Sub-actions in Action: Text-Guided Hand-Role Alignment for Sub-Action Recognition

Sadegh Rahmaniboldaji[1]
s.rahmani@surrey.ac.uk
Filip Rybansky[2]
f.rybansky2@newcastle.ac.uk
Quoc Vuong[2]
quoc.vuong@newcastle.ac.uk
Frank Guerin[1]
f.guerin@surrey.ac.uk
Andrew Gilbert[1]
a.gilbert@surrey.ac.uk

[1] University of Surrey
Surrey, UK
[2] University of Newcastle
Newcastle, UK

Abstract

Our objective is to enhance VLMs' ability to distinguish the roles of the left and right hands in egocentric actions. We introduce **sub-action detection**, a task that requires models to explicitly recognise the left- and right-hand roles and their manipulated objects within an ongoing high-level action. To address this challenge, we propose **BiScopeNet**, a framework that augments existing video-LLMs without modifying their vision encoders or language models. BiScopeNet incorporates two components: **FocusAlign**, a gated text-guided attention-based module that selectively aligns video features with textual context to emphasise action-relevant cues, and **MetaLoss**, a masked, object-weighted loss that prioritises visually grounded terms. To supervise sub-action detection, we introduce **BiScopeCap-100K**, a large-scale video-level dataset curated from existing egocentric datasets, providing dense captions that jointly describe high-level actions, left/right-hand roles, and manipulated objects, enriched with spatial metadata for grounding. Across multiple large-scale pretrained VLMs, BiScopeNet outperforms state-of-the-art baselines, improving overall accuracy by 18% and achieving gains of up to more than 15% on hand-specific metrics, while producing substantially fewer unstructured outputs under LLM-as-judge evaluation.

1 Introduction

Online video content and embodied agents increasingly demand structured video understanding beyond coarse action labels, as real-world manipulation tasks often involve multiple sub-actions. However, in many action recognition approaches, a caption typically describes the high-level action (e.g., "a person is picking up a pan"). In contrast, sub-action capture the

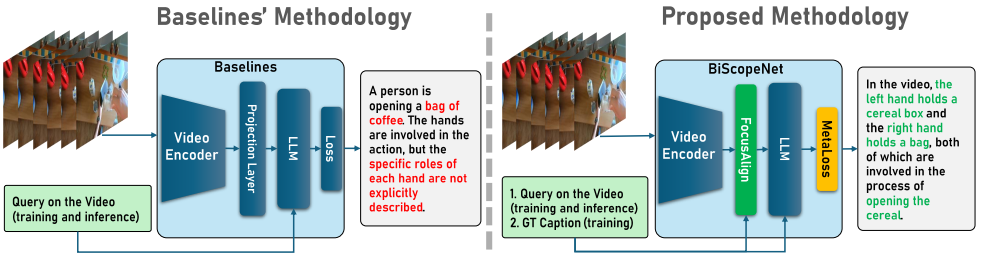


Figure 1: Comparison of BiScopeNet and baseline models, showing how the new data incorporation approach, along with the proposed projection and loss mechanisms, improves hand sub-action recognition, overall action understanding, and object grounding in generated captions.

fine-grained hand-specific contributions within the same activity (e.g. "a person is picking up a pan. The right hand holds the pan and the left hand holds the lid"). Capturing such fine-grained hand-role semantics is important for applications that require precise interaction understanding, including human–robot collaboration, AR/VR environments, and educational or professional training scenarios, where identifying each hand's contribution can improve task interpretation and responsiveness.

Recent advances in video descriptors [3, 8, 9, 29, 31, 32, 33, 48, 56] have significantly improved video understanding on various large-scale benchmarks [11, 19, 20]. However, progress is still primarily evaluated at the level of high-level action recognition or captioning. As a result, models often fail to distinguish hand-specific sub-actions, such as determining which hand manipulates an object and which provides support during an ongoing activity. In egocentric settings, motion blur, occlusion, and visual clutter further complicate the extraction of such fine-grained cues. In this work, we explicitly model sub-actions at the hand level by introducing a structured framework that differentiates between left- and right-hand roles within ongoing high-level actions.

