SAGE: SYNCHRONIZED ACTION-GAZE RECOGNITION AND ANTICIPATION FOR HUMAN BEHAVIOR UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Human object interaction (HOI), gaze pattern, and their anticipation are intricately linked, providing valuable insights into cognitive processes, intentions, and behavior. However, most existing models handle gaze and actions separately, missing both their interdependence and the advantages of a unified solution. This paper presents a novel unified framework, SAGE (**S**ynchronized **A**ction-**G**az**E**), which integrates simultaneous recognition and anticipation of both HOI and human gaze into a single unified end-to-end trainable model. Our approach leverages a transformer-based architecture and incorporates gaze data into spatiotemporal attention mechanisms to simultaneously predict current and future human actions and gaze behavior. We explore this bidirectional relationship between gaze and actions under different scenarios, whether requiring a close-up, detailed view (egocentric) or a wider, more contextual view (exocentric), making our framework versatile for various applications. Additionally, due to lack of datasets for comprehensive analysis of both HOI and gaze in exocentric videos, we establish a new benchmark *Exo-Cook* to facilitate further research in this domain. Experiments on three benchmark datasets—VidHOI, EGTEA Gaze+, and Exo-Cook—show that jointly modeling gaze and actions across current and future frames achieves consistently strong results, often surpassing specialized state-of-the-art models tailored to individual tasks. By unifying actions and attention in a comprehensive way, our work lays the groundwork for more intuitive human-machine interaction.

1 INTRODUCTION

Consider a person preparing a meal with a skillet (Figure 4 in Appendix A.1). While placing an empty bowl on the table, their gaze remains on the skillet, revealing the object relevant to the current interaction and signaling that they may next transfer the food into the bowl. Conversely, anticipating this action suggests continued attention to the skillet. More generally, gaze and actions are strongly coupled: gaze provides spatial cues about ongoing interactions and temporal cues about future behavior. Humans naturally exploit this interplay to jointly understand current and future behavior. Enabling intelligent systems to do the same, rather than relying on separate task-specific pipelines, is important for natural human–machine interaction, with applications in proactive human–robot collaboration, assistive manufacturing, and AR/VR.

Yet, most existing approaches treat these tasks separately: some recognize HOI and actions (Ji et al., 2021; Cong et al., 2021; Tu et al., 2022; Chiou et al., 2021; Mascaro et al., 2023; Ni et al., 2023; Hao et al., 2022; Wang et al., 2020) or anticipate them (Roy et al., 2024; Girdhar & Grauman, 2021; Zhong et al., 2023; Liu et al., 2020; Cong et al., 2021; Ni et al., 2023), while others model gaze detection (Lai et al., 2024a; Li et al., 2021; Huang et al., 2018; Tafasca et al., 2024; Chong et al., 2020) or gaze anticipation (Lai et al., 2024b; Zhang et al., 2017). Few attempt to integrate both, and those that do either use disjointed pipelines or focus only on egocentric views (Li et al., 2015; Ma et al., 2016; Singh et al., 2016; Ni et al., 2023; Li et al., 2018; 2021; Min & Corso, 2021; Huang et al., 2020). Consequently, the *spatial* dependency between gaze and current human–object interactions and the *temporal* dependency between anticipated gaze and future actions remain insufficiently explored within a unified framework.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce *SAGE*, a unified framework for *S*ynchronized *A*ction and *Gaz**E*** recognition and anticipation. Rather than treating gaze and actions as independent outputs of a multi-task network, SAGE explicitly models their dependencies across both observed and future time. For current behavior, predicted gaze provides an object-level spatial prior for action understanding, capturing the correlation between where a person looks and the objects involved in the interaction. For future behavior, SAGE first anticipates gaze and uses it to condition action anticipation, explicitly modeling the temporal dependency between future attention and future actions. We instantiate these two dependencies through *Gaze-Conditioned Spatial Attention* (GCSA) and *Gaze-Conditioned Temporal Prediction* (GCTP), respectively. Importantly, these modules serve as mechanisms for realizing the proposed joint gaze–action formulation rather than independent task-specific predictors. The same formulation supports both egocentric and exocentric videos through view-specific feature construction while preserving the joint modeling framework.

Joint gaze–action learning also requires data in which both signals are aligned. Existing datasets containing both gaze and action annotations are predominantly egocentric (Li et al., 2021; Grauman et al., 2022). Although Ego-Exo4D (Grauman et al., 2024) provides synchronized egocentric and exocentric recordings with gaze and action information, its annotations are not directly structured for joint gaze–action recognition and anticipation. We therefore construct *Exo-Cook*, the first benchmark to enable joint gaze-action research in third-person settings. We transform and align gaze, action, human, and object annotations to provide 32,050 clips for joint gaze–action learning and evaluation (Section 3).

Our contributions are fourfold. **First**, we jointly model gaze and actions for both current and future, explicitly capturing their spatial and temporal dependencies and enabling gaze-conditioned action uncertainty analysis at inference. **Second**, we instantiate these dependencies through our proposed GCSA and GCTP modules, which use gaze to condition current human–object interaction understanding and future action anticipation, respectively. **Third**, unlike prior works focusing on either egocentric or exocentric views, our modular framework supports and is evaluated on both through separate training, enabling broader applicability. **Fourth**, we demonstrate SAGE's efficacy against state-of-the-art methods and introduce Exo-Cook, a new benchmark for joint HOI and gaze analysis in third-person videos.

2 Related Works

2.1 Gaze Detection and Anticipation

Gaze detection estimates a person's current visual attention and has been studied in both egocentric and exocentric settings. Following early gaze-following datasets (Recasens et al., 2015; 2017), existing methods have progressed from combining head and saliency cues (Chong et al., 2018; Horanyi et al., 2023) to incorporating richer scene information such as depth (Tafasca et al., 2023; Bao et al., 2022; Hu et al., 2023; Tafasca et al., 2024) and human pose (Gupta et al., 2022). Although gaze has been used to support action recognition (Huang et al., 2020; Li et al., 2018; 2021), exploiting action–gaze dependencies for joint prediction remains less explored. In contrast, SAGE explicitly models the spatial dependency between current gaze and actions.

Gaze anticipation predicts future shifts of visual attention and has primarily been studied in egocentric settings, where gaze often precedes object interactions. Existing methods (Lai et al., 2024b; Zhang et al., 2017) forecast gaze from temporal scene and motion cues but model it independently of future actions. However, these methods treat gaze anticipation as an isolated task and do not account for its interplay with actions. More importantly, to our knowledge, no existing work has explored gaze anticipation in exocentric videos—leaving a critical gap in third-person behavior modeling.

2.2 Recognizing and Anticipating Actions

To infer actions from appearance and motion, two-stream models such as I3D-2Stream (Li et al., 2021) and R34-2Stream (Sudhakaran & Lanz, 2018) capture RGB and motion cues, while SAP (Wang et al., 2020), GC-TSM (Hao et al., 2022), and ACE (Ghoddoosian et al., 2025) improve egocentric and procedural understanding. Gaze complements these visual-temporal methods by highlighting task-relevant regions, from early saliency or gaze based approaches with handcrafted features (Wang et al., 2016; Chen et al., 2014; Mathe & Sminchisescu, 2012; Shapovalova et al.,

2013) to recent HOI models that integrate gaze heatmaps into multimodal transformers (Ni et al., 2023); related attention-based methods instead learn such regions without explicit gaze supervision (Sharma et al., 2016).

Meanwhile, action anticipation aims to forecast future behavior from partial observations. Existing methods often model HOIs using spatio-temporal graphs (Jain et al., 2016; Materzynska et al., 2020; Ou et al., 2022; Teng et al., 2021; Wang & Gupta, 2018), supported by rich datasets (Girdhar & Grauman, 2021; Damen et al., 2022; Teed & Deng, 2020) and architectural advances like AVT (Girdhar & Grauman, 2021), MeMViT (Wu et al., 2022), RAFTformer (Girase et al., 2023), UADT (Guo et al., 2024), InAViT (Roy et al., 2024), AGA (Tai et al., 2026) as well as several other methods (Wang et al., 2024; Mittal et al., 2024; Zhao et al., 2024; Zhang et al., 2024). Recent trends also explore goal-conditioned reasoning (Roy & Fernando, 2024), motion primitives (Dessalene et al., 2023), and multimodal cues such as audio (Zhong et al., 2023).

2.3 Joint Action-Gaze Modeling

Prior works have explored joint gaze–action modeling primarily in egocentric videos. Li et al. (2018; 2021) and Huang et al. (2020) jointly estimate gaze and actions, using predicted gaze to spatially guide visual features for action recognition. However, these methods focus on observed gaze–action relations and do not explicitly associate gaze with individual humans in exocentric scenes. More recently, Özdel et al. (2024) use observed gaze in exocentric view to guide action anticipation through a visual-semantic graph, but do not anticipate gaze jointly with future actions. Yet, despite its predictive power, gaze remains underutilized in action anticipation. No existing approach models the temporal link between future gaze and future actions, or between current gaze-action modeling with future gaze-action.

3 Exo-Cook Benchmark

Few datasets provide both gaze and action annotations, with existing joint gaze–action benchmarks largely restricted to egocentric views (Li et al., 2021; Huang et al., 2020). Ego-Exo4D (Grauman et al., 2024) provides synchronized egocentric and multiple exocentric views with rich annotations, but lacks a benchmark for joint gaze–action learning. We therefore construct *Exo-Cook* from its "cooking" domain, where gaze is strongly associated with object-centric actions, enabling joint gaze–action recognition and anticipation from exocentric views. The other seven domains exhibit weaker gaze–action correspondence (e.g., soccer, basketball) or more diffuse gaze patterns (e.g., music). This choice is also consistent with prior works (Li et al., 2021; Huang et al., 2020), which evaluate exclusively on cooking datasets. Starting from 658 cooking videos and 189,225 textual action descriptions, we construct Exo-Cook through the following pipeline - **(a) Human and object localization:** We detect the camera wearer's body and head following (Ni et al., 2023) and obtain interactable-object bounding boxes from Ego-Exo4D instance masks. **(b) Gaze projection:** We compute 3D gaze intersections from egocentric gaze measurements, transform them into each exocentric camera using the provided camera poses, and project them onto the image plane as Gaussian gaze heatmaps. **(c) Action labeling:** We extract candidate verb–noun pairs from textual action descriptions using spaCy (Explosion, 2024), select semantically relevant pairs using BERT (Devlin et al., 2019) similarity, and cluster their embeddings into 10 action categories (Figure 7 in Appendix A.3.3), followed by manual verification. **(d) Temporal alignment:** We align gaze, action, and human–object annotations with video timestamps to construct joint gaze–action clips, while tracking human–object trajectories across frames. Exo-Cook contains **32,050** valid clips from five exocentric cameras, each annotated with gaze, action, and human–object information. We use 25,650 clips for training, 3,200 for validation, and 3,200 for testing. The split assignment follows official Ego-Exo4D (Grauman et al., 2024) split setting, with clips from the same source recording restricted to a single split. Detailed preprocessing, annotation validation, class/viewpoint statistics, and semantic consistency analysis are provided in Appendix A.3.

4 SAGE

Given an observed video clip $I_{0:t}$, SAGE jointly models the current action y_t and future actions $\boldsymbol{y}_{t+1:t+M}$. Here, y denotes an action represented as either an HOI category (exocentric) or a verb–

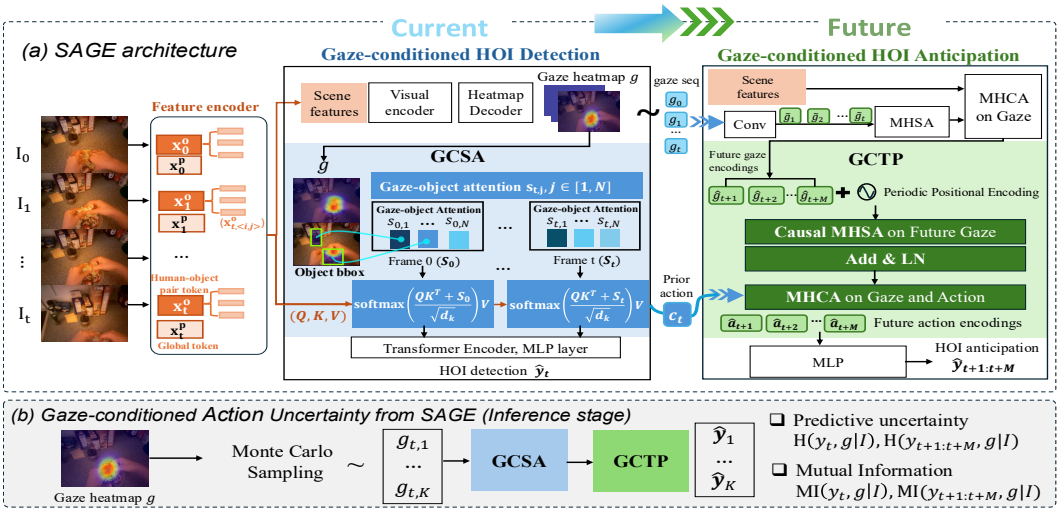


Figure 1: **Overview of SAGE.** We first present (a) SAGE architecture, which jointly performs gaze detection, HOI detection, and their respective anticipation. A feature encoder extracts human–object interaction tokens, while the gaze detection module predicts a probabilistic heatmap used by **GCSA** for HOI detection, where **GCSA** replaces the standard multi-head self-attention (MHSA) operation within the Transformer encoder layer. For forecasting, a Gaze Anticipation module predicts future gaze from observed frames and past gaze, and **GCTP** models future HOIs conditioned on current interactions and anticipated gaze. We also show (b) gaze-conditioned action uncertainty analysis via sampling from the predicted gaze heatmap at inference.

noun pair (egocentric). Rather than predicting actions independently of gaze, we introduce the gaze sequence $\boldsymbol{g}_{0:t+M}$ as an explicit conditioning variable and marginalize over its distribution:

$$p(\boldsymbol{y}_{t:t+M} \mid I_{0:t}) = \int_G p(\boldsymbol{y}_{t:t+M} \mid G, I_{0:t}) p(G \mid I_{0:t})\, dG, \tag{1}$$

where $G \triangleq \boldsymbol{g}_{0:t+M}$ and $\boldsymbol{y}_{t:t+M} = [y_t, \ldots, y_{t+M}]$.

We decompose Eq. 1 into *present* $(0{:}t)$, where visual observations are available, and *future* $(t{+}1{:}t{+}M)$, where actions and gaze must be anticipated. This yields four components:

$$\underbrace{p(\boldsymbol{g}_{0:t} \mid I_{0:t})}_{\text{SAGE-1: current gaze}},\ \underbrace{p(y_t \mid \boldsymbol{g}_{0:t}, I_{0:t})}_{\text{SAGE-2: current action}},\ \underbrace{p(\boldsymbol{g}_{t+1:t+M} \mid \boldsymbol{g}_{0:t}, I_{0:t})}_{\text{SAGE-3: future gaze}},\ \underbrace{p(\boldsymbol{y}_{t+1:t+M} \mid y_t, \boldsymbol{g}_{t+1:t+M}, I_{0:t})}_{\text{SAGE-4: future action}}.$$

SAGE-1 and SAGE-3 model current and future gaze, while SAGE-2 and SAGE-4 explicitly condition current and future action inference on the predicted gaze. For SAGE-4, we adopt $p(\boldsymbol{y}_{t+1:t+M} \mid y_t, \boldsymbol{g}_{0:t+M}, I_{0:t}) \approx p(\boldsymbol{y}_{t+1:t+M} \mid y_t, \boldsymbol{g}_{t+1:t+M}, I_{0:t})$, assuming that future gaze summarizes the gaze information relevant to future actions given the current action representation and observed video. Conditioning on the full gaze sequence does not improve performance (Appendix A.5.6).

