in a real-world assistive robotics scenario to demonstrate the potential of our approach in enhancing human–robot interaction in practical applications. Specifically, we used two near-egocentric videos from the Mobile Aloha project (Fu et al., 2024), each about 2 min long, featuring a robot performing household chores with approximately 30 and 15 steps, respectively. We queried the model with the free-form description of three different steps, without any prior fine-tuning on this data. Our approach achieves efficient execution, with an inference time of 125 ms per sample, ensuring real-time processing capabilities and supporting embedded applications as Fig. 7 shows.

7. Conclusions

Video understanding is a key perception task for assistive applications, where robotic platforms or wearable devices frequently rely on video data to function. In this paper, we presented Bayesian-VSLNet, a fully probabilistic formulation of step grounding in long, untrimmed videos. Concretely, the network outputs a likelihood distribution over video segments, which is then refined with a temporal prior to produce a posterior distribution that guides inference. This posterior enables the model to pinpoint the queried step even when it appears multiple times, directly addressing the challenges posed by repetitive and cyclic actions.

Our main limitation comes in very long videos, where a sparse-sampling technique is not enough. We achieve the state-of-the art on the Ego4D Goal-Step dataset, where this approach won the step grounding challenge at the CVPR 2024 EgoVis workshop—and further validate its generality on additional benchmarks. Finally, we demonstrate qualitative results on real-world robotics data, showing its potential for more natural interactions and robot learning in assistive applications.

CRediT authorship contribution statement

Carlos Plou: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lorenzo Mur-Labadia:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jose J. Guerrero:** Resources, Funding acquisition. **Ruben Martinez-Cantin:** Writing – review & editing, Validation, Supervision, Resources, Project administration. **Ana C. Murillo:** Writing – review & editing, Validation, Supervision, Resources, Project administration.

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.

Acknowledgments

This work was supported by PID2024-159284NB-I00, PID2021-125514NB-I00, PID2024-158322OB-I00, PID2021-125209OB-I00, and AIA2025-1635 grants funded by MCIN/AEI/10.13039/501100011033 ERDF/NextGenerationEU/PRTR, and two DGA scholarships and project T45_23R.

Data availability

All data used in this work is publicly available as described in the abstract.

References

Aytar, Y., Pfaff, T., Budden, D., Paine, T., Wang, Z., De Freitas, N., 2018. Playing hard exploration games by watching youtube. Adv. Neural Inf. Process. Syst. 31.

Bahl, S., Gupta, A., Pathak, D., 2022. Human-to-robot imitation in the wild. arXiv preprint arXiv:2207.09450.

Bain, M., Nagrani, A., Varol, G., Zisserman, A., 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738.

Baker, B., Akkaya, I., Zhokov, P., Huizinga, J., Tang, J., Ecoffet, A., Houghton, B., Sampedro, R., Clune, J., 2022. Video pretraining (vpt): Learning to act by watching unlabeled online videos. Adv. Neural Inf. Process. Syst. 35, 24639–24654.

Ballester, I., Fontán, A., Civera, J., Strobl, K.H., Triebel, R., 2021. DOT: Dynamic object tracking for visual SLAM. In: 2021 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 11705–11711.

Bansal, S., Arora, C., Jawahar, C., 2022. My view is the best view: Procedure learning from egocentric videos. In: European Conference on Computer Vision. Springer, pp. 657–675.

Bärmann, L., Waibel, A., 2022. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1560–1568.

Bonanno, C., Ragusa, F., Furnari, A., Farinella, G.M., 2023. Hero: A multi-modal approach on mobile devices for visual-aware conversational assistance in industrial domains. In: International Conference on Image Analysis and Processing. Springer, pp. 424–436.

Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J., 2015. Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition. pp. 961–970.

Cao, Z., Wang, Z., Xie, S., Liu, A., Fan, L., 2024. Smart help: Strategic opponent modeling for proactive and adaptive robot assistance in households. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18091–18101.

Capi, G., Kitani, M., Ueki, K., 2014. Guide robot intelligent navigation in urban environments. Adv. Robot. 28 (15), 1043–1053.

Chang, M., Gupta, A., Gupta, S., 2020. Semantic visual navigation by watching youtube videos. Adv. Neural Inf. Process. Syst. 33, 4283–4294.

Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al., 2018. Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 720–736.

De Rossi, G., Minelli, M., Roin, S., Falezza, F., Sozzi, A., Ferraguti, F., Setti, F., Bonfè, M., Secchi, C., Muradore, R., 2021. A first evaluation of a multi-modal learning system to control surgical assistant robots via action segmentation. IEEE Trans. Med. Robot. Bionics 3 (3), 714–724.