To enable sub-action detection of fine-grained videos, we propose **BiScopeNet**, which consists of two components: **FocusAlign**, a text-guided projection module, and **MetaLoss**, a selective training loss. FocusAlign refines the visual representation before it reaches the language model. Conventional linear or MLP-based projection layers [31, 32] treat visual features uniformly and limit fine-grained alignment with textual cues. Query-based methods such as Q-Former [25] and Flamingo's perceiver resampler [1] compress visual tokens into a fixed set of latent queries prior to cross-modal fusion. Unlike these approaches, our design introduces a structure with visual and textual information exchanged via an attention mechanism and refined via an adaptive, gated residual connection, passing features through an auxiliary alignment MLP. Rather than relying on a single interaction mechanism, our design jointly models cross-modal interactions while preserving text as the guiding signal and dynamically reweighting visual representations. This combination enables the model to prioritise action- and sub-action-relevant visual features, maintaining robust alignment even during inference when textual input is minimal.

In parallel, we introduced MetaLoss, a masked and object-weighted loss that ignores spatial metadata tokens used for structural guidance while increasing the contribution of grounded object tokens. Unlike standard cross-entropy loss, which treats all tokens equally, this selective formulation promotes stronger visual grounding in generated captions. Prior selective objectives such as RHO-LOSS [34] and RHO-1 [28] rely on auxiliary models or adaptive scoring mechanisms to prioritise examples or tokens during training. In contrast,

MetaLoss derives masking decisions deterministically from annotation structure, requiring no auxiliary reference model, while additionally introducing an explicit object-weighting term ($\lambda_o = 3$) that amplifies gradient signals on visually grounded object tokens, a mechanism absent from prior selective training formulations. Figure 1 contrasts our approach with baseline architectures, and detailed ablations are provided in Section 3.1.2.

Effective learning also requires supervision that reflects hand-specific structure. Existing egocentric datasets only partially satisfy this requirement. EPIC-Kitchens/visor [11, 14] and Ego4D [19] provide large-scale coverage but focus primarily on high-level action labels without explicit hand-role supervision. HOI-QA [4] includes sparse hand annotations limited to selected frames, and HD-Epic [38] offers richer captions yet lacks explicit modelling of both high-level action labels and per-hand roles within unified video-level descriptions. Consequently, current benchmarks do not adequately support joint reasoning over high-level actions and hand-specific sub-actions.

To establish a sub-action evaluation benchmark, we introduce **BiScopeCap-100K**, a video-level corpus that provides dense captions describing high-level actions, along with explicit left- and right-hand roles, and is augmented with spatial metadata for grounding. The dataset serves as structured supervision for action- and sub-action detection. A detailed comparison with existing datasets is provided in Table 1.

Our main contributions are summarised as follows:

1. **Sub-action Detection Task:** We introduce a new task, **sub-action detection**, which requires models to explicitly recognise left- and right-hand roles within an ongoing action, alongside the high-level action itself.

2. **BiScopeNet:** We propose a framework composed of two components: **FocusAlign**, a text-guided module that aligns video features with textual context to emphasise action and sub-action cues, and **MetaLoss**, a masked, object-weighted objective that suppresses structural tokens and strengthens visual grounding for fine-grained captioning.

3. **Dataset and SOTA results:** We introduce **BiScopeCap-100K**[1], a large-scale video-level benchmark with dense captions covering both global actions and fine-grained hand-specific sub-actions, enriched with spatial metadata for structured supervision. using this fine-grained supervision, BiScopeNet achieves state-of-the-art performance over recent video descriptors.

2 Related Work

This section reviews prior work relevant to fine-grained egocentric action captioning. We first discuss foundational video modelling and the evolution from action recognition to video captioning, then examine approaches for egocentric action understanding. Finally, we review visual–language fusion strategies in Video-LLMs and recent efforts toward semantic evaluation using LLM-based judges.

Foundations of Video Modelling. Early video understanding established volumetric spatiotemporal feature learning with 3D CNNs such as C3D and I3D [6, 51]. Subsequent work improved long-range temporal reasoning via non-local operations [55] and dual-pathway architectures such as SlowFast [16], while efficient scaling strategies such as X3D [15] balanced speed and accuracy. Transformer-based models, including TimeSformer [5], ViViT

[1]The dataset will be released publicly upon acceptance.

[2], and VidTr [59], introduced global self-attention, enhancing the modelling of short, rapid motions critical for egocentric hand manipulations. Despite these advances, most approaches focus on clip-level action recognition rather than fine-grained, language-grounded descriptions.

From Action Recognition to Video Captioning.