SAGE also provides an additional benefit at inference. Eq. 1 allows action predictions to be marginalized over gaze uncertainty. We apply Monte-Carlo sampling during inference to generate K gaze perturbations in a single forward pass: $G^k = (\boldsymbol{g}^k_{0:t}, \boldsymbol{g}^k_{t+1:t+M})$, then we compute sample average for current and future actions: $p(\boldsymbol{y}_{t:t+M} \mid I_{0:t}) \approx \frac{1}{K} \sum_{k=1}^{K} p(y_t \mid \boldsymbol{g}^k_{0:t}, I_{0:t})\, p(\boldsymbol{y}_{t+1:t+M} \mid y_t, \boldsymbol{g}^k_{t+1:t+M}, I_{0:t})$. We use this Monte Carlo approximation during inference for action uncertainty estimation and action-gaze sensitivity analysis (Figure 1 (b)); details are provided in Section 5.4 and Appendix A.2.4.

4.1 Feature Encoding and Viewpoint Adaptability

We represent each frame by human–object interaction tokens $\mathbf{X}_t = [\mathbf{x}_{t,\langle i,j \rangle}; \mathbf{x}^p_t]$, where $\mathbf{x}_{t,\langle i,j \rangle}$ encodes the visual and spatial relationship between human patch i and object patch j, and $\mathbf{x}^p_t$ is a learnable global context token. Note that unlike prior works (Li et al., 2018; Huang et al., 2020; Li et al., 2021), the token construction is adapted to the viewpoint: TPV tokens encode human appear-

ance, location, and human–object relations, whereas FPV tokens encode hand–object relations with a single implicit human ($n_t^h = 1$). This viewpoint-specific encoding allows the subsequent joint gaze–action formulation to remain unchanged across FPV and TPV. Detailed token construction is provided in Appendix A.2.

4.2 Modeling Present: Gaze-Conditioned HOI Detection

We first model the present components SAGE-1 and SAGE-2. Given observed video $I_{0:t}$, SAGE-1 estimates current gaze $p(\mathbf{g}_{0:t} \mid I_{0:t})$. We adapt existing gaze detection models (Lai et al., 2024a; Chong et al., 2020; Tafasca et al., 2024; Ryan et al., 2025) into SAGE-1 for predicting gaze $\mathbf{g}_{0:t}$, where each g_t is a normalized heatmap representing a probabilistic distribution over gaze locations at time t. Rather than treating gaze and action as parallel prediction tasks, SAGE-2 explicitly models $p(y_t \mid \mathbf{g}_{0:t}, I_{0:t})$, such that the predicted gaze directly conditions the representation used for current action detection.

Gaze-Conditioned Spatial Attention (GCSA). Gaze is known as an important human attention cue for analyzing human action and intent. Prior joint gaze–action models (Li et al., 2018; Huang et al., 2020; Li et al., 2021) typically use gaze heatmaps to spatially reweight visual features at pixel level for action recognition. In contrast, we introduce GCSA, which integrates gaze into transformer attention through **object-level** gaze relations. For each object j in frame t with bounding box, SAGE assigns a gaze relevance score $s_{t,j}$ computed as $s_{t,j} = \mathrm{Norm}\,\{\sum_{(h,w)\in \mathrm{bbox}_{t,j}} g_t(h,w)\}$. which measures the probability mass of gaze falling inside the object region. Collecting all scores yields a gaze relevance vector $\mathbf{s}_t$, which is formulated into an additive attention bias matrix $\mathbf{S}_t \in \mathbb{R}^{(N+1)\times(N+1)}$, where N is the maximum number of human–object pair tokens. We inject the gaze bias $\mathbf{S}_t$ into the self-attention operation to form Gaze-Conditioned Spatial Attention:

$$\mathrm{GCSA}(Q_t, K_t, V_t, \mathbf{S}_t) = \mathrm{Softmax}\left(\frac{Q_t K_t^\top + \mathbf{S}_t}{\sqrt{d_k}}\right) V_t. \tag{2}$$

By augmenting the attention logits rather than directly masking features (Li et al., 2018; Huang et al., 2020; Li et al., 2021), GCSA biases the model towards interactions involving gaze-relevant objects while still preserving contextual cues from the full scene. GCSA replaces the standard multi-head self-attention (MHSA) operation within each Transformer encoder layer. Specifically, the gaze-conditioned attention bias is added to the attention logits before the softmax operation. Thus, gaze-relevant interactions receive higher attention while information from the full scene remains available. Applying GCSA to the interaction tokens yields the gaze-conditioned representation

$$\mathbf{c}_t = f_{\mathrm{spatial}}(\mathbf{X}_t, \mathbf{S}_t), \tag{3}$$

where f_{spatial} is implemented by stacking multiple transformer encoder layers (layer number N_d) equipped with GCSA in the MHSA layer.

HOI Detection. We use one transformer encoder layer to model the temporal feature in the gaze-aware features $\{\mathbf{c}_0, \ldots, \mathbf{c}_t\}$. Since gaze biases the spatial interactions at each frame, the temporal encoder propagates gaze-informed cues over time. Following by an MLP classifier $\phi(\cdot)$, the action probability is predicted as $p(y_t \mid \boldsymbol{g}_{0:t}, I_{0:t}) = \mathrm{Softmax}(\phi(\mathrm{TransformerEnc}_{N_d}([\mathbf{c}_0, \cdots, \mathbf{c}_t])))$.

4.3 Modeling Future: Gaze-conditioned HOI Anticipation

Prior joint gaze–action models such as (Li et al., 2021; Huang et al., 2020) focus on *only recognition*, that only model the correlation between gaze and HOI at the same time step but do not consider how gaze and HOI jointly evolve over time. In contrast, SAGE models the temporal relation between future gaze and HOI given current action. Under the factorization in Eq. 1, HOI anticipation is modeled as $p(\boldsymbol{y}_{t+1:t+M} \mid I_{0:t}) = \int p(\boldsymbol{y}_{t+1:t+M} \mid y_t, \boldsymbol{g}_{t+1:t+M}, I_{0:t})\, p(\boldsymbol{g}_{t+1:t+M} \mid \boldsymbol{g}_{0:t}, I_{0:t})\, dG$ which are decomposed into two steps.

Future Gaze Modeling. SAGE-3 estimates $p(\mathbf{g}_{t+1:t+M} \mid \mathbf{g}_{0:t}, I_{0:t})$ from observed gaze dynamics and interaction context. Each observed gaze heatmap is embedded into a latent representation with temporal positional encoding. The gaze sequence is temporally aggregated and conditioned on the observed human–object features to produce future gaze representations:

$$\hat{\mathbf{g}}_{t+1:t+M} = f_{\mathrm{gaze}}(\mathbf{g}_{0:t}, \mathbf{X}_{0:t}), \tag{4}$$

where f_{gaze} is implemented using temporal self-attention, cross-attention with interaction features, and a heatmap decoder. Architectural details are provided in Appendix A.2.3.

Gaze-Conditioned Temporal Prediction (GCTP). The key role of SAGE-4 is to explicitly realize $p(\mathbf{y}_{t+1:t+M} \mid y_t, \mathbf{g}_{t+1:t+M}, I_{0:t})$. Given the current gaze-conditioned interaction representation $\mathbf{c}_t$ from SAGE-2, GCTP integrates the anticipated gaze sequence $\hat{\mathbf{g}}_{t+1:t+M}$ with the current interaction context to construct future action representations:

$$\hat{\mathbf{a}}_{t+1:t+M} = f_{\mathrm{GCTP}}(\mathbf{c}_t, \hat{\mathbf{g}}_{t+1:t+M}), \tag{5}$$

followed by an action classifier to obtain $p(\mathbf{y}_{t+1:t+M} \mid y_t, \hat{\mathbf{g}}_{t+1:t+M}, I_{0:t})$. GCTP implements the conditional dependency in Eq. 5 by first encoding temporal relationships among anticipated gaze locations and then using them to condition the current interaction representation. We implement these operations with temporal self-attention and cross-attention in Transformer, respectively. By explicitly modeling future gaze as a conditioning variable, GCTP bridges current gaze–action recognition with future action anticipation. Similar to GCSA, GCTP mechanism can be applied to both egocentric and exocentric videos. To the best of our knowledge, this is the first framework that jointly models current and future gaze–action dependencies for HOI anticipation.

4.4 Loss function

SAGE is trained jointly for current and future gaze and action prediction. We use pixel-wise MSE (Tafasca et al., 2024; 2023) for current and future gaze heatmaps: $\mathcal{L}_{\mathrm{hm},1} = \sum_{m=0}^{t} \|\hat{g}_m - g_m^{\mathrm{gt}}\|_2^2$, $\mathcal{L}_{\mathrm{hm},2} = \sum_{m=t+1}^{t+M} \|\hat{g}_m - g_m^{\mathrm{gt}}\|_2^2$. We apply cross-entropy losses $\mathcal{L}_{\mathrm{act},1}$ and $\mathcal{L}_{\mathrm{act},2}$ for current and future action predictions, respectively. For exocentric videos, an additional binary cross-entropy loss $\mathcal{L}_{\mathrm{io}}$ (Chong et al., 2020; Tafasca et al., 2024) supervises whether the predicted gaze lies inside the frame. The overall training objective is

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{hm},1} + \lambda_2 \mathcal{L}_{\mathrm{hm},2} + \lambda_3 \mathcal{L}_{\mathrm{io}} + \lambda_4 \mathcal{L}_{\mathrm{act},1} + \lambda_5 \mathcal{L}_{\mathrm{act},2}. \tag{6}$$

The loss weights and additional training details are provided in Appendix A.2.5.

5 Experiments

Datasets & Metrics. To evaluate our framework, we use three datasets. **Vid-HOI** (Chiou et al., 2021), an exocentric dataset for video-based HOI detection and anticipation, provides annotated sequences of human–object interactions. **EGTEA Gaze+** (Li et al., 2018), widely used for egocentric gaze and action tasks, contains over 28 hours of video across 86 action classes from 32 participants. **Exo-Cook**, introduced in this work, is the first exocentric dataset curated for joint HOI and gaze analysis. Following (Ni et al., 2023), we evaluate HOI detection and anticipation on Vid-HOI and Exo-Cook using mean average precision (mAP), top-5 recall, precision, accuracy, and F1-score. For egocentric action recognition, we report Top-1 Accuracy on split 3 and the average on all three splits (Hao et al., 2022) for fair comparison. For action anticipation, we report mean class accuracy (Roy et al., 2024), while gaze detection and anticipation are evaluated with F1-score, recall, and precision.

Training proceeds in two stages: we first train SAGE-12 for 5 epochs on current gaze and action prediction, then initialize the full SAGE model with these weights and jointly train all components using current and future annotations for a total of 25 epochs. SAGE processes a 20-frame sequence in 0.63 s on an NVIDIA RTX A6000. Additional implementation and efficiency details are provided in Appendix A.2 and Appendix A.5.1.

5.1 Ablation Study

We conduct an ablation study on EGTEA Gaze+ in Table 1 to disentangle the effects of joint supervision, GCSA, and GCTP. Beyond SAGE-1–4, SAGE-12, SAGE-34, and full SAGE, we construct joint-training baselines that remove GCSA/GCTP while retaining identical multi-task supervision, isolating the effect of explicitly modeling gaze–action dependencies. For current gaze–action modeling, adding GCSA to SAGE-12 improves gaze detection and action recognition, demonstrating benefits beyond joint supervision. Similarly, adding GCTP to SAGE-34 substantially improves action anticipation while slightly improving gaze anticipation, showing that future gaze provides useful temporal cues for action anticipation. Finally, compared with a four-task joint-training baseline without GCSA and GCTP, full SAGE improves gaze detection, action recognition, gaze anticipation, and action anticipation. Thus, the performance gains are not only from multi-task supervision,

Table 1: **Ablation study of GCSA and GCTP** under task-based supervision on EGTEA Gaze+. "w/o" and "w/" denote without and with the module, respectively.

	Models	Gaze Detection			Action Recognition	Gaze Anticipation			Action Anticipation
		F1	Rec	Prec	Acc	F1	Rec	Prec	Acc
Current	SAGE-1	44.8	61.2	35.3	-	-	-	-	-
	SAGE-2	-	-	-	63.1	-	-	-	-
	SAGE-12 (w/o GCSA)	44.8	61.4	35.3	63.0	-	-	-	-
	SAGE-12 (w/ GCSA)	46.3	61.9	36.4	63.9				
Future	SAGE-3	-	-	-	-	35.4	49.5	28.2	-
	SAGE-4	-	-	-	-	-	-	-	47.3
	SAGE-34 (w/o GCTP)	-	-	-	-	34.8	49.4	27.4	44.6
	SAGE-34 (w/ GCTP)	-	-	-	-	35.2	49.4	27.5	52.6
	SAGE (w/o GCSA, GCTP)	44.8	61.4	35.3	63.1	34.5	48.8	26.7	44.0
	SAGE (w/ GCSA, GCTP)	**46.8**	**62.1**	**36.8**	**65.0**	**37.0**	**54.9**	**32.7**	**58.4**

Table 2: **SAGE results on Exo-Cook dataset.** †adapted for gaze anticipation.

Method	τ_a	Gaze Detection			HOI Detection					Gaze Anticipation			HOI Anticipation				
		F1	Rec	Prec	mAP	Rec	Prec	Acc	F1	F1	Rec	Prec	mAP	Rec	Prec	Acc	F1
VideoAttn (Chong et al., 2020)	-	55.6	72.8	54.8	-	-	-	-	-	-	-	-	-	-	-	-	-
Sharingan (Tafasca et al., 2024)	-	56.6	72.8	55.3	-	-	-	-	-	-	-	-	-	-	-	-	-
Gaze-LLE (Ryan et al., 2025)	-	56.8	72.4	55.0	-	-	-	-	-	-	-	-	-	-	-	-	-
ST-Gaze (Ni et al., 2023)	-	-	-	-	42.5	77.0	61.9	59.0	66.9	-	-	-	-	-	-	-	-
VideoAttn (Chong et al., 2020)†	-	-	-	-	-	-	-	-	-	42.2	54.6	42.5	-	-	-	-	-
Sharingan (Tafasca et al., 2024)†	-	-	-	-	-	-	-	-	-	40.1	52.7	40.4	-	-	-	-	-
Gaze-LLE (Ryan et al., 2025)†	-	-	-	-	-	-	-	-	-	38.6	50.3	36.2	-	-	-	-	-
ST-Gaze (Ni et al., 2023)	1	-	-	-	-	-	-	-	-	-	-	-	41.1	76.2	58.2	53.7	66.1
SAGE (VideoAttn)	1	55.8	73.2	**55.7**	44.2	79.0	62.6	58.5	68.9	48.5	62.1	49.6	42.4	76.6	59.0	54.6	66.9
SAGE (Sharingan)	1	**57.7**	**73.4**	55.5	**44.6**	**79.4**	**63.1**	**60.2**	**69.4**	**49.2**	**62.8**	**50.0**	**42.5**	**76.8**	**59.2**	**54.8**	**67.2**

but from explicitly modeling gaze–action dependencies through the proposed GCSA and GCTP modules.