Deng, Z., Li, X., Li, X., Tong, Y., Zhao, S., Liu, M., 2024. VG4d: Vision-language model goes 4D video recognition. In: 2024 IEEE International Conference on Robotics and Automation. ICRA.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dionigi, A., Devo, A., Guiducci, L., Costante, G., 2022. E-vat: An asymmetric end-to-end approach to visual active exploration and tracking. IEEE Robot. Autom. Lett. 7 (2), 4259–4266.

Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J., Schuurmans, D., Abbeel, P., 2024. Learning universal policies via text-guided video generation. Adv. Neural Inf. Process. Syst. 36.

Escontrela, A., Adeniji, A., Yan, W., Jain, A., Peng, X.B., Goldberg, K., Lee, Y., Hafner, D., Abbeel, P., 2024. Video prediction models as rewards for reinforcement learning. Adv. Neural Inf. Process. Syst. 36.

Fang, X., Liu, D., Fang, W., Zhou, P., Xu, Z., Xu, W., Chen, J., Li, R., 2024. Fewer steps, better performance: Efficient cross-modal clip trimming for video moment retrieval using language. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, (2), pp. 1735–1743.

Fard, M.J., Ameri, S., Chinnam, R.B., Ellis, R.D., 2016. Soft boundary approach for unsupervised gesture segmentation in robotic-assisted surgery. IEEE Robot. Autom. Lett. 2 (1), 171–178.

Fernando, B., Gavves, E., Oramas, J.M., Ghodrati, A., Tuytelaars, T., 2015. Modeling video evolution for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5378–5387.

Flaborea, A., Di Melendugno, G.M.D., Plini, L., Scofano, L., De Matteis, E., Furnari, A., Farinella, G.M., Galasso, F., 2024. PREGO: online mistake detection in procedural egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18483–18492.

Fu, Z., Zhao, T.Z., Finn, C., 2024. Mobile ALOHA: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. arXiv.

Fukuda, T., Nakauchi, Y., Noguchi, K., Matsubara, T., 2005. Sequential human behavior recognition for cooking-support robots. J. Robot. Mechatronics 17 (6), 717.

Furnari, A., Farinella, G.M., 2020. Rolling-unrolling lstms for action anticipation from first-person video. IEEE Trans. Pattern Anal. Mach. Intell. 43 (11), 4021–4036.

Gao, J., Sarkar, B., Xia, F., Xiao, T., Wu, J., Ichter, B., Majumdar, A., Sadigh, D., 2024. Physically grounded vision-language models for robotic manipulation. In: 2024 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 12462–12469.

Girdhar, R., Singh, M., Ravi, N., Van Der Maaten, L., Joulin, A., Misra, I., 2022. Omnivore: A single model for many visual modalities. In: CVPR. pp. 16102–16112.

Goel, K., Brunskill, E., 2019. Learning procedural abstractions and evaluating discrete latent temporal structure. In: International Conference on Learning Representations.

Goyal, M., Modi, S., Goyal, R., Gupta, S., 2022. Human hands as probes for interactive object understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3293–3303.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al., 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR. pp. 18995–19012.

Guo, H., Wang, H., Ji, Q., 2024. Bayesian evidential deep learning for online action detection. In: European Conference on Computer Vision. Springer, pp. 283–301.

Hou, Z., Ji, L., Gao, D., Zhong, W., Yan, K., Li, C., Chan, W.K., Ngo, C.W., Duan, N., Shou, M.Z., 2023. Groundnlq@ ego4d natural language queries challenge 2023. arXiv preprint arXiv:2306.15255.

Hou, Z., Zhong, W., Ji, L., Gao, D., Yan, K., Chan, W.K., Ngo, C.W., Shou, Z., Duan, N., 2022. An efficient coarse-to-fine alignment framework@ ego4d natural language queries challenge 2022. arXiv preprint arXiv:2211.08776.

Karnan, H., Warnell, G., Xiao, X., Stone, P., 2022. Voila: Visual-observation-only imitation learning for autonomous navigation. In: 2022 International Conference on Robotics and Automation. ICRA, IEEE, pp. 2497–2503.

Kothandaraman, D., Lin, M., Manocha, D., 2023. Diffar: Differentiable frequency-based disentanglement for aerial video action recognition. In: 2023 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 8254–8261.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Guestrin, C., 2017. Dense-captioning events in videos. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 706–715.