Video captioning extends action recognition by generating temporally grounded natural language. Many standard captioning datasets lack the fine-grained, manipulation-centered annotations needed for egocentric sub-action narratives [11, 19, 22, 43, 61]. Recent Video-LLMs, including Momentor [39], TimeChat [42], VTG-LLM [21], AVicuna [47], and Vid2Seq [57] leverage pretrained LLMs to generate coherent narratives with improved temporal reasoning. However, the strong linguistic capabilities of LLM-based approaches do not necessarily guarantee faithful grounding in subtle visual dynamics, particularly for fine-grained interactions [50].

Egocentric Action Understanding and Manipulation Captioning.

Egocentric video understanding presents unique challenges, including limited field of view, frequent hand–object occlusions, and rapid interactions. Exo2EgoDVC [35] addresses limited labelled data by aligning egocentric cooking videos with YouCook2, while EgoVideo [37] emphasises modelling fine-grained hand–object state transitions beyond high-level action labels, focusing on scene and background details rather than on sub-actions. Motion-aware captioning methods such as the Motion-Augmented Caption Model (M-ACM) [45] further improve motion-sensitive descriptions using dual-pathway streams and human mesh recovery. Despite these advances, most approaches operate at event level or generate high-level narratives. Fine-grained sub-action detection, capturing subtle hand roles, contact states, and short temporal transitions, remains underexplored in captioning architectures, despite its importance for precise manipulation semantics.

Visual–Language Fusion in Video-LLMs.

A key design choice in Video-LLMs is how visual features are integrated into language models [46], either via Analyzer → LLM pipelines [26] that convert videos into intermediate text/structures, or Embedder → LLM that project visual features directly into the LLM token space. Within the latter, prior work uses linear or MLP projectors [24, 27, 31, 32, 33, 50] or Q-Former [25]. Q-Former compresses visual content into query tokens, potentially discarding fine-grained spatiotemporal cues. Separately, Flamingo [1] compresses visual tokens into a fixed set of latent queries via a perceiver resampler and introduces gated cross-attention layers interleaved within the LLM decoder. While effective for general vision-language tasks, the perceiver compression reduces spatial granularity before cross-modal fusion begins, potentially discarding the fine-grained, localised cues that hand-role disambiguation requires In contrast, our proposed projection layer, FocusAlign, operates on the full spatiotemporal grid with caption-conditioned gating to suppress irrelevant background regions, preserving fine-grained hand–object dynamics without aggressive compression.

Semantic Evaluation with LLM Judges. As captioning becomes more fine-grained, traditional n-gram metrics and evaluation protocols such as CIDEr [53], METEOR [23], and SODA/F1 [17] struggle to capture semantic faithfulness and temporal consistency. Recent studies show that LLM-based judges provide scalable evaluation aligned with human reasoning. A large-scale study in the TREC 2024 RAG Track found that LLMs achieved high correlation with human assessments and perfect agreement on support evaluation cases [49]. Similarly, VideoAutoArena and ARGUS use GPT-4o to detect hallucinations and temporal inconsistencies in Video-LLMs [30, 41]. Earlier works such as [31] also employed LLMs in their evaluation pipelines. Motivated by these findings, we adopt an LLM-as-a-judge

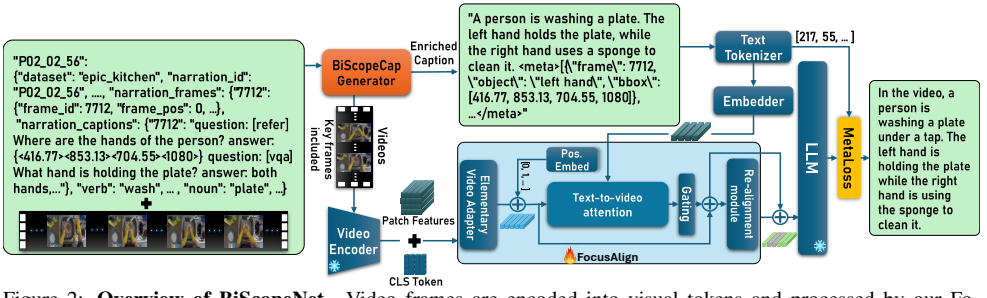


Figure 2: **Overview of BiScopeNet.** Video frames are encoded into visual tokens and processed by our FocusAlign, which integrates text-guided cross-attention, adaptive gating, and MLP refinement to prioritise action-relevant patches. The resulting representation is projected into the LLM space to generate fine-grained captions that capture hand interactions underlying the main action.

framework to assess semantic fidelity in egocentric action/sub-action captioning.