5.2 SAGE ON EXOCENTRIC VIDEOS

Exo-Cook Dataset. We evaluate SAGE on four tasks in Table 2. Since no prior methods directly benchmark on Exo-Cook, we adapt existing models as baselines. For gaze detection and anticipation, we train and evaluate VideoAttn (Chong et al., 2020), Sharingan (Tafasca et al., 2024), and Gaze-LLE (Ryan et al., 2025), and integrate the first two into SAGE as SAGE (VideoAttn) and SAGE (Sharingan). For HOI detection and anticipation, we use ST-Gaze (Ni et al., 2023) as the first baseline. We set the anticipation horizon $\tau_a = 1s$. SAGE (Sharingan) outperforms ST-Gaze on HOI detection by +2.1 mAP and +1.0 accuracy while achieving the highest F1. It also achieves the best overall HOI anticipation performance. For gaze detection, SAGE (Sharingan) improves F1 over Sharingan by +1.1 points and achieves the best overall gaze anticipation performance. Comparing SAGE (VideoAttn) and SAGE (Sharingan) further shows that stronger gaze architectures improve joint performance. Qualitative results are shown in Figure 2, with additional results in Appendix A.6.
Vid-HOI Dataset. We take ST-Gaze (Ni et al., 2023) and STTran (Cong et al., 2021) as baseline models for HOI detection and anticipation in Table 3 and Table 4. Gaze is *not* evaluated on Vid-HOI as no gaze labels are provided. For HOI detection, SAGE achieves the best mAP, precision and F1 values. The improved performance of ST-Gaze* (Ni et al., 2023) in recall and accuracy is attributed to the use of word embeddings as an additional input modality. We individually analyze the impact of GCSA in Appendix A.5.4. For HOI anticipation, SAGE outperforms STTran (Cong et al., 2021) and ST-Gaze (Ni et al., 2023) across all anticipation horizons (1s, 3s, 5s) for mAP, precision, Accuracy and F1. Despite modest margins in Tables 3, 4, the gains exceed cross-seed variability (Appendix A.5.5) and remain meaningful as our unified model jointly solves multiple tasks.

5.3 SAGE ON EGOCENTRIC VIDEOS

We conduct a comprehensive evaluation on the EGTEA Gaze+ dataset (Li et al., 2018) in Table 5, covering four tasks as described below. For the anticipation setup, we use the standard horizon of $\tau_a = 0.5s$ and report results for additional horizons in Appendix A.5.3 and A.5.7. We also report cross-seed variance of SAGE in Appendix A.5.5.
Gaze Detection & Anticipation. We compare our gaze estimation performance against existing egocentric gaze estimation models, including GLC (Lai et al., 2024a), MViT (Fan et al., 2021), and I3D-R50 (Feichtenhofer et al., 2019). As shown in Table 5, SAGE achieves the best F1 score,

Table 3: **HOI detection** in oracle mode on Vid-HOI dataset. * includes word embedding module.

Method	τ_a	mAP	Rec	Prec	Acc	F1
STTran (Cong et al., 2021)	-	28.32	-	-	-	-
ST-Gaze* (Ni et al., 2023)	-	38.46	**73.62**	59.16	**53.76**	60.57
SAGE (Ours)	1	**38.65**	72.44	59.22	52.22	61.96
	3	38.18	72.21	**59.95**	52.18	**62.02**
	5	38.13	72.01	59.92	52.18	61.88

Table 4: **HOI anticipation** in oracle mode on Vid-HOI dataset. * includes word embedding module.

Method	τ_a	mAP	Rec	Prec	Acc	F1
STTran (Cong et al., 2021)	1	29.09	**74.76**	41.36	36.61	50.48
	3	27.59	**74.79**	40.86	36.42	50.16
	5	27.32	**75.65**	41.18	36.92	50.66
ST-Gaze* (Ni et al., 2023)	1	35.71	71.28	59.38	51.06	62.06
	3	32.19	71.09	60.17	51.63	62.39
	5	32.30	70.67	58.99	50.79	61.59
SAGE (Ours)	1	**37.71**	72.88	**60.24**	**51.86**	**62.72**
	3	**34.22**	72.36	**60.98**	**52.88**	**63.05**
	5	**32.64**	71.96	**59.88**	**51.46**	**62.14**

Table 5: **SAGE results on EGTEA Gaze+.** †adapted for gaze anticipation.

Method	Gaze Detection			Action Recognition			Gaze Anticipation			Action Anticipation
	F1	Rec	Prec	S1	S3 (Top-1)	Avg. (Top-1)	F1	Rec	Prec	Mean-cls Acc.
I3D-R50 (Feichtenhofer et al., 2019)	40.9	57.2	31.8	-	-	-	-	-	-	-
MViT (Fan et al., 2021)	43.0	57.8	35.4	-	-	-	-	-	-	-
GLC (Lai et al., 2024a)	44.8	61.2	35.3	-	-	-	-	-	-	-
I3D-2Stream (Li et al., 2021)	-	-	-	-	53.6	54.2	-	-	-	
R34-2Stream (Sudhakaran & Lanz, 2018)	-	-	-	-	58.6	60.8	-	-	-	-
SAP (Wang et al., 2020)	-	-	-	-	62.0	62.7	-	-	-	-
GC-TSM (Hao et al., 2022)	-	-	-	-	62.6	**65.1**	-	-	-	-
I3D-R50 (Feichtenhofer et al., 2019)†	-	-	-	-	-	-	34.3	46.5	29.6	-
MViT (Fan et al., 2021)†	-	-	-	-	-	-	31.5	44.8	28.5	-
CSTS-Visual (Lai et al., 2024b)	-	-	-	-	-	-	31.3	46.2	28.1	-
GLC (Lai et al., 2024a)†	-	-	-	-	-	-	32.8	48.5	28.7	-
AFFT (Zhong et al., 2023)	-	-	-	-	-	-	-	-	-	35.2
AVT (Girdhar & Grauman, 2021)	-	-	-	-	-	-	-	-	-	35.2
AGA (Tai et al., 2026)	-	-	-	-	-	-	-	-	-	37.4
MF (Patrick et al., 2021)	-	-	-	-	-	-	-	-	-	56.9
ORVIT-MF (Herzig et al., 2022)	-	-	-	-	-	-	-	-	-	57.2
InAViT (Roy et al., 2024)	-	-	-	-	-	-	-	-	-	58.2
MCN (Huang et al., 2020)	-	-	-	55.7	-	-	-	-	-	-
EGTEA (Li et al., 2021)	33.84	41.93	28.41	57.2	-	-	-	-	-	-
TSF-B (Zhao et al., 2023)	-	-	-	-	-	65.6	-	-	-	-
Ego-VPA (Wu et al., 2025)	-	-	-	-	-	73.4	-	-	-	-
LaViLa-L (Pei et al., 2025)	-	-	-	-	-	76.0	-	-	-	-
LaViLa (Zhao et al., 2023)	-	-	-	-	-	81.8	-	-	-	-
EgoVideo-G (Pei et al., 2025)	-	-	-	-	-	80.0	-	-	-	-
SAGE (Ours)	**46.8**	**62.1**	**36.8**	**61.8**	**63.1**	65.0	**37.0**	**54.9**	**32.7**	**58.4**

recall, and precision, with an average gain of 1.5 points over GLC across all metrics. For gaze anticipation, since no existing baseline exists on EGTEA Gaze+, we establish four baselines by re-training GLC, MViT, I3D-R50 and CSTS-Visual (Lai et al., 2024b)) with ground-truth future gaze heatmaps. SAGE significantly outperforms these baselines across F1, recall, and precision, with an average gain of 4.9 points, demonstrating the benefit of joint gaze-action modeling and establishing a strong baseline for gaze anticipation on EGTEA Gaze+. We study sensitivity of SAGE to different gaze models in Appendix A.5.2.

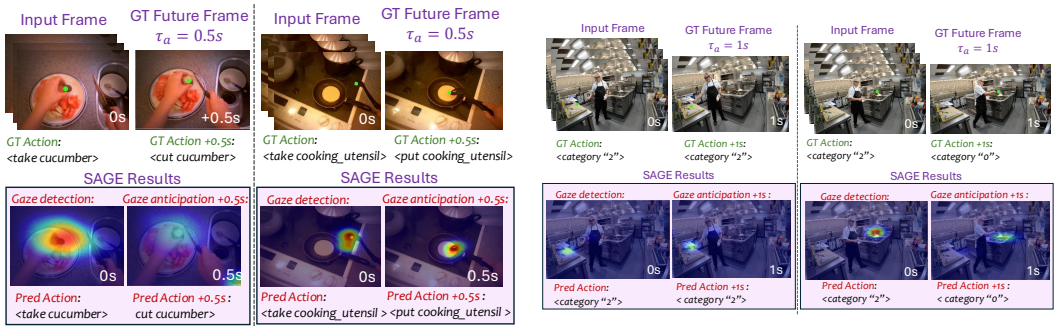


Figure 2: **Qualitative results.** SAGE on EGTEA Gaze+ (left) and Exo-Cook (right), including action and gaze predictions for current and future frames for two samples each. The green dot in first rwo represents ground truth gaze.

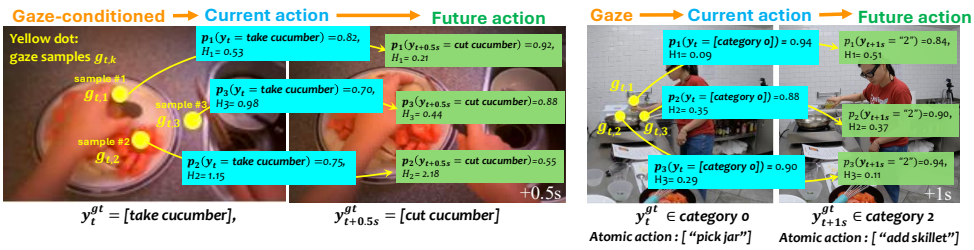


Figure 3: **Gaze-conditioned action uncertainty.** We show two examples of uncertainty quantification for current action y_t and future action y_{t+M}. Sampling three gaze locations produces varying probabilities and entropies: in EGTEA Gaze+ (left), both current and future HOIs are gaze-sensitive, whereas in Exo-Cook (right) predictions remain relatively stable, indicating robust action inference.

Action Recognition & Anticipation. In Table 5, we compare SAGE with state-of-the-art methods on EGTEA Gaze+ using average accuracy across three test splits. SAGE outperforms GCTSM (Hao et al., 2022) on S3 and achieves comparable average accuracy while outperforming other recent baselines. Since SAGE uses neither foundation models (LLMs or VLMs) nor large-scale external pretraining, we exclude recent models (e.g., LaViLa (Zhao et al., 2023), Ego-VPA (Wu et al., 2025), EgoVideo (Pei et al., 2025) etc) from direct comparison but report them in Table 5 for completeness. Moreover, unlike SAGE, these methods are limited to action recognition and do not report results on additional tasks. For action anticipation, the full SAGE model outperforms prior models including AVT (Girdhar & Grauman, 2021), and InAViT (Roy et al., 2024), achieving the highest mean-class accuracy of 58.4. Compared to transformer-based methods such as MF and AVT, SAGE demonstrates a clear advantage in capturing gaze-conditioned temporal dependencies for anticipation. We show qualitative results on EGTEA Gaze+ in Figure 2.

5.4 Action Uncertainty from Gaze

Li et al. (2021) first modeled gaze uncertainty for action recognition, but considers only gaze-distribution uncertainty for current actions, without explicitly quantifying action uncertainty or its effect on future predictions. In contrast, SAGE quantifies uncertainty for both current and future actions. As shown in Figure 1 (b), gaze locations are sampled through Monte Carlo sampling and propagate to action detection and anticipation through GCSA and GCTP module. Action uncertainty can be computed by marginalizing over the gaze samples. Given K Monte Carlo gaze samples $\{g^k\}_{k=1}^K$ from predicted gaze heatmap, we obtain action distributions $\mathbf{p}^k = p(y \mid g^k, I)$ and compute conditional entropy $H(y \mid I)$ and mutual information $\mathrm{MI}(y, g \mid I)$ for action. (1) The predictive distribution is approximated as $\bar{\mathbf{p}} = \frac{1}{K}\sum_{k=1}^K \mathbf{p}^k$. The total predictive uncertainty is measured by the entropy: $H(y \mid I) = -\sum_c \bar{p}_c \log \bar{p}_c$. (2) The mutual information (MI) of HOI predictions over gaze, represents how sensitive HOI prediction is to gaze uncertainty: $\mathbb{E}[H] = \frac{1}{K}\sum_{k=1}^K(-\sum_c p_c^k \log p_c^k)$, $\mathrm{MI}(y, g \mid I) = H(y \mid I) - \mathbb{E}[H]$. Figure 3 illustrates examples from both egocentric (EGTEA Gaze+) and exocentric (Exo-Cook) datasets, showing gaze-conditioned uncertainty for current (y_t) and future (y_{t+M}) actions. On EGTEA Gaze+, small gaze changes produce variations in prediction confidence and entropy for both current and future actions, whereas Exo-Cook exhibits substantially lower sensitivity to gaze. Additional uncertainty analysis is provided in Appendix A.6.1

6 Conclusion

We present SAGE, a unified end-to-end framework for joint recognition and anticipation of human actions and gaze. Through the proposed GCSA and GCTP modules, SAGE models gaze-action dependencies across current and future horizons for comprehensive human behavior understanding. Its modular design supports both egocentric and exocentric settings, validated on VidHOI, EGTEA Gaze+, and our new Exo-Cook benchmark. Experiments show that joint gaze-action modeling matches or surpasses specialized state-of-the-art methods across multiple tasks, providing a foundation for human behavior analysis and intuitive human–machine interaction.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and improving readability and presentation of the manuscript. We did not use generative AI to develop the proposed methodology, design experiments, generate data, or interpret experimental results. All AI-assisted text was reviewed and revised by the authors to ensure technical accuracy, consistency with the original research, and appropriate attribution. We take full responsibility for the final content of this work, including any material produced with the assistance of generative AI.

ETHICS STATEMENT

This work does not involve new human-subject data collection. Exo-Cook is derived exclusively from the publicly available Ego-Exo4D dataset Grauman et al. (2024), whose data were collected under established consent, privacy, and ethical protocols. We use the data solely for research on gaze and human-action understanding. While gaze and action modeling can benefit applications such as assistive systems and human–machine interaction, such models may also raise privacy concerns if applied to individuals without appropriate consent. We encourage responsible use consistent with applicable privacy and ethical standards.

REPRODUCIBILITY STATEMENT

Section 3 describes the construction of the Exo-Cook benchmark, including the preprocessing procedures and dataset distribution. Detailed preprocessing and annotation procedures, annotation validation, action clustering analysis, class/viewpoint statistics, and semantic consistency analysis are provided in Appendix A.3. Section 4 presents the SAGE formulation and model components, while layer-wise architecture details are provided in Appendix A.2.3. The modeling assumption used in SAGE-4 is examined through ablation studies in Appendix A.5.6. The Monte Carlo procedure used for gaze-conditioned uncertainty analysis is introduced in Section 5.4 and detailed steps are described in Appendix A.2.4. Complete implementation details and hyperparameters of SAGE are provided in Appendix A.2. Additional experimental results, ablation studies, and cross-seed variance analyses are included in the appendix. We will release the Exo-Cook benchmark and code upon acceptance.

REFERENCES

Jun Bao, Buyu Liu, and Jun Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14126–14135, 2022.

Wei Chen, Caiming Xiong, Ran Xu, and Jason J. Corso. Actionness ranking with lattice conditional ordinal random fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 748–755, 2014.

Meng-Jiun Chiou, Chun-Yu Liao, Li-Wei Wang, Roger Zimmermann, and Jiashi Feng. St-hoi: A spatial-temporal baseline for human-object interaction detection in videos. In *Proceedings of the 2021 ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, pp. 9–17, 2021.

Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *European Conference on Computer Vision*, pp. 383–398, 2018.

Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5396–5406, 2020.

Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *IEEE/CVF International Conference on Computer Vision*, pp. 16372–16382, 2021.

Dima Damen et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130:33–55, 2022.

Eadom Dessalene, Michael Maynord, Cornelia Fermüller, and Yiannis Aloimonos. Therbligs in action: Video understanding through motion primitives. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10618–10626. IEEE, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Explosion. spacy: Industrial-strength natural language processing in python. `https://github.com/explosion/spaCy`, 2024. Accessed: 2025-08-02.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 6824–6835, 2021.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18770–18780, 2022.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, 2019.

Reza Ghoddoosian, Nakul Agarwal, Isht Dwivedi, and Behzad Dariush. Ace: Action concept enhancement of video-language models in procedural videos. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 9503–9513, 2025.

Harshayu Girase, Nakul Agarwal, Chiho Choi, and Karttikeya Mangalam. Latency matters: Real-time action forecasting transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18759–18769, 2023.

Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *IEEE/CVF International Conference on Computer Vision*, pp. 13505–13515, 2021.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Hongji Guo, Nakul Agarwal, Shao-Yuan Lo, Kwonjoon Lee, and Qiang Ji. Uncertainty-aware action decoupling transformer for action anticipation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18644–18654, 2024.

Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5037–5046. IEEE, 2022.

Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. Group contextualization for video recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 928–938, 2022.

Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3148–3159, 2022.

Nora Horanyi, Linfang Zheng, Eunji Chong, Aleš Leonardis, and Hyung Jin Chang. Where are they looking in the 3d space? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2678–2687. IEEE, 2023.

Zhengxi Hu, Yuxue Yang, Xiaolin Zhai, Dingye Yang, Bohan Zhou, and Jingtai Liu. Gfie: A dataset and baseline for gaze-following from 2d to 3d in indoor environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8907–8916, June 2023.

Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *European Conference on Computer Vision*, pp. 754–769, 2018.

Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020.

Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5308–5317, 2016.

Jingwei Ji, Rishi Desai, and Juan Carlos Niebles. Detecting human-object relationships in videos. In *IEEE/CVF International Conference on Computer Vision*, pp. 8106–8116, 2021.

G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, TaoXie, J. Fang, imyhxy, K. Michael, V. Lorna, D. Montes, J. Nadar, Laughing, tkianai, yxNONG, P. Skalski, Z. Wang, A. Hogan, C. Fati, L. Mammana, AlexWang1900, D. Patel, D. Yiwei, F. You, J. Hajek, L. Diaconu, and M.T. Minh. Yolov5. `https://github.com/ultralytics/yolov5/`, 2022.

Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision*, 132(3):854–871, 2024a.

Bolin Lai, Fiona Ryan, Wenqi Jia, Miao Liu, and James M Rehg. Listen to look into the future: Audio-visual egocentric gaze anticipation. In *European Conference on Computer Vision*, pp. 192–210. Springer, 2024b.

Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 287–295, 2015.

Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *European Conference on Computer Vision*, pp. 619–635, 2018.

Yin Li, Miao Liu, and James M Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6731–6747, 2021.

Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *European Conference on Computer Vision*, pp. 704–721. Springer, 2020.

Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1894–1903, 2016.

Esteve Valls Mascaro, Daniel Sliwowski, and Dongheui Lee. Hoi4abot: Human-object interaction anticipation for human intention reading collaborative robots. In *Conference on Robot Learning*, 2023.

Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1049–1059, 2020.

Stefan Mathe and Cristian Sminchisescu. Dynamic eye movement datasets and saliency models for visual action recognition. In *European Conference on Computer Vision*, pp. 842–856, 2012.

Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1069–1078, 2021.

Himangi Mittal, Nakul Agarwal, Shao-Yuan Lo, and Kwonjoon Lee. Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18580–18590, 2024.

Zhifan Ni, Esteve Valls Mascaró, Hyemin Ahn, and Dongheui Lee. Human–object interaction prediction in videos through gaze following. *Computer Vision and Image Understanding*, 233: 103741, 2023.

Yangjun Ou, Li Mi, and Zhenzhong Chen. Object-relation reasoning graph for action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20101–20110. IEEE, 2022.

Süleyman Özdel, Yao Rong, Berat Mert Albaba, Yen-Ling Kuo, Xi Wang, and Enkelejda Kasneci. Gaze-guided graph neural network for action anticipation conditioned on intention. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, pp. 1–9, 2024.

Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12493–12506, 2021.

Baoqi Pei, Yifei Huang, Jilan Xu, Guo Chen, Yuping He, Lijin Yang, Yali Wang, Weidi Xie, Yu Qiao, Fei Wu, et al. Modeling fine-grained hand-object dynamics for egocentric video representation learning. In *International Conference on Learning Representations*, 2025.

Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *IEEE/CVF International Conference on Computer Vision*, pp. 1435–1443, 2017.

Debaditya Roy and Basura Fernando. Predicting the next action by modeling the abstract goal. In *International Conference on Pattern Recognition*, 2024.

Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer for egocentric action anticipation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6740–6750, 2024.

Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M. Rehg. Gaze-lle: Gaze target estimation via large-scale learned encoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Nataliya Shapovalova, Michalis Raptis, Leonid Sigal, and Greg Mori. Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization. In *Advances in Neural Information Processing Systems*, pp. 2407–2415, 2013.

Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. In *International Conference on Learning Representations Workshops*, 2016.

Suriya Singh, Chetan Arora, and CV Jawahar. First person action recognition using deep learned descriptors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2620–2628, 2016.

Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. In *British Machine Vision Conference*, 2018.

Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *IEEE/CVF International Conference on Computer Vision*, pp. 20935–20946, 2023.

Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Sharingan: A transformer architecture for multi-person gaze following. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2008–2017, 2024.

Tsung-Ming Tai, Sofia Casarin, Andrea Pilzer, Werner Nutt, and Oswald Lanz. Action-guided attention for video action anticipation. In *International Conference on Learning Representations*, volume 2026, pp. 50815–50838, 2026.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pp. 402–419, 2020.

Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *IEEE/CVF International Conference on Computer Vision*, pp. 13688–13697, 2021.

Danyang Tu, Wei Sun, Xiongkuo Min, Guangtao Zhai, and Wei Shen. Video-based human-object interaction detection from tubelet tokens. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23345–23357, 2022.

Limin Wang, Yu Qiao, Xiaoou Tang, and Luc Van Gool. Actionness estimation using hybrid fully convolutional networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2708–2717, 2016.

Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. In *European Conference on Computer Vision*, pp. 142–160, 2024.

Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Symbiotic attention with privileged information for egocentric action recognition. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 12249–12256, 2020.

Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *European Conference on Computer Vision*, pp. 399–417, 2018.

Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13587–13597. IEEE, 2022.

Tz-Ying Wu, Kyle Min, Subarna Tripathi, and Nuno Vasconcelos. Ego-vpa: Egocentric video understanding with parameter-efficient adaptation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9258–9268. IEEE, 2025.

Ce Zhang, Changcheng Fu, Shijie Wang, Nakul Agarwal, Kwonjoon Lee, Chiho Choi, and Chen Sun. Object-centric video representation for long-term action anticipation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6751–6761, 2024.

Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4372–4381, 2017.

Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? In *International Conference on Learning Representations*, volume 2024, pp. 56677–56697, 2024.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597, 2023.

Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6068–6077, 2023.

A Appendix

A.1 Motivation of SAGE

Consider a person preparing a meal with a skillet, as shown in Figure 4. Their gaze remains focused on the skillet while they place an empty bowl on the table. Even before any hand movement begins, this visual attention provides a strong cue that the next action is likely to involve picking up the skillet and transferring the food into the bowl. Their gaze is also expected to remain on the skillet, as it is the object involved in the upcoming action. Humans perform this reasoning effortlessly. By observing where someone is looking and how they interact with objects, we can simultaneously recognize current actions and anticipate future ones. These abilities arise from a unified cognitive process, highlighting the strong interdependence between gaze and action understanding. For intelligent systems to interact naturally with people, they must unify perception and prediction—understanding both current and future behavior without relying on separate task-specific pipelines. From assistive robots to driver monitoring and AR assistants, many applications require understanding current actions and anticipating future ones. Gaze reveals intention, while human–object interactions capture engagement with the environment. Together, they provide rich insight for understanding human behavior.

A.2 Implementation Details

A.2.1 Human-object-pair Feature Extraction.

We refer to ST-Gaze (Ni et al., 2023) to pre-process Vid-HOI and Exo-Cooking datasets. An object module is applied to process the video sequence and detects N bounding boxes $\{b_{t,j}\}$ along with their corresponding classes. Human bounding boxes are also detected. In addition, human head are detected by YOLO-v5 model (Jocher et al., 2022) and paired with their body bounding boxes. DeeoSort model is applied to associate these detections with previous detections to establish trajectories for each detected human $\{H_i\}$ and object $\{O_j\}$. Video feature are extracted by ResNet.

A.2.2 Viewpoint Adaptability.

As mentioned in main paper (Section 4.1), SAGE can adapt to egocentric and exocentric viewpoint with differing feature construction process. Egocentric (FPV) and exocentric (TPV) scenarios have typically been studied separately in gaze and action understanding, as their visual characteristics differ substantially and are rarely addressed within a single model. Consequently, prior work does not report results across both settings. To the best of our knowledge, SAGE is the first framework that supports both FPV and TPV within a unified architecture for four tasks: gaze detection, HOI detection, gaze anticipation, and HOI anticipation. The framework adapts viewpoint-specific modalities during feature encoding and gaze estimation while maintaining a consistent network flow and output representation. Note that this work does not claim cross-view generalization between FPV and TPV domains.

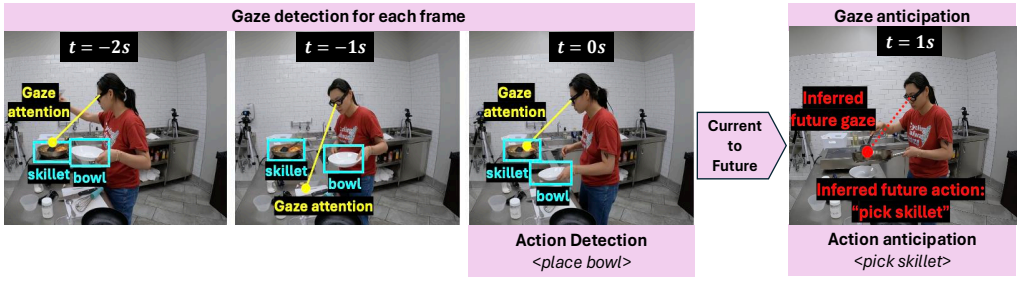


Figure 4: **Correlation of action–gaze over time in Exo-Cook.** Gaze reveals both the current focus of attention and future intention, providing complementary cues for recognizing ongoing actions and anticipating subsequent human-object interactions.

During feature encoding, the key difference between FPV and TPV arises in the construction of human–object interaction tokens. For each frame t, we extract human and object bounding boxes and construct human–object pair tokens

$$\mathbf{x}_t^o = [\mathbf{x}_{t,\langle 1,1\rangle}, \ldots, \mathbf{x}_{t,\langle i,j\rangle}, \ldots, \mathbf{x}_{t,\langle n_t^h, n_t^o\rangle}],$$

where n_t^h and n_t^o denote the numbers of detected humans and objects. Each token $\mathbf{x}_{t,\langle i,j\rangle}$ encodes the interaction between human i and object j, including appearance features and spatial relationships. The representation of $\mathbf{x}_{t,\langle i,j\rangle}$ differs slightly between viewpoints. In TPV videos, multiple subjects may appear in the scene ($i \geq 1$), and the token encodes scene appearance, human body features, body location, and the spatial relationship between humans and objects. These cues are important for recognizing interactions from a third-person perspective. In contrast, FPV videos lack explicit global body location because the camera is attached to the actor. Instead, interaction tokens encode the spatial relationships between detected hands and objects, with the human index fixed to $n_t^h = 1$. Another difference arises in gaze estimation. In TPV videos, SAGE predicts gaze using head appearance and orientation features following prior work such as (Chong et al., 2020). In FPV videos, gaze is interpreted as an attention region within the egocentric scene, and the gaze model predicts a spatial attention distribution from scene features. Despite these viewpoint-specific adaptations, the subsequent gaze-conditioned reasoning modules (GCSA and GCTP) remain identical across FPV and TPV, enabling a unified framework for joint gaze–action modeling.

A.2.3 Model Architecture.

SAGE jointly performs four tasks and the detailed SAGE architecture can be divided into four parts:

SAGE-1 for Gaze Detection. We integrate existing gaze detection models into SAGE framework to produce differentiable gaze heatmaps. The gaze detection module comprises a visual encoder and a decoder that outputs gaze fixation heatmaps. The input sequence is divided into non-overlapping patches and each patch is flattened and projected into a D-dimensional embedding space via a linear mapping. These video tokens are then processed by transformer layers composed of multiple self-attention blocks. To generate gaze heatmaps, a transformer decoder—based on multiscale self-attention blocks from MViT (Fan et al., 2021)—upsamples the encoded features. The decoder outputs feature maps of shape $T'H'W' \times D'$, followed by a softmax operation over the last dimension to obtain the final gaze heatmap $\boldsymbol{g}_{0:t}$. For the egocentric setting, we adopt the GLC (Lai et al., 2024a) model as our gaze detection architecture. For exocentric videos, we use the architecture proposed by (Chong et al., 2020), initializing our model with their pretrained weights.