Kuehne, H., Arslan, A., Serre, T., 2014. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 780–787.

Kukleva, A., Kuehne, H., Sener, F., Gall, J., 2019. Unsupervised learning of action classes with continuous temporal embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12066–12074.

Kung, Y.C., Zhang, A., Wang, J., Biswas, J., 2024. Looking inside out: Anticipating driver intent from videos. In: 2024 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 5608–5614.

Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H., 2017. Unsupervised representation learning by sorting sequences. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 667–676.

Lei, J., Berg, T.L., Bansal, M., 2021. Detecting moments and highlights in videos via natural language queries. In: NeurIPS, vol. 34, pp. 11846–11858.

Li, J., Shi, X., Chen, F., Stroud, J., Zhang, Z., Lan, T., Mao, J., Kang, J., Refaat, K.S., Yang, W., et al., 2023. Pedestrian crossing action recognition and trajectory prediction with 3d human keypoints. In: 2023 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 1463–1470.

Liao, Y., Xie, J., Geiger, A., 2022. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Trans. Pattern Anal. Mach. Intell. 45 (3), 3292–3310.

Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., Xu, E.Z., Gao, D., Tu, R.C., Zhao, W., Kong, W., et al., 2022. Egocentric video-language pretraining. Adv. Neural Inf. Process. Syst. 35, 7575–7586.

Liu, Y., Chen, C., Wang, Z., Yi, L., 2024. CrossVideo: Self-supervised cross-modal contrastive learning for point cloud video understanding. In: 2024 IEEE International Conference on Robotics and Automation. ICRA.

Lu, Z., Elhamifar, E., 2024. FACT: Frame-action cross-attention temporal modeling for efficient action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 18175–18185.

Mai, J., Hamdi, A., Giancola, S., Zhao, C., Ghanem, B., 2023. Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 45–57.

Mazzamuto, M., Ragusa, F., Furnari, A., D'Ambra, I., Guarriera, A., Sorbello, A., Farinella, G.M., 2024. A mixed reality application to help impaired people re-habilitate outside clinical environments. In: 2024 IEEE International Conference on Metrology for EXtended Reality, Artificial Intelligence and Neural Engineering. MetroXRAINE, IEEE, pp. 42–47.

Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W., 2022. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robot. Autom. Lett. (RA-L) 7 (3), 7327–7334.

Meli, D., Fiorini, P., 2021. Unsupervised identification of surgical robotic actions from small non-homogeneous datasets. IEEE Robot. Autom. Lett. 6 (4), 8205–8212.

Meltzoff, A.N., 1993. The role of imitation in understanding persons and developing a theory of mind. In: Understanding Other Minds: Perspectives from Autism. Oxford University Press, pp. 335–366.

Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J., 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2630–2640.

Mur-Labadia, L., Guerrero, J.J., Martinez-Cantin, R., 2023. Multi-label affordance mapping from egocentric vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5238–5249.

Mur-Labadia, L., Martinez-Cantin, R., Guerrero, J.J., Farinella, G.M., Furnari, A., 2024. AFF-ttention! affordances and attention models for short-term object interaction anticipation. In: European Conference on Computer Vision. Springer, pp. 167–184.

Nagarajan, T., Li, Y., Feichtenhofer, C., Grauman, K., 2020. Ego-topo: Environment affordances from egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 163–172.

Nagarajan, T., Ramakrishnan, S.K., Desai, R., Hillis, J., Grauman, K., 2024. EgoEnv: Human-centric environment representations from egocentric video. NeurIPS, NeurIPS.vol. 36,

Pei, B., Chen, G., Xu, J., He, Y., Liu, Y., Pan, K., Huang, Y., Wang, Y., Lu, T., Wang, L., et al., 2024. EgoVideo: Exploring egocentric foundation model and downstream adaptation. arXiv preprint arXiv:2406.18070.

Pieroni, R., Specchia, S., Corno, M., Savaresi, S.M., 2024. Multi-object tracking with camera-LiDAR fusion for autonomous driving. arXiv preprint arXiv:2403.04112.

Plou, C., Gallego, N., Sabater, A., Montijano, E., Urcola, P., Montesano, L., Martinez-Cantin, R., Murillo, A.C., 2024. EventSleep: Sleep activity recognition with event cameras. URL https://arxiv.org/abs/2404.01801, arXiv:2404.01801.

Pramanick, S., Song, Y., Nag, S., Lin, K.Q., Shah, H., Shou, M.Z., Chellappa, R., Zhang, P., 2023. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In: ICCV. pp. 5285–5297.