3 Methodology

We present BiScopeNet, a framework for *sub-action detection* in egocentric video captioning. The design of BiScopeNet is driven by two principles: (i) sub-action aware cross-modal alignment, and (ii) selective supervision that prioritises visually grounded hand–object cues. We first describe the overall architecture, then detail the two key components: FocusAlign, which performs task-conditioned cross-modal feature alignment, and MetaLoss, which enforces structured and object-aware supervision.

3.1 BiScopeNet Framework

BiScopeNet is a backbone-agnostic multimodal captioning framework designed specifically for sub-action detection, i.e., distinguishing left- and right-hand roles within an ongoing high-level action. Rather than modifying the decoder or LLM parts, our approach intervenes at the projection stage between the vision encoder and the language model. This is demonstrated in Figure 2, where the video frames are encoded into spatiotemporal tokens, refined by FocusAlign, and then projected into the LLM space for caption generation. Specifically, FocusAlign conditions visual token re-weighting on the task description to amplify action-relevant spatial cues, while MetaLoss enforces structured supervision by masking auxiliary metadata and upweighting object-grounded tokens. Together, these components bridge the gap between generic visual representations and the fine-grained sub-action semantics required for accurate hand-role disambiguation.

3.1.1 Text-Video features alignment.

Sub-action detection requires preserving fine-grained spatial–temporal cues while aligning them with linguistic semantics describing hand roles. Standard linear or MLP projection layers [51, 52] treat visual tokens uniformly, limiting their ability to emphasise sub-action–relevant evidence. To address this, we introduce FocusAlign, a module that performs a more structured cross-modal interaction, with visual and textual information exchanged via attention mechanism and refined via a gated residual connection. Rather than compressing

visual tokens into fixed queries [1, 25] or propagating them unchanged [31, 32], FocusAlign re-weights existing video tokens conditioned on textual context. Formally, let the video encoder produce a sequence of embeddings $\mathbf{V} \in \mathbb{R}^{B\times N_v\times d_v}$ and the language model provide token embeddings $\mathbf{T} \in \mathbb{R}^{B\times N_t\times d_t}$. The video features are first adapted to the text embedding dimension using a learnable projection, Elementary Video Adapter:

$$\hat{\mathbf{V}} = \mathrm{MLP}(\mathrm{LN}(\mathbf{V})) \in \mathbb{R}^{B\times N_v\times d_t} \tag{1}$$

To preserve positional consistency across frames, a learnable positional encoding $\mathbf{P}_v$ is added:

$$\mathbf{V}' = \hat{\mathbf{V}} + \mathbf{P}_v. \tag{2}$$

Next, cross-modal alignment is achieved through an attention mechanism that computes mappings between video and text sequences. Specifically, the video-to-text attention context is defined as:

$$\mathbf{A}_{t\to v} = \mathrm{softmax}\left(\frac{\mathbf{Q}_V\mathbf{K}_T^\top}{\sqrt{d_h}} + \mathbf{M}_t\right), \tag{3}$$

where $\mathbf{Q}_V = \mathbf{W}_q\mathbf{V}'$, $\mathbf{K}_T = \mathbf{W}_k\mathbf{T}$, d_h is the attention head dimension, and $\mathbf{M}_t$ is an additive attention mask for the text tokens.

The attention output is then computed as:

$$\mathbf{O}_V = \mathbf{A}_{t\to v}\mathbf{V}_T, \quad \text{where } \mathbf{V}_T = \mathbf{W}_v\mathbf{T}. \tag{4}$$

The resulting video-attended representation is integrated via a gated residual connection:

$$\tilde{\mathbf{V}} = \mathbf{V}' + \tanh(g_v)\cdot\mathbf{O}_V, \tag{5}$$

where g_v is a learnable scalar gate that controls the fusion strength. This connection retains the original visual representation alongside the text-guided attention output, while the learnable gate prevents premature over-reliance on textual context early in training, allowing the model to gradually shift focus toward action-relevant visual cues.

Finally, a feed-forward refinement is applied to re-align the fused representation with the LLM's token embedding statistics:

$$\mathbf{Z}_V = \tilde{\mathbf{V}} + \mathrm{FFN}(\mathrm{LN}(\tilde{\mathbf{V}})). \tag{6}$$

Where FFN corresponds to the re-alignment module in Figure 2. The residual connection here ensures the cross-modal fusion is preserved while the FFN non-linearly re-projects the fused features into the LLM's expected token embedding space without overwriting the hand–object grounding signals accumulated in the previous step.