SAGE-2 for HOI Detection. Let $\mathbf{X}_t \in \mathbb{R}^{(N+1)\times D}$ denote the input tokens at frame t, consisting of N human–object pair tokens and one global token, where D is the model dimension. The spatial encoder contains N_d Transformer layers. Let $\mathbf{H}_t^{(0)} = \mathbf{X}_t$. For layer $l = 1, \ldots, N_d$, the projections are

$$Q_t^l = \mathbf{H}_t^{(l-1)} W_Q^l, \quad K_t^l = \mathbf{H}_t^{(l-1)} W_K^l, \quad V_t^l = \mathbf{H}_t^{(l-1)} W_V^l, \tag{7}$$

where $W_Q^l, W_K^l, W_V^l \in \mathbb{R}^{D\times D}$. For H attention heads, each head has dimension $d_k = D/H$. The gaze bias matrix is $\mathbf{S}_t$. Each Transformer layer updates the token representations with the GCSA operation:

$$\mathrm{GCSA}(Q_t^l, K_t^l, V_t^l, \mathbf{S}_t) = \mathrm{Softmax}\left(\frac{Q_t^l (K_t^l)^\top + \mathbf{S}_t}{\sqrt{d_k}}\right) V_t^l. \tag{8}$$

$$\widetilde{\mathbf{H}}_t^l = \mathrm{LN}\Big(\mathbf{H}_t^{(l-1)} + \mathrm{GCSA}(Q_t^l, K_t^l, V_t^l, \mathbf{S}_t)\Big), \tag{9}$$

$$\mathbf{H}_t^l = \mathrm{LN}\Big(\widetilde{\mathbf{H}}_t^l + \mathrm{FFN}^l(\widetilde{\mathbf{H}}_t^l)\Big), \tag{10}$$

where $\mathrm{FFN}^l : \mathbb{R}^D \to \mathbb{R}^{D_{\mathrm{ff}}} \to \mathbb{R}^D$ is a feed-forward network. After N_d layers, the spatial encoder outputs $\mathbf{C}_t = \mathbf{H}_t^{(N_d)} \in \mathbb{R}^{(N+1)\times D}$. The frame-level gaze-aware HOI feature is obtained from the global token $\mathbf{c}_t \in \mathbb{R}^D$. Given the sequence of frame features $[\mathbf{c}_0, \ldots, \mathbf{c}_t] \in \mathbb{R}^{(t+1)\times D}$, we apply a temporal Transformer encoder with one layer

$$\mathbf{z}_{0:t} = \mathrm{TransformerEnc}_1([\mathbf{c}_0, \ldots, \mathbf{c}_t]) \in \mathbb{R}^{(t+1)\times D}. \tag{11}$$

Let $\mathbf{z}_t \in \mathbb{R}^D$ denote the final token corresponding to time t. The HOI detection head is an MLP $\phi : \mathbb{R}^D \to \mathbb{R}^{C_y}$, where C_y is the number of HOI classes. The action probability is predicted as

$$p(y_t \mid g_{0:t}, I_{0:t}) = \mathrm{Softmax}(\phi(\mathbf{z}_t)) . \tag{12}$$

SAGE-3 for Gaze Anticipation. Each gaze heatmap $g_k \in \mathbb{R}^{64\times 64}$ is embedded into a latent token with periodic position encoding (PPE) (Fan et al., 2022) : $\tilde{\mathbf{g}}_k = \mathrm{Conv}(g_k) + \mathrm{PPE}(k)$, where $\tilde{\mathbf{g}}_k \in \mathbb{R}^D$, D is the model dimension and $PPE(\cdot)$ is denoted as:

$$\begin{aligned} PPE_{(t,2i)} &= \sin((t \bmod P)/(10000)^{2t/d}) \\ PPE_{(t,2i+1)} &= \cos((t \bmod P)/(10000)^{2t/d}) \end{aligned} \tag{13}$$

Stacking all tokens yields the gaze sequence $[\tilde{\mathbf{g}}_0, \ldots, \tilde{\mathbf{g}}_t] \in \mathbb{R}^{(t+1)\times D}$. We use a Transformer encoder with N_g layers to model the temporal dependencies of gaze trajectories:

$$\mathbf{h}^g_{0:t} = \mathrm{TransformerEnc}_{N_g}([\tilde{\mathbf{g}}_0, \ldots, \tilde{\mathbf{g}}_t]) \in \mathbb{R}^{(t+1)\times D}. \tag{14}$$

Let $\mathbf{X}^o_{0:t} \in \mathbb{R}^{(t+1)\times N\times D}$ denote the human–object interaction tokens, $\mathbf{H}^g = \{\mathbf{h}^g_{0:t}\}$. A Cross-attention layer is applied to integrate video interaction context with gaze features:

$$\mathbf{z}^g_{0:t} = \mathrm{MHCA}(\mathbf{H}^g, \mathbf{X}^o_{0:t}) \in \mathbb{R}^{(t+1)\times D}. \tag{15}$$

The future gaze heatmaps are generated by a decoder

$$\hat{g}_m = \mathcal{D}_\phi(\mathbf{z}^g_t), \quad m = t+1, \ldots, t+M, \tag{16}$$

SAGE-4 for HOI Anticipation. Let $\hat{\mathbf{G}} = \{\hat{\mathbf{g}}_{t+1:t+M}\} \in \mathbb{R}^{M\times D}$ denote the embedded future gaze tokens, where M is the prediction horizon and D is the model dimension. Temporal correlations among future gaze shifts are modeled by a MHSA layer. For the l-th layer, the projections are

$$Q^l = \hat{\mathbf{G}}^{(l-1)} W^l_Q, \quad K^l = \hat{\mathbf{G}}^{(l-1)} W^l_K, \quad V^l = \hat{\mathbf{G}}^{(l-1)} W^l_V, \tag{17}$$

where $W^l_Q, W^l_K, W^l_V \in \mathbb{R}^{D\times D}$. For H attention heads, each head has dimension $d_k = D/H$. The self-attention output is

$$\mathrm{MHSA}(\hat{\mathbf{G}}^{(l-1)}) = \mathrm{Softmax}\left(\frac{Q^l (K^l)^\top}{\sqrt{d_k}}\right) V^l. \tag{18}$$

The historical temporal gaze representation is $\mathbf{A}^l = \mathrm{LN}\Big(\hat{\mathbf{G}}^{(l-1)} + \mathrm{MHSA}(\hat{\mathbf{G}}^{(l-1)})\Big)$. To condition future HOI prediction on the current interaction context, we apply cross-attention between the gaze tokens and the current gaze-conditioned HOI feature $\mathbf{c}_t \in \mathbb{R}^D$, with $Q_c = \mathbf{A}^l W^c_Q, \quad K_c = \mathbf{c}_t W^c_K, \quad V_c = \mathbf{c}_t W^c_V$, The MHCA is calculated as:

$$\mathrm{MHCA}(\mathbf{A}^l, \mathbf{c}_t) = \mathrm{Softmax}\left(\frac{Q_c K_c^\top}{\sqrt{d_k}}\right) V_c; \ \hat{\mathbf{A}}^l = \mathrm{LN}\big(\mathbf{A}^l + \mathrm{MHCA}(\mathbf{A}^l, \mathbf{c}_t)\big) . \tag{19}$$

In conclusion, the gaze-conditioned temporal representation (matrix $\hat{\mathbf{a}}_{t+1:t+M}$ denoted as $\hat{\mathbf{A}}$) for future HOI feature is generated from

$$\hat{\mathbf{A}} = \mathrm{TransformerEnc}_{N_a}(\hat{\mathbf{G}}, \mathbf{c}_t) \in \mathbb{R}^{M\times D}. \tag{20}$$

and the anticipated HOI at step $t+M$ is predicted by

$$\hat{y}_{t+M} = \mathrm{Softmax}(\mathrm{MLP}(\hat{\mathbf{a}}_{t+M})). \tag{21}$$

A.2.4 Monte Carlo Gaze Sampling.

To propagate gaze uncertainty into both current and future action prediction, we approximate the marginal likelihood in Eq.1 (main paper) using Monte Carlo sampling over gaze trajectories. The gaze detector predicts a heatmap for each observed frame independently. Let $\hat{g}_\tau \in \mathbb{R}^{H_g\times W_g}$ denote

the predicted gaze heatmap at time τ, where $\tau = 0, \ldots, t$. We first normalize each heatmap into a spatial distribution

$$p_\tau(h, w) = \frac{\hat{g}_\tau(h, w)}{\sum_{h', w'} \hat{g}_\tau(h', w')}. \tag{22}$$

Assuming conditional independence across observed frames, the present gaze distribution is

$$p(\boldsymbol{g}_{0:t} \mid I_{0:t}) = \prod_{\tau=0}^{t} p(g_\tau \mid I_\tau). \tag{23}$$

For each Monte Carlo sample $k \in \{1, \ldots, K\}$, we independently sample one gaze fixation location from each frame-wise distribution, The sampled present gaze trajectory is then

$$\boldsymbol{g}^k_{0:t} = \big(g^k_0, g^k_1, \ldots, g^k_t\big). \tag{24}$$

Conditioned on the sampled present gaze trajectory $\boldsymbol{g}^k_{0:t}$ and observed video features, the gaze anticipation module predicts future gaze heatmaps

$$\hat{\boldsymbol{g}}^k_{t+1:t+M} \sim p(\boldsymbol{g}_{t+1:t+M} \mid \boldsymbol{g}^k_{0:t}, I_{0:t}). \tag{25}$$

For each future step $t+m$, we again normalize the predicted heatmap into a 2D gaussian distribution, sample one gaze location. Combining Eq 24 and Eq 25 yields a full sampled gaze trajectory

$$G^k = \big(\boldsymbol{g}^k_{0:t}, \boldsymbol{g}^k_{t+1:t+M}\big). \tag{26}$$

For each sampled trajectory G^k, the current and future action probabilities are evaluated as

$$f(G^k) = p(y_t \mid \boldsymbol{g}^k_{0:t}, I_{0:t})\; p(\boldsymbol{y}_{t+1:t+M} \mid y_t, \boldsymbol{g}^k_{t+1:t+M}, I_{0:t}). \tag{27}$$

The marginal likelihood is approximated by

$$\widehat{p}(y_{t:t+M} \mid I_{0:t}) = \frac{1}{K} \sum_{k=1}^{K} f(G^k). \tag{28}$$

During training, since the gaze detection model is loaded from a pre-trained model, we use a small number of samples ($K = 1$) for efficiency. During inference, we can sample multiple gaze trajectories ($K \geq 2$)for analyzing action uncertainty conditioning on gaze.

A.2.5 Training Details.

Based on the complexity of the full SAGE model and the difficulty in training for all the tasks. We initialize some models with pre-trained model and multi-stage training process.

- The Gaze detection module is initialized with the pre-trained model from VideoAttn (Chong et al., 2020) (on Vid-HOI dataset) or GLC (Lai et al., 2024a).
- We first train joint model SAGE-12 for 5 epochs with gaze labels and action labels for the current sequence.
- Then we load the weights from SAGE-12 and train the full model SAGE with future gaze and action annotations. We train 25 epochs in total.
- In the loss function $\mathcal{L} = \lambda_1 \mathcal{L}_{hm,1} + \lambda_2 \mathcal{L}_{hm,2} + \lambda_3 \mathcal{L}_{io} + \lambda_4 \mathcal{L}_{act,1} + \lambda_5 \mathcal{L}_{act,2}$, we set $\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 0.5, \lambda_4 = 1.5, \lambda_5 = 1.5$. Note that $\lambda_3 = 0$ in egocentric case. These values were determined empirically through initial tuning experiments and remained fixed for all reported results.

A.3 Exo-Cook Benchmark Creation

We construct a benchmark based on a subset from Ego-Exo 4D. We describe more details for preprocessing and generating the labels in this section. On cooking videos, we try BERT model (Devlin et al., 2019) to generate sentence embedding and cluster text into action categories. We extract the text descriptions from Ego-Exo 4D cooking videos and generate 30250 video clips and the corresponding action descriptions. We generate text embeddings for each video clip and apply K-means algorithm to perform action clustering based on the semantic meaning of the text.



Figure 5: Visualization of the best exocentric views (based on Ego-Exo4D annotations) at different timestamps in one cooking scenario with four camera views. We show that at two different stamps (183.30927s and 432.88395s), the best exocentric views (highlighted in red) come from Cam02 and Cam03 respectively. We generate Exo-Cook clips based on the best exocentric views.

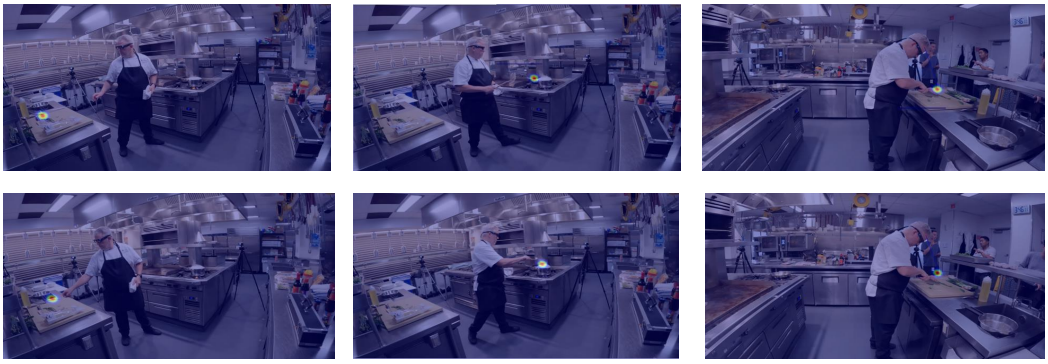


Figure 6: Individual examples (six) of created ground-truth gaze heatmap from Exo-Cook.

A.3.1 View Diversity of Exo-Cook.

The original Ego-Exo4D (Grauman et al., 2024) provide multiple (4-5) exocentric videos from tripods placed around the subject. All views are time-synchronized. Exo-Cook is sampled from Ego-Exo4D cooking videos based on the atomic action descriptions that specify what occurs at a single time point. The timestamp in the description is linked to a `best_exo` field, which denotes the exocentric camera view that provides the clearest visual evidence of the event. Exo-Cook clips are sampled from the best exocentric capturing using the atomic action timestamp. In all, Exo-Cook includes 8,778, 8,721, 6,429, 6,253, and 1,869 clips from Cam01–Cam05 respectively. We show an example of Exo-Cook clips from different view points in Figure 5. For each cooking scene, we sample 20 frames at different timestamps from the best exocentric view with valid action annotations, which will be used as the input sequence for SAGE. Each Exo-Cook clip contains a input sequence and future segments. For the anticipation task, we sample 4 frames within future 1 second. We generate annotations for each clip, including gaze heatmap, human and object bounding boxes and action labels. We introduce the annotation pipeline in the following sections.

A.3.2 Gaze Label Creation.

Ego-Exo 4D provide raw gaze annotations from the eye-tracking device (Project Aria). To generate 2D gaze labels for third-person views in Exo-Cook, we follow a calibrated projection pipeline that transforms the 3D gaze direction from the eye-tracking device (Project Aria) into 2D image coordinates of external cameras.

Gaze Ray Extraction. For each frame, we extract the 3D gaze origin and direction vectors from the left and right eyes using Aria's 'scene gaze' and 'eye gaze' streams. These are defined in the Aria scene camera coordinate system and describe binocular gaze rays:

$$\text{Ray}_L = (\boldsymbol{o}_L, \boldsymbol{d}_L), \quad \text{Ray}_R = (\boldsymbol{o}_R, \boldsymbol{d}_R)$$

3D Gaze Estimation in Aria Frame. Each frame of Project Aria includes left and right eye gaze vectors with respect to the device's coordinate frame. We estimate the 3D point of gaze $\boldsymbol{G}_{3D}^{\text{aria}}$ by computing the 3D intersection (or midpoint approximation) of the two gaze rays Ray_L and Ray_R.

Coordinate Transformation. To localize the gaze in the third-person camera frame, we use the known extrinsic calibration between the Aria device and each external camera. Let $\boldsymbol{T}_{\text{aria}\to\text{cam}}$ denote the 6DoF pose of the Aria glasses with respect to a third-person camera. We transform the 3D gaze point into the third-person coordinate system via:

$$\boldsymbol{G}_{3D}^{\text{cam}} = \boldsymbol{T}_{\text{aria}\to\text{cam}} \cdot \boldsymbol{G}_{3D}^{\text{aria}}$$

3D-to-2D Projection. Given the intrinsic camera matrix $\boldsymbol{K}$ of the third-person camera and the transformed gaze point $\boldsymbol{G}_{3D}^{\text{cam}}$, we project the gaze to the image plane:

$$\boldsymbol{G}_{2D}^{\text{cam}} = \Pi(\boldsymbol{K}, \boldsymbol{G}_{3D}^{\text{cam}})$$

where Π denotes the standard perspective projection. This gives us the 2D gaze fixation point in pixel coordinates.

Heatmap Generation. Following prior work (Chong et al., 2020), we convert each gaze fixation point into a 2D Gaussian heatmap $\boldsymbol{M}^{\text{pseudo}} \in \mathbb{R}^{H\times W}$. The heatmap is centered at $\boldsymbol{G}_{2D}^{\text{cam}}$ with an isotropic Gaussian kernel:

$$\sigma = \frac{W_{\text{hm}} + H_{\text{hm}}}{2} \cdot \frac{3}{64}$$

where W_{hm} and H_{hm} are the dimensions of the heatmap. For a 64×64 heatmap, this yields $\sigma = 3$ pixels. We show examples of generated gaze heatmaps in Figure. 6.