Radevski, G., Grujicic, D., Blaschko, M., Moens, M.F., Tuytelaars, T., 2023. Multimodal distillation for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5213–5224.

Ramakrishnan, S.K., Al-Halah, Z., Grauman, K., 2023. Naq: Leveraging narrations as queries to supervise episodic memory. In: CVPR. pp. 6694–6703.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M., 2013. Grounding action descriptions in videos. In: Transactions of the Association for Computational Linguistics, vol. 1, pp. 25–36.

Schmeckpeper, K., Rybkin, O., Daniilidis, K., Levine, S., Finn, C., 2021. Reinforcement learning with videos: Combining offline observations with interaction. In: Conference on Robot Learning. PMLR, pp. 339–354.

Seo, M., Cho, D., Lee, S., Park, J., Kim, D., Lee, J., Ju, J., Noh, H., Choi, D.G., 2021. A self-supervised sampler for efficient action recognition: Real-world applications in surveillance systems. IEEE Robot. Autom. Lett. 7 (2), 1752–1759.

Sermanet, P., Ding, T., Zhao, J., Xia, F., Dwibedi, D., Gopalakrishnan, K., Chan, C., Dulac-Arnold, G., Maddineni, S., Joshi, N.J., Florence, P., Han, W., Baruch, R., Lu, Y., Mirchandani, S., Xu, P., Sanketi, P., Hausman, K., Shafran, I., Ichter, B., Cao, Y., 2024. RoboVQA: Multimodal long-horizon reasoning for robotics. In: 2024 IEEE International Conference on Robotics and Automation. ICRA, pp. 645–652. http://dx.doi.org/10.1109/ICRA57147.2024.10610216.

Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D., 2023. Tridet: Temporal action detection with relative boundary modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18857–18866.

Song, Y., Byrne, E., Nagarajan, T., Wang, H., Martin, M., Torresani, L., 2024. Ego4d goal-step: Toward hierarchical understanding of procedural activities. In: NeurIPS, vol. 36.

Subedar, M., Krishnan, R., Meyer, P.L., Tickoo, O., Huang, J., 2019. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6301–6310.

Vahdani, E., Tian, Y., 2022. Deep learning-based action detection in untrimmed videos: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 45 (4), 4302–4320.

Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al., 2022. Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191.

Wang, X., Xian, R., Guan, T., de Melo, C.M., Nogar, S.M., Bera, A., Manocha, D., 2023. Aztr: Aerial video action recognition with auto zoom and temporal reasoning. In: 2023 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 1312–1318.

Wong, B., Chen, J., Wu, Y., Lei, S.W., Mao, D., Gao, D., Shou, M.Z., 2022. Assistq: Affordance-centric question-driven task completion for egocentric assistant. In: European Conference on Computer Vision. Springer, pp. 485–501.

Yang, F., Odashima, S., Masui, S., Jiang, S., 2023. Is weakly-supervised action segmentation ready for human-robot interaction? No, let's improve it with action-union learning. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 9800–9807.

Yang, J., Yu, E., Li, Z., Li, X., Tao, W., 2022. Quality matters: Embracing quality clues for robust 3d multi-object tracking. arXiv preprint arXiv:2208.10976.

Yao, L., Fu, C., Li, S., Zheng, G., Ye, J., 2023. SGDViT: saliency-guided dynamic vision transformer for UAV tracking. In: 2023 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 3353–3359.

Yi, Z., et al., 2022. Unified fully and timestamp supervised temporal action segmentation via sequence-to-sequence learning. In: European Conference on Computer Vision. ECCV.

Zhang, H., Sun, A., Jing, W., Zhen, L., Zhou, J.T., Goh, R.S.M., 2021. Natural language video localization: A revisit in span-based question answering framework. IEEE Trans. Pattern Anal. Mach. Intell. http://dx.doi.org/10.1109/TPAMI.2021.3060449.

Zhang, H., Sun, A., Jing, W., Zhou, J.T., 2020. Span-based localizing network for natural language video localization. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 6543–6554, URL https://www.aclweb.org/anthology/2020.acl-main.585.

Zhang, C.L., Wu, J., Li, Y., 2022. Actionformer: Localizing moments of actions with transformers. In: European Conference on Computer Vision. Springer, pp. 492–510.

Zhu, C., Sheng, W., 2011. Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living. IEEE Trans. Syst. Man, Cybernetics-Part A: Syst. Humans 41 (3), 569–573.