This task-conditioned alignment selectively amplifies visual evidence relevant to hand-specific sub-actions, improving role disambiguation without altering the decoder or token budget. Figure 2 shows how the FocusAlign refines video embeddings via input text. Also, the forward pass of the FocusAlign is detailed in the supplementary material.

3.1.2 Selective Supervision for Sub-Action Learning

Conventional language modelling losses assign equal importance to all tokens, regardless of their semantic or structural role. However, our training data contains both natural-language captions and structured metadata spans (e.g., <meta>...</meta>) that describe auxiliary information such as object positions or spatial annotations. These metadata tokens are not intended to contribute directly to the language modelling objective. Prior selective training strategies [28, 34] address token or sample importance through adaptive weighting or prioritisation mechanisms, often relying on auxiliary scoring procedures or reference models. In contrast, we introduce **MetaLoss**, a lightweight supervision strategy that selectively masks metadata spans and upweights object-related tokens to emphasise visually grounded semantics. Importantly, our method requires no auxiliary reference model, as masking positions are determined directly from the annotation structure during dataset curation.

Let the model produce logits $\mathbf{y} \in \mathbb{R}^{B\times T\times V}$ and the corresponding target tokens be $\mathbf{t} \in \{0,\dots,V-1\}^{B\times T}$, where B is the batch size, T is the sequence length, and V is the vocabulary size. We define a binary mask $m_i \in \{0,1\}$ for each token position i as:

$$m_i = \begin{cases} 0, & \text{if } t_i \text{ lies in meta span,} \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

This mask excludes tokens between <meta> and </meta> from loss computation, preventing the model from wasting capacity on structural markup and encouraging it to focus on semantically meaningful content (actions, hands, and objects). We construct the object set to mask O by taking the top-K object nouns from BiScopeCap captions and mapping them to tokenizer IDs. This yields a compact vocabulary of frequently manipulated objects (e.g. pan, tap, knife, mug) rather than a large open-vocabulary lexicon. Next, we introduce a per-token weighting term w_i to emphasise object mentions:

$$w_i = \begin{cases} \lambda_o, & \text{if } t_i \in \mathcal{O}, \\ 1, & \text{otherwise,} \end{cases} \tag{8}$$

Where $\mathcal{O}$ denotes the set of vocabulary indices corresponding to object terms derived from metadata provided with the ground-truth captions, and $\lambda_o > 1$ is a coefficient controlling the degree of emphasis. We set $\lambda_o = 3$ with ablation results comparing MetaLoss to standard cross-entropy shown in Table 5. This weighting mechanism increases the gradient magnitude associated with visually grounded words, thereby strengthening multimodal alignment. Token-level cross-entropy for position i is computed as:

$$\mathrm{CE}(\mathbf{y}_i, t_i) = -\log \mathrm{softmax}(\mathbf{y}_i)_{t_i}. \tag{9}$$

Combining masking and weighting, the overall MetaLoss objective is formulated as:

$$\mathcal{L}_{\text{meta}} = \frac{\sum_i m_i w_i \,\mathrm{CE}(\mathbf{y}_i, t_i)}{\sum_i m_i w_i}. \tag{10}$$

Intuitively, MetaLoss extends standard cross-entropy by (i) masking <meta> spans (e.g., structured punctuation or bounding-box markup) and (ii) upweighting object tokens to emphasise semantically meaningful content. It requires no modification to the optimiser or logits, relying only on per-token scalar weights and a binary mask. This encourages focus on visually relevant tokens while maintaining stable, normalised optimisation. Because

MetaLoss reweights terms based on textual metadata rather than instance-specific patterns, it is data-agnostic and applicable to captioning tasks with structured targets. The complete algorithm is provided in the supplementary material.

4 Experiments

We introduce **BiScopeCap** and evaluate our framework for sub-action detection in egocentric video captioning. We first outline model configurations and implementation details, followed by dataset description and experimental results.

Implementation Details: All experiments are implemented in PyTorch and HuggingFace Transformers and trained on three NVIDIA A100 GPUs. We evaluate BiScopeNet across multiple VLM backbones, including NVILA [29] and VideoChatGPT [31]. Unless otherwise stated, all models are trained using AdamW (without weight decay) with a learning rate of 2×10^{-5} and batch size 8. The train–test split follows Table 1.