To handle occlusions or failed triangulation cases, we discard gaze points where the intersection error between eye rays exceeds a threshold. Additionally, all transformations and projections are timestamp-aligned to ensure synchronization between Aria and third-person views.

Table 6: Verb-object extraction and semantic object selection using spaCy and BERT. In the third column, we show the cosine similarity between the spaCy object and the object name mentioned in Exo-Cook raw annotation.

Sentence	Verb Object	spaCy Object List	Object Name
C places the spoon in the bowl in his right hand.	places, spoon	spoon (0.652), bowl (1.000), hand (0.437)	Bowl
C slices the garlic on the chopping board with his right hand.	slices, garlic	garlic (0.424), board (0.869), hand (0.341)	Chopping board
C adds the sliced garlic into the frying pan with his right hand.	adds, garlic	garlic (0.461), pan (1.000), hand (0.378)	Pan
C adds oil into the frying pan with the oil bottle in his left hand.	adds, oil	oil (0.552), pan (0.449), bottle (1.000), hand (0.437)	Bottle

A.3.3 Action Label Creation

Action Description Alignment. Each atomic description is associated with the following key metadata:

- `start_frame` and `end_frame`: specifying the frame-level temporal boundaries of the described action.
- `source_view`: indicating which third-person or egocentric camera the annotation corresponds to.
- `text`: raw textual description of the action.

To align each action with global time stamp, we use the provided frame indices and convert them to time given the video frame rate. We first sample the video frames near the `start_frame` for action recognition. Then we define a time offset $\tau_a = 1s$ beyond the `end_frame` to define the target for action anticipation.

Action Simplification with spaCy. We use spaCy (Explosion, 2024) to perform syntactic parsing and extract verb-object pairs from each sentence. For example, given the description *"C slices the garlic on the chopping board with his right hand"*, spaCy identifies `slices` as the verb and `garlic` as the primary object. In more complex sentences containing multiple noun phrases (e.g., "C adds the sliced garlic into the frying pan with his right hand"), spaCy is used to list all candidate objects (e.g., `garlic`, `pan`, `hand`). To identify the most semantically relevant object, we compute the cosine similarity between the BERT (Devlin et al., 2019) embedding of each (verb,noun) pair and the embedding of the full sentence. The object with the highest similarity is selected as the primary target. For instance, in the garlic example, `pan` receives the highest score, thus being identified as the object most aligned with the described action context. In Table. 6, we show the Cosine Similarity between different sentences extracted from the annotation file of Ego-Exo 4D.

Action Clustering. As described above, we repeat the action description alignment and simplification process for all the samples we cropped from Exo-Cook. After applying spaCy for extracting essential semantic components for each clip, we employ the pre-trained BERT (Devlin et al., 2019) model to transform these extracted components into high-dimensional text embeddings. Once text embeddings for all sequences are obtained, we apply K-means clustering to group similar embeddings. To determine the optimal number of clusters, we use the Elbow Method by analyzing the inertia, defined as the sum of squared distances between each data point and the centroid of its assigned cluster. The inertia values across different choices of K are illustrated in Figure 8. Based on this analysis, we set the number of clusters to $K = 10$ and therefore, we have 10 categories of actions in Exo-Cook. In Figure 7, we present two representative examples from every category, accompanied by their respective (verb, noun) pairs, illustrating how each cluster category captures distinct semantic interpretations.

To further evaluate the semantic consistency of the resulting categories, we embed each action label (verb–noun pair) using a pretrained Sentence-BERT model (all-mpnet-base-v2) and compute pairwise cosine similarities. The average intra-category similarity (0.61) is substantially higher than the inter-category similarity (0.22), yielding a separation gap of 0.39 (2.77×) with statistical significance ($p < 10^{-4}$, permutation test). Consistently, t-SNE visualizations and qualitative examples (Figure 7) show well-separated, interpretable clusters, which are also manually verified. To evaluate reliability, we conduct a manual study on 1000 randomly sampled narrations (98% with multiple nouns), comparing manually selected (verb, noun) pairs with those from our pipeline. The BERT-based object selection achieves 80.6% accuracy, with high inter-annotator agreement (0.91) between two independent annotators. This accuracy is partly limited by cases where the correct noun lacks an object mask in Ego-Exo4D, and would be much higher otherwise.

A.3.4 Bounding Box for Human and Object.

We follow the pipeline in (Ni et al., 2023) and use YOLOv5 (Jocher et al., 2022) to detect full-body and head bounding boxes of the camera wearer in third-person views.

Bounding Box Detection. The first row of Figure 9 illustrates examples of head bounding box detections (highlighted in green) within Exo-Cook images, capturing every person appeared in the scene. Similarly, the second row presents object bounding box detections, marked in yellow, covering all relevant objects within the scene.

Bounding Box Matching. As our primary interest is the activity of the "major participant" equipped with the Project Aria device, we implement bounding box matching based on the projected location of the device. Specifically, the head bounding box containing the 2D projection of the Project Aria's center is designated as the major participant. Consequently, as depicted in the second row of Figure 9, any unrelated head bounding boxes will be removed.

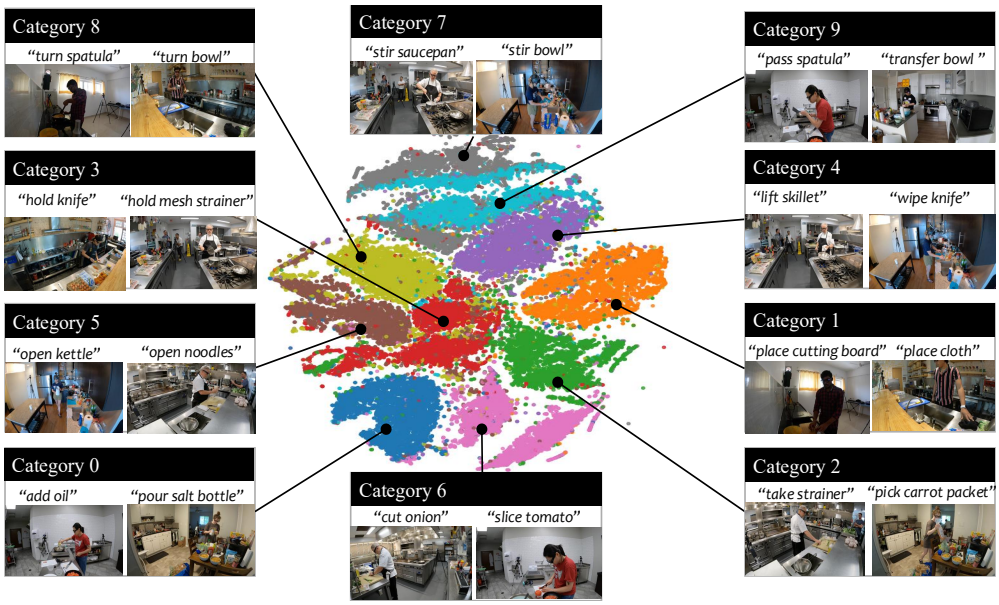


Figure 7: We illustrate the semantic interpretations of the clusters in the Exo-Cook dataset through some representative samples across 10 action categories. The clusters are obtained using K-Means on the (verb, noun) embeddings.

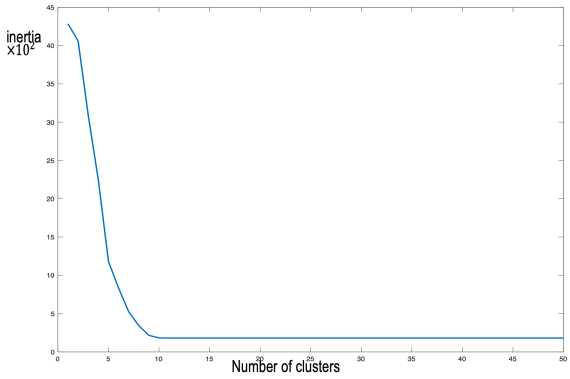


Figure 8: The inertia plot from the Elbow Method applied to all text embeddings from cooking video clips indicates the selection of $K = 10$ clusters.

A.4 Baseline Model for Gaze Anticipation

A.4.1 Exocentric View.

Due to lack of baseline methods on Gaze Anticipation task, we propose to adapt existing models as baseline. In Table 1 of the main paper, we evaluate gaze anticipation performance on the Exo-Cook dataset by adapting two baseline models: VideoAttn (Chong et al., 2020) and Sharingan (Tafasca et al., 2024). These baseline models, initially designed for gaze detection tasks, were modified and retrained to predict gaze heatmaps at a future timestamp (1 second ahead) based solely on the last frame of a given sequence. The results show the advantage of employing the temporal modeling in our SAGE model. Specifically, our temporal-aware SAGE significantly outperforms the adapted baseline models across all gaze anticipation metrics. For instance, the SAGE model (with Sharingan as its gaze detection backbone) achieves an F1 score of 49.2, notably higher than adapted VideoAttn

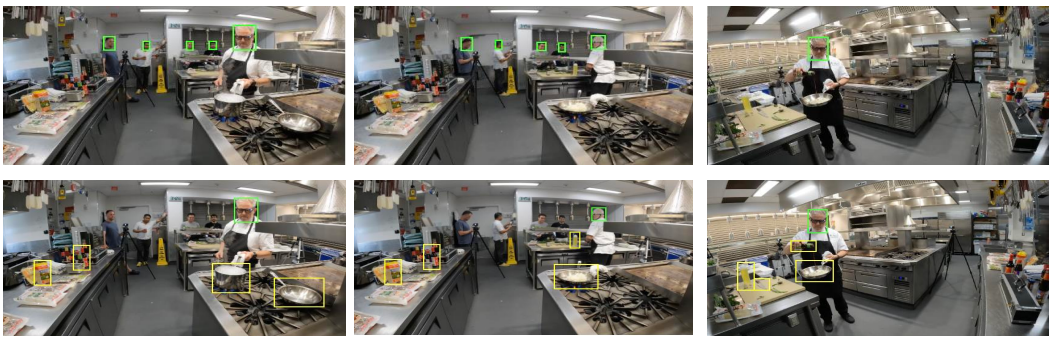


Figure 9: Examples (three) of bounding box detection for human head and object in Exo-Cook images. The first row represents head bounding box (green color) detection for all subjects in the scene. The second row shows the head bounding box for the key participant and the objct bounding boxes (yellow color).

(42.2) and adapted Sharingan (40.1). Similarly, recall (62.8 vs. 54.6 and 52.7) and precision (50.0 vs. 42.5 and 40.4) metrics also show substantial improvements.

A.4.2 Egocentric View.

Similarly, we show comparison results in Table 4 of the main paper. CSTS (Lai et al., 2024b) is the state-of-the art model for egocentric gaze anticipation. However, CSTS (Lai et al., 2024b) is not trained/evaluated on EGTEA Gaze+ as they require audio data. We adapt their backbone model, denoted as "CSTS-Visual (Lai et al., 2024b)" in Table. 4 to gaze anticipation task on EGTEA Gaze+. In addition to CSTS-Visual, we adapt another three existing models for gaze anticipation, denoted as GLC† (Lai et al., 2024a), MViT† (Fan et al., 2021) and I3D-R50† (Feichtenhofer et al., 2019). These models were retrained to predict gaze heatmaps at 0.5 seconds into the future. We re-train the four models by using the grounf-truth heatmaps at the 0.5 seconds later. However, their performance was substantially lower compared to SAGE.

A.5 Extensive Experimental Analysis

A.5.1 Model Efficiency.

Although efficiency is not our focus, SAGE achieves strong recognition and anticipation accuracy while maintaining high computational efficiency, processing a 16-frame input sequence (on EGTEA Gaze+) in just 0.63 s on an NVIDIA RTX A6000. The full SAGE (egocentric) model contains 333.2M parameters, representing only a modest 3% increase compared to the combined next-best task-specific baselines (GLC, GC-TSM, GLC-adapted, and InAViT), despite jointly addressing all four tasks within a single architecture. In addition, SAGE exhibits comparable inference efficiency, with a slightly faster runtime (by 2%) than independently executing the four task-specific models. We also compare the computational complexity for one round of iterative inference of SAGE vs other SOTA methods in terms of GFLOPS (Giga Floating-Point Operations Per Second) in Table 7. As shown, the proposed model achieves higher accuracy while reducing computational complexity by 11.3% compared to the *combined* next-best task-specific baselines. This demonstrates an improved accuracy–compute tradeoff, delivering better predictive performance at lower computational cost, enabling enhanced suitability for real-world deployment under compute and power constraints.

A.5.2 Sensitivity of SAGE to Gaze Model.

In the main paper, we compare the performance of SAGE with state-of-the-art models across different task domains. In the exocentric view, the effectiveness of gaze is more apparent, as the gaze model can explicitly identify the visual target of the person within the scene. Intuitively, the gaze-based attention closely aligns with the object being interacted with, making it a strong prior for modeling human-object interactions. In Table 1 of the main paper, we show the sensitivity of SAGE performance to the accuracy of gaze model. We incorporate two different gaze backbone model, de-

Table 7: Comparison of computational cost (GFLOPs). SAGE achieves an 11.3% reduction in total GFLOPs compared to the combined cost of prior SOTA models.

Model	GFLOPS	
	Individual	Total
GLC (Lai et al., 2024a)	72.2	601.5
GC-TSM (Hao et al., 2022)	66.5	
GLC (Lai et al., 2024a)	70.8	
InAViT (Roy et al., 2024)	392	
SAGE (Ours)	533.4	533.4 ↓11.3%

Table 8: Sensitivity of SAGE to gaze modules on the EGTEA Gaze+ dataset

Model	Gaze Model	Gaze Detection			Action Recogn.	Gaze Anticipation			Action Anticipation
		F1	Rec	Prec	Top-1 Acc	F1	Rec	Prec	Mean-cls Acc
SAGE	I3D-R50 Feichtenhofer et al. (2019)	42.2	58.5	33.2	64.6	34.7	47.4	30.2	57.8
	MViT Fan et al. (2021)	44.5	58.8	35.5	64.6	**38.8**	54.2	32.5	58.0
	GLC Lai et al. (2024a)	**46.8**	**62.1**	**36.8**	**65.0**	37.0	**54.9**	**32.7**	**58.4**

noted as SAGE (Gaze Detection Model: VideoAttn) and SAGE (Gaze Detection Model: Sharingan) in Table 1, into SAGE architecture. The results show that better gaze detection model in SAGE enhances performances for other three tasks.

Similarly, to explore the sensitivity of SAGE to the quality of gaze estimation in egocentric videos, we compare three different gaze backbones in Table 8: I3D-R50 (Feichtenhofer et al., 2019), MViT (Fan et al., 2021), and GLC (Lai et al., 2024a). We observe that improvements in gaze detection performance consistently lead to better action recognition and anticipation results. For instance, GLC achieves the best gaze detection F1 score (46.8), which corresponds to the highest top-1 accuracy in action recognition (65.0) and action anticipation (58.4). In contrast, models with weaker gaze detection performance, such as I3D-R50 (F1 = 42.2), yield the lowest accuracy in both action recognition and anticipation. These results suggest that precise gaze localization is crucial for enhancing action detection and anticipation in egocentric settings. Intuitively, gaze reflects the subject's focus of interest and serves as an informative prior for modeling human-object interactions. The results in Table 8 prove that better gaze attention enhances the performance for each task. It validates the effectiveness of our GCSA module.

A.5.3 Sensitivity of SAGE to Anticipation Horizon.

Following InAViT (Roy et al., 2024), we explore SAGE performance on longer time horizons. In Table 9, we train SAGE on different time horizons {0.5s, 1s, 1.5s, 2s}. Notably, SAGE consistently outperforms InAViT on all time horizons. As the anticipation horizon increases, there are greater uncertainty for future action prediction. The Top-1 accuracy of action anticipation in InAViT decreases from 67.8 to 64.1 when the anticipation time increases to 2.0s. As SAGE is anticipating future actions with multiple intermediate time stamps (i.e., $[y_{t+1}, \cdots, y_{t+M}]$), SAGE maintains more stable performance over longer time spans.

A.5.4 Extensive Results On Vid-HOI.

In Table 10, we benchmark HOI detection performance on the Vid-HOI dataset using ST-Gaze (Ni et al., 2023) and STTran (Cong et al., 2021) as baselines. Since Vid-HOI does not provide gaze

Table 9: Top-1 accuracy for SAGE at longer anticipation horizon on the EGTEA Gaze+ dataset.

Method	τ_a			
	0.5	1.0	1.5	2.0
InAViT Roy et al. (2024)	67.8	66.9	65.8	64.1
SAGE (Ours)	**68.0**	**67.3**	**66.2**	**64.5**

Table 10: HOI detection in Oracle mode on Vid-HOI dataset Chiou et al. (2021). * includes word embedding module.

Method	τ_a	mAP	Rec	Prec	Acc	F1
STTran Cong et al. (2021)	-	28.32	-	-	-	-
ST-Gaze* Ni et al. (2023)	-	38.46	**73.62**	59.16	**53.76**	60.57
ST-Gaze Spatial* Ni et al. (2023)	-	36.29	71.03	59.38	51.72	61.24
SAGE-12 (Ours)	-	38.08	72.11	59.90	52.12	**62.03**
SAGE (Ours)	1	**38.65**	72.44	59.22	52.22	61.96
	3	38.18	72.21	**59.95**	52.18	62.02
	5	38.13	72.01	59.92	52.18	61.88

labels, gaze evaluation is omitted. To specifically evaluate the impact of Gaze-Conditioned Attention in the spatial domain, we introduce a spatial-only variant of ST-Gaze by removing its temporal layer (referred to as ST-Gaze Spatial). Additionally, we include SAGE-12 in the comparison. Notably, SAGE-12 achieves the highest F1 score (62.03), underscoring the importance of integrating gaze-based attention mechanisms into HOI recognition tasks. Furthermore, our full SAGE model attains the best mAP (38.65) and Recall (72.44) scores at the 1-second anticipation setting, and the highest Precision (59.95) at the 3-second setting.

A.5.5 SAGE Performance Variance

To evaluate the stability of SAGE, we repeat the experiments using three random seeds and report the mean and standard deviation of the evaluation metrics. Tables 11 and 12 summarize the results on EGTEA Gaze+ and Vid-HOI, respectively. Across gaze detection, action detection, gaze anticipation, and action anticipation, SAGE exhibits consistently small variation across runs. On EGTEA Gaze+, the standard deviations range from 0.09 to 0.35 across all reported metrics. Similarly, on Vid-HOI, the standard deviations are at most 0.51 for HOI detection and 0.24 for HOI anticipation. These results indicate that the performance of SAGE is stable across different random initializations. Following prior methods, the main paper reports the best run for direct comparison, while the results below provide the corresponding cross-seed variability.

Table 11: Performance variance of SAGE on EGTEA Gaze+ over three random seeds. We report mean $\pm$ standard deviation together with the result reported in the main paper.

Model	**Gaze Det.**			**Action Det.**	**Gaze Ant.**			**Action Ant.**
	F1	Rec.	Prec.	Top-1 Acc.	F1	Rec.	Prec.	Mean-cls Acc.
SAGE (reported)	46.8	62.1	36.8	65.0	37.0	54.9	32.7	58.4
SAGE (3 seeds)	46.7 ± 0.09	61.8 ± 0.24	36.6 ± 0.14	64.8 ± 0.11	36.6 ± 0.26	54.8 ± 0.35	32.5 ± 0.12	58.1 ± 0.23

Table 12: Performance variance of SAGE on Vid-HOI over three random seeds for HOI detection and anticipation ($\tau_a = 1$).

Setting	mAP	Rec.	Prec.	Acc.	F1
Detection (reported)	38.65	72.44	59.22	52.22	61.96
Detection (3 seeds)	37.22 ± 0.42	72.01 ± 0.39	59.26 ± 0.45	51.65 ± 0.28	62.06 ± 0.51
Anticipation (reported)	37.71	72.88	60.24	51.86	62.72
Anticipation (3 seeds)	37.59 ± 0.24	72.84 ± 0.12	59.98 ± 0.12	51.62 ± 0.24	62.54 ± 0.23

A.5.6 SAGE Ablation Study.

In the main paper, we show ablation study of model combination on an egocentric dataset, i.e. EGTEA Gaze+ dataset. In this section, we provide a more comprehensive analysis of different component of SAGE. We first introduce different combination of SAGE modules. In Section 4 of

Table 13: Empirical validation of the conditional independence assumption between past gaze and future actions.

Models	Gaze Detection			Action Recogn.	Gaze Anticipation			Action Anticipation
	F1	Rec	Prec	Acc	F1	Rec	Prec	Acc
SAGE* ($\boldsymbol{y}_{t+1:t+M}\|\boldsymbol{g}_{0:t+M}, y_t$)	46.6	62.0	36.5	65.0	36.3	54.2	32.3	57.7
SAGE ($\boldsymbol{y}_{t+1:t+M}\|\boldsymbol{g}_{t+1:t+M}, y_t$)	**46.8**	**62.1**	**36.8**	65.0	**37.0**	**54.9**	**32.7**	**58.4**

the main paper, we decompose the problem of joint estimation of current and future action as:

$$\begin{aligned} p(\boldsymbol{y}_{t:t+M} \mid I_{0:t}) &= \int_G p(\boldsymbol{y}_{t:t+M}, \boldsymbol{g}_{0:t+M} \mid I_{0:t})\, dG \\ &= \int_G p(\boldsymbol{y}_{t:t+M} \mid \boldsymbol{g}_{0:t+M}, I_{0:t})\, p(\boldsymbol{g}_{0:t+M} \mid I_{0:t})\, dG \\ &= \int_G p(y_t \mid \boldsymbol{g}_{0:t}, I_{0:t})\, p(\boldsymbol{y}_{t+1:t+M} \mid y_t, \boldsymbol{g}_{t+1:t+M}, I_{0:t}) \\ &\quad \cdot\, p(\boldsymbol{g}_{0:t} \mid I_{0:t})\, p(\boldsymbol{g}_{t+1:t+M} \mid \boldsymbol{g}_{0:t}, I_{0:t})\, dG \end{aligned} \tag{29}$$

Causal Assumption. In Section 4 of the main paper, we adopt a simplifying causal assumption that future actions $\boldsymbol{y}_{t+1:t+M}$ are conditionally independent of past gaze $\boldsymbol{g}_{0:t}$ given the predicted future gaze $\boldsymbol{g}_{t+1:t+M}$. This leads to the approximation

$$p(\boldsymbol{y}_{t+1:t+M} \mid y_t, \boldsymbol{g}_{0:t+M}, I_{0:t}) \approx p(\boldsymbol{y}_{t+1:t+M} \mid y_t, \boldsymbol{g}_{t+1:t+M}, I_{0:t}).$$

The assumption is introduced to prevent an unnecessarily larger model architecture, as it removes the need to explicitly propagate the full historical gaze sequence into the action anticipation module. Intuitively, once the current action representation and the predicted future gaze trajectory are available, historical gaze observations provide limited additional information for predicting future actions.

To validate this assumption, we evaluate a variant that conditions future action prediction on the full gaze sequence $\boldsymbol{g}_{0:t+M}$. As shown in Table 13, the variant (denoted as SAGE*) performs slightly worse than the proposed SAGE model in both gaze anticipation and action anticipation, while introducing additional model complexity. Therefore, in our final implementation, the action anticipation module only conditions on future gaze encodings.

In the main paper, we analyze the distinct effects of the GCSA and GCTP modules by reporting the performance of SAGE-12 and SAGE-34. Beyond these settings, SAGE can also be configured in additional combinations to perform specific tasks with or without the GCSA and GCTP components. Here, we extend the ablation study from the main paper by exploring two additional SAGE configurations.

- **SAGE-123:** The model recognizes current actions conditioned on current gaze and additionally predicts future gaze based on the observed gaze trajectory. However, as formulated in Eq. 29, the future gaze variables $\boldsymbol{g}_{t+1:t+M}$ are redundant for the action recognition task, since we assume that given $\boldsymbol{g}_{0:t}$, the current action y_t is conditionally independent of $\boldsymbol{g}_{t+1:t+M}$. We include this configuration to investigate the potential correlation between future gaze and future actions.
- **SAGE-124:** The model recognizes both current and future actions conditioned on current gaze, without involving gaze anticipation. The GCTP module is not activated. Given the formulation of Eq 29, we need to compute future actions conditioning on current action and current gaze features. We random initialize a list of learnable tokens to replace the future gaze encoding in the GCTP module.

A.5.7 Ablation Study on EGTEA Gaze+.

In Table 14, we conduct ablations on an egocentric dataset, i.e. EGTEA Gaze+ dataset.

Table 14: SAGE ablations of GCSA and GCTP on EGTEA Gaze+ dataset.

Models	Gaze Detection			Action Recogn.	Gaze Anticipation			Action Anticipation
	F1	Rec	Prec	Acc	F1	Rec	Prec	Acc
SAGE-1	44.8	61.2	35.3	-	-	-	-	-
SAGE-2	-	-	-	63.1	-	-	-	-
SAGE-12	46.3	61.9	36.4	63.9				
SAGE-123	46.2	61.5	36.2	64.1	36.5	54.3	32.1	-
SAGE-124	45.8	61.7	36.3	64.4	-	-	-	56.3
SAGE	**46.8**	**62.1**	**36.8**	**65.0**	**37.0**	**54.9**	**32.7**	**58.4**

Table 15: Effectiveness of GCTP on EGTEA Gaze+ across different anticipation horizons τ_a.

Models	τ_a	Gaze Detection			Action Recogn.	Action Anticipation
		F1	Rec	Prec	Acc	Acc
SAGE-124	0.5	45.8	61.7	36.3	64.4	56.3
	1.0	45.6	61.5	36.2	64.1	55.7
	1.5	45.9	61.8	36.5	64.1	54.6
	2.0	45.6	61.2	36.3	63.6	53.2
SAGE	0.5	46.8	62.1	36.8	65.0	58.4
	1.0	46.7	62.0	36.7	64.7	57.4
	1.5	46.2	61.9	36.2	64.1	56.3
	2.0	46.5	62.5	36.5	63.7	56.2

GCSA effectiveness. Incorporating the GCSA module significantly improves both gaze detection and action recognition in Table 14. Compared to SAGE-2, which performs action recognition without gaze conditioning, SAGE-12 achieves higher accuracy (63.9% vs. 63.1%) and better gaze detection (46.3 F1 vs. 44.8). This demonstrates that gaze attention provides discriminative spatial priors for action understanding and further stabilizes gaze localization.

GCTP effectiveness. In Table 14, we compare SAGE-124 with the full SAGE model (with GCTP). While SAGE consistently outperforms SAGE-124 across all tasks, the improvement is particularly notable for action anticipation. This shows that the GCTP module effectively captures predictive temporal dependencies between future gaze and future actions. In addition, we also study how future gaze contributes to action anticipation on longer time horizons. We compare the results of SAGE-124 with the full SAGE model on additional τ_a (1.0s, 1.5s and 2.0s) as shown in Table 15. As τ_a increases, action anticipation accuracy decreases notably for both SAGE-124 and SAGE, reflecting the growing uncertainty of long-term prediction. Nevertheless, using future gaze for action anticipation in SAGE consistently achieves better performances than SAGE-124 across all horizons, validating the contribution of GCTP module.

We also analyze how gaze anticipation performance is affected by the presence of the action anticipation task across different anticipation horizons τ_a in Table 16. In SAGE-123, no action anticipation is performed, and the gaze anticipation module learns independently of future action supervision. In contrast, SAGE jointly optimizes both gaze and action anticipation, where the action anticipation module is conditioned on the future gaze encoding, allowing the action loss to propagate gradients back through the gaze anticipation pathway. This joint optimization through SAGE provides an additional supervisory signal that slightly improves gaze anticipation quality over SAGE-123 (e.g., F1 of 37.0 vs. 36.5 at $\tau_a = 0.5s$). Additionally, as τ_a increases, the uncertainty in long-term temporal prediction of future gaze grows, leading to degraded gaze anticipation performance for both SAGE and SAGE-123.

A.5.8 Ablation Study on Exo-Cook.

In Table 17, we conduct similar ablations on an exocentric dataset, i.e. Exo-Cook dataset. We use Sharingan (Tafasca et al., 2024) as the gaze model (SAGE-1).

Table 16: SAGE Ablations on EGTEA Gaze+ of how gaze anticipation performance affected by future actions across different anticipation horizons τ_a.

Models	τ_a	Gaze Detection			Action Recogn.	Gaze Anticipation		
		F1	Rec	Prec	Acc	F1	Rec	Prec
SAGE-123	0.5	46.2	61.5	36.2	64.1	36.5	54.3	32.1
	1.0	46.1	61.4	36.2	63.8	35.7	53.7	31.5
	1.5	45.6	61.5	36.2	63.2	35.1	52.8	31.0
	2.0	45.6	61.5	36.2	62.6	34.4	51.9	30.2
SAGE	0.5	46.8	62.1	36.8	65.0	37.0	54.9	32.7
	1.0	46.7	62.0	36.7	64.7	36.1	54.0	32.0
	1.5	46.2	61.9	36.2	64.1	35.5	53.6	31.8
	2.0	46.5	62.5	36.5	63.7	34.9	52.2	30.6

Table 17: SAGE ablations of GCSA and GCTP on Exo-Cook Dataset.

Models	Gaze Detection			HOI Detn.	Gaze Anticipation			HOI Anticipation
	F1	Rec	Prec	Acc	F1	Rec	Prec	Acc
SAGE-1	56.6	72.8	55.3	-	-	-	-	-
SAGE-2	-	-	-	58.9	-	-	-	-
SAGE-12	57.5	**73.6**	**56.0**	59.9				
SAGE-123	57.5	73.2	55.4	60.2	48.8	62.4	49.8	-
SAGE-124	57.5	73.2	55.6	59.9	-	-	-	52.5
SAGE	**57.7**	73.4	55.5	**60.2**	**49.2**	**62.8**	**50.0**	**54.8**

GCSA effectiveness. SAGE-1 and SAGE-2 perform gaze detection and HOI detection (oracle mode) independently and we observe that combining them into the joint model SAGE-12 through GCSA leads to consistent improvements in both the tasks, which effectively integrates gaze into spatial attention, allowing the model to better capture human-object interactions.

GCTP effectiveness. We observe that SAGE outperforms SAGE-124 in action anticipation, indicating that the GCTP module effectively captures predictive temporal dependencies between future gaze and future actions. We also find that SAGE achieves slightly better gaze anticipation results than SAGE-123, owing to the joint optimization of both future gaze and future action—consistent with the trend observed on the EGTEA Gaze+ dataset (Table 16). It should be noted that since the optimal SAGE model is selected based on best HOI detection performance, the checkpoint yielding the best gaze detection result may differ. As shown in Table 5, the F1 and Recall of SAGE are slightly lower than those of SAGE-12, which also suggests that gaze tends to be noisier and less directly linked to actions in exocentric videos. Overall, the progressive improvements from SAGE-1/2 to SAGE-12 and finally to the full SAGE model demonstrates that both GCSA and GCTP are essential for fully leveraging the bidirectional relationship between gaze and HOIs in both current and future frames.

A.6 Qualitative Results of SAGE

Figure 10 provides qualitative comparisons of our proposed model **SAGE** with four competitive baselines—GC-TSM (Hao et al., 2022), GLC (Lai et al., 2024a), GLC (anticipation), and In-AViT (Roy et al., 2024)—on the EGTEA Gaze+ dataset. The evaluation covers four tasks: gaze detection, gaze anticipation, action recognition, and action anticipation, with an anticipation horizon of $\tau_a = 0.5$ seconds. Each column shows the last frame (at $0s$) in the input sequence and its corresponding ground-truth future frame at $+0.5s$. The left example illustrates a two-step activity transition from *taking a cucumber* to *cutting the cucumber*. The middle example captures the action progression from *taking a cooking utensil* to *putting it down*, while the right example depicts a temporally stable action—*taking bread*. SAGE can performs all four tasks. The gaze heatmap predicted from SAGE for current and future frame is well-aligned with the interaction object in the scene. In the meantime, SAGE can predict current and future actions. GC-TSM does not model gaze and only predicts the current action, failing to anticipate future actions. GLC produces accurate gaze

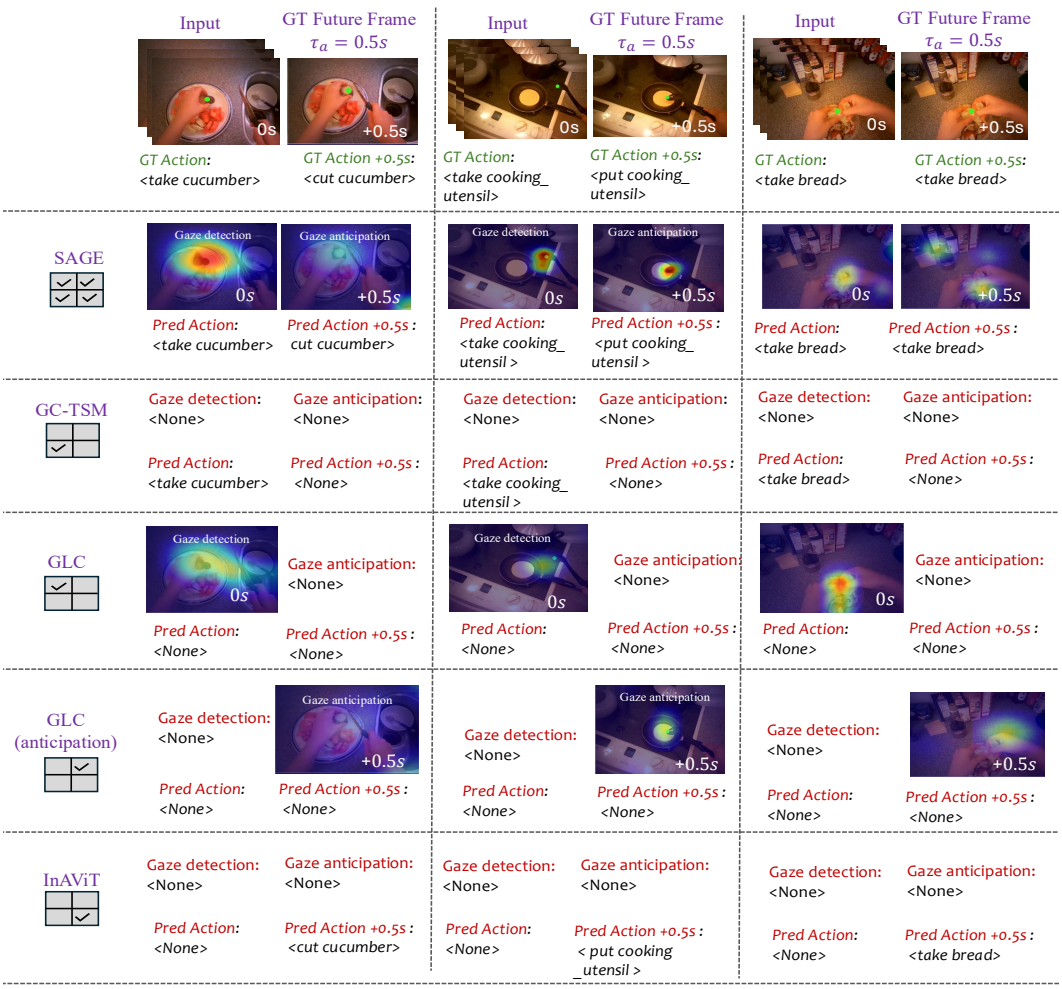


Figure 10: Visualization of SAGE results on the EGTEA Gaze+ dataset, illustrating action and gaze estimation for both current and future frames. The leftmost column displays checkbox representation corresponding to each method, indicating the output settings available. Our method (SAGE) supports all four prediction settings: gaze detection, action recognition, gaze anticipation, and action anticipation.

detection at $0s$, while GLC (anticipation) provides plausible future gaze at $+0.5s$; however, neither model completes all four tasks jointly. Notably, both variants fail in action anticipation. InAViT does not model gaze and can predict future actions.

In Figure 11, we compare SAGE with four state-of-the-art models in different task domain—VideoAttn (Chong et al., 2020), ST-Gaze (Ni et al., 2023), and Sharingan (Tafasca et al., 2024)—using identical input video sequences. Notably, ST-Gaze provides two separate pipelines for HOI detection and HOI anticipation. While these baseline models are typically designed for individual tasks such as gaze detection or HOI classification, SAGE is the only model can jointly performing all four tasks. The figure showcases three representative examples from the Exo-Cook dataset, displaying the last frame at the current time ($0s$) and a future frame at the anticipation horizon $\tau_a = 1s$. The leftmost example illustrates a stable action sequence, while the middle and right examples involve action transitions. Since actions are labeled with categorical indices, we compare category predictions across models. As discussed in the main paper, no existing model is specifically designed for exocentric gaze anticipation; hence, we adapt Sharingan (Tafasca et al., 2024) for this task. Sharingan is the SOTA model for exocentric gaze target detection in images. However, the adapted Sharingan fails to accurately predict future gaze heatmaps across all three examples. Meanwhile, ST-Gaze can perform either HOI detection or anticipation, but not both simultaneously,

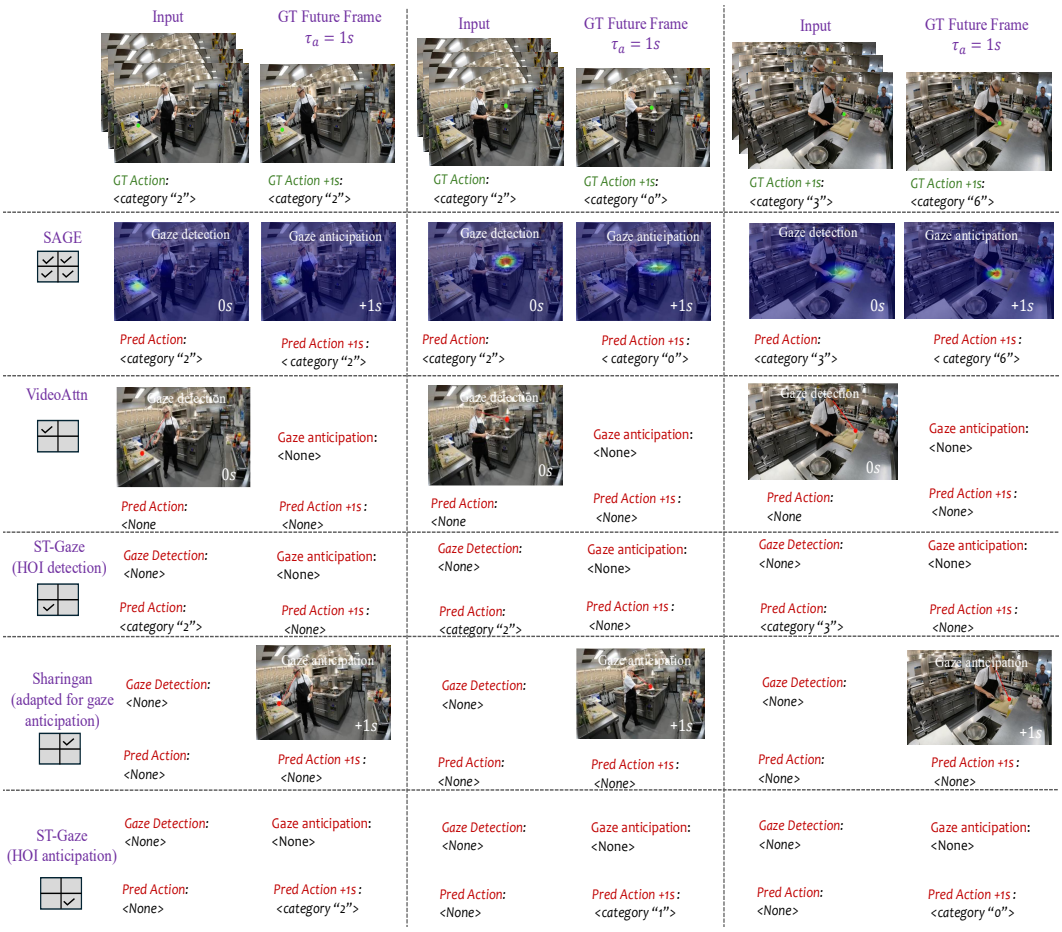


Figure 11: Visualization of SAGE results on Exo-Cook, including action and gaze results for current and future frames. The leftmost column displays checkbox representation corresponding to each method, indicating the output settings available. Our method (SAGE) supports all four prediction settings: gaze detection, action recognition, gaze anticipation, and action anticipation.

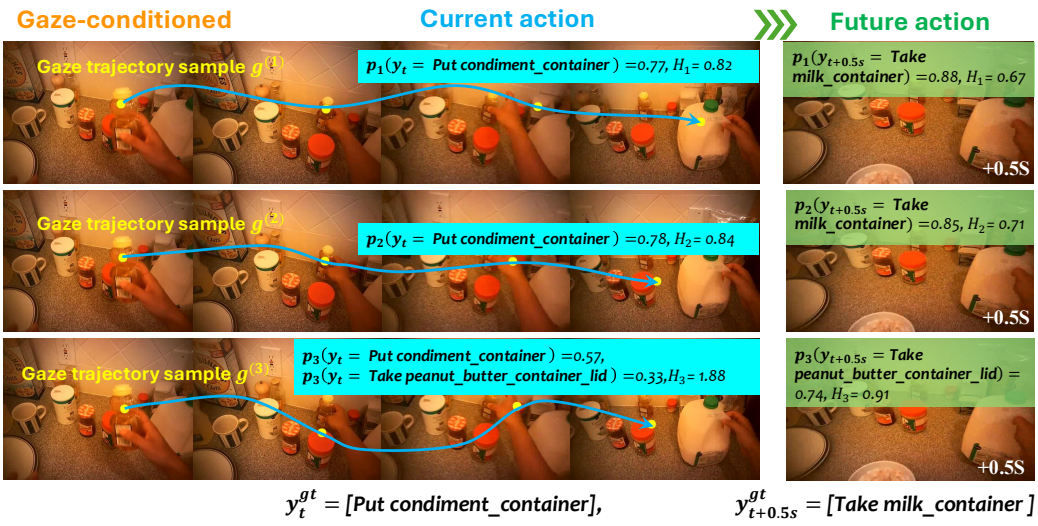
and it fails to predict the correct HOI category in the rightmost example. These results highlight the advantage of SAGE in jointly modeling spatial and temporal dynamics for comprehensive human activity understanding.

A.6.1 Action Uncertainty from Gaze.

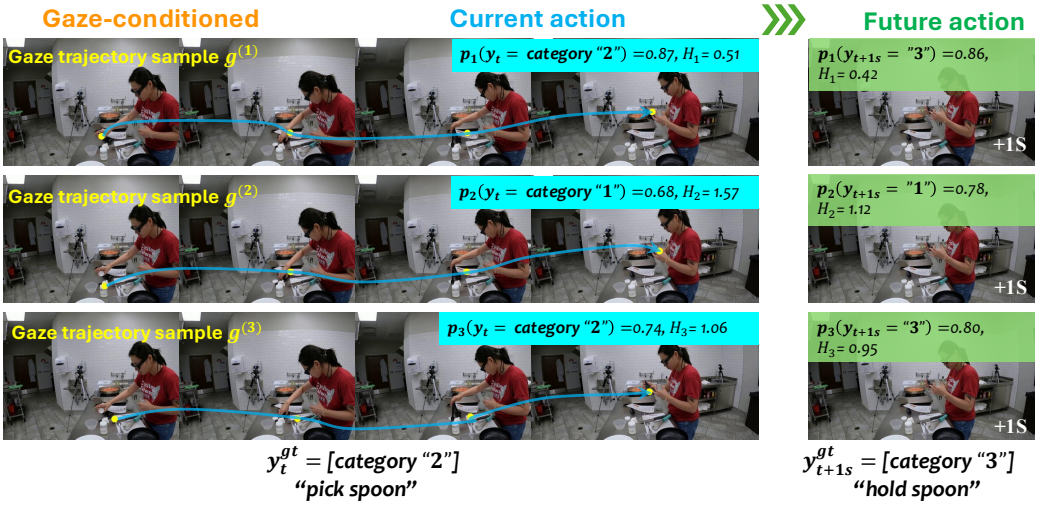
We visualize how variations in gaze affect both current action recognition and future action anticipation. For each input clip, we draw $K = 3$ gaze samples from the predicted gaze distribution and feed them into the current-action and future-action modules. We then compare the resulting action probabilities $p(y_t \mid g^k_{0:t}, I_{0:t})$ and $p(y_{t+M} \mid y_t, g^k_{t+1:t+M}, I_{0:t})$ across samples.

Figure 12 illustrates how variations in gaze influence both current action recognition and future action anticipation. For each clip, we sample multiple gaze trajectories from the predicted gaze distribution and evaluate the corresponding action probabilities.

In Fig. 12 (a), the predicted actions are sensitive to gaze variations. Different gaze samples produce different probability distributions and entropy values. In particular, when the sampled gaze trajectory falls on the peanut butter container (in sample $g^{(3)}$), the model becomes uncertain and may confuse between *Take-peanut-butter-container-lid* and *Take-milk-container* for the future action. This example shows that gaze provides an important cue for disambiguating object-level interactions. Figure 12 (b) demonstrates action-gaze dependency in the exocentric setting. When the



(a) Gaze-conditional action uncertainty example from EGTEA Gaze+.



(b) Gaze-conditional action uncertainty example from Exo-Cook.

Figure 12: Qualitative visualization of gaze-conditioned action sensitivity. For each clip, three sampled gaze trajectories are used to predict the current action y_t and future action y_{t+M}, with corresponding probabilities and entropy H. (a) On EGTEA Gaze+, action predictions vary noticeably with gaze, with certain gaze samples increasing uncertainty and causing confusion between similar actions. (b) On Exo-Cook, different gaze samples lead to distinct predictions; misleading intermediate fixations can result in incorrect current and future action predictions.

sampled gaze trajectory temporarily shifts to the recipe sheet (sample $g^{(2)}$), the model predicts an incorrect action category for both the current and future actions. This indicates that misleading gaze observations can propagate through the interaction reasoning process and affect action inference. The examples demonstrate that SAGE captures how gaze variations influence action predictions and provides an interpretable estimate of gaze-conditioned uncertainty for both action recognition and anticipation.