For NVILA, which employs Qwen2-VL [54] as the LLM and SigLIP [58] as the vision encoder, we fine-tune the model for 3 epochs using 8 uniformly sampled frames per video. For VideoChatGPT, which employs Vicuna 1.5 [10] as the LLM and CLIP [40] as the vision encoder, we train for 12 epochs using 100 uniformly sampled frames per video. BiScopeNet is integrated into these architectures without modifying the core language or vision encoders. During fine-tuning, both the vision encoder and the language model remain frozen, and only FocusAlign is trained, ensuring that performance gains arise from improved video-text alignment rather than backbone optimisation. Additionally, we conduct an ablation study on the effect of different vision encoders within the VideoChatGPT backbone, as detailed in Section 4.4.3.

4.1 BiScopeCap dataset

Action recognition benchmarks span egocentric and exocentric settings. Charades-Ego [43] provides paired first- and third-person videos for viewpoint transfer but focuses on high-level action categories without modelling hand-specific roles. Ego-Exo4D [20] further supports action recognition and human–object interaction across paired views. Within the egocentric domain, the EPIC-KITCHENS family [11, 12, 13, 14] provides large-scale recordings of unscripted daily activities and has become a standard benchmark for action recognition and anticipation. Similarly, HD-EPIC [38] extends these annotations with richer multimodal signals and detailed descriptions. While these datasets enable high-level action recognition, they are not designed for sub-action detection, which requires explicit left–right hand role modelling. As shown in Table 1, they lack hand-role attribution and spatially grounded video-level captions capturing fine-grained manipulation semantics. The closest effort, HOI-QA, provides frame-level hand–object annotations for selected frames only, without temporally coherent video-level supervision. To address this gap, we introduce BiScopeCap, which provides dense video-level captions that describe high-level actions and each hand's role, augmented with textual spatial metadata.

Dataset Curation: BiScopeCap is built through a unified annotation pipeline to support sub-action detection. Instead of new recordings, we standardise and enrich data from EPIC-KITCHENS, EPIC-VISOR [11, 12, 13, 14], Ego4D [19], and HD-EPIC [38]. Frame-level hand–object annotations from HOI-QA [4] are consolidated into temporally coherent, video-level captions using Gemini [18] only for structured summarisation. All captions are re-

Table 1: Comparison between existing egocentric datasets and our proposed **BiScopeCap**, along with the distribution of samples across its constituent subsets.

Dataset	Action Label	Hand Role Labels	Spatial Metadata	Coarse-grained Captions	Fine-grained Captions	Domain Source	# Train Samples	# Test Samples
Epic-Kitchens / Visor	✓	✗	✗	✗	✗	Kitchen / Daily Tasks	32,727	3,729
Ego4D	✓	✗	✗	✗	✗	Diverse Activities	23,598	2,622
HOI-QA	✓	✗	✓	✗	✓	Kitchen / Daily Tasks	–	–
HD-Epic	✓	✓	✗	✗	✓	Kitchen / Daily Tasks	37,229	4,137
BiScopeCap (ours)	✓	✓	✓	✓	✓	Diverse Activities	**93,554**	**10,488**

Table 2: LLM-as-a-judge scoring of captions compared with a human user.

Method	# Samples	% ActRec ↑	% RH↑	% LH↑	% ObjRec↑
GPT Scoring	100	40.3	14.0	31.0	52.0
Human Scoring	100	39.6	16.6	31.3	55.0

formulated into a unified format integrating (i) high-level actions, (ii) fine-grained hand roles, and (iii) structured spatial metadata. This harmonised strategy produces a single, task-aligned corpus explicitly designed for sub-action detection. To illustrate how our annotation strategy supports sub-action detection, we summarise each source's contribution. EPIC-KITCHENS and Ego4D provide narrations and temporally localised actions, while EPIC-VISOR adds segmentation masks that HOI-QA uses to extract bounding boxes for hands and objects along with keyframe captions. These signals, narrations, temporal boundaries, and hand–object regions, are unified into structured video-level captions using Google Gemini 2.0-Flash [18], conditioned on human narrations and annotated regions to preserve explicit grounding and accurate hand-role attribution. For HD-EPIC, Although it includes hand motion descriptions in video-level, these often omit the high-level action and lack structured metadata. We therefore reformulate and augment its annotations using available hand and object segmentations to align with our unified sub-action–oriented caption format.

Clips are segmented by narration timestamps and resized to the model's input resolution, with bounding boxes temporally and spatially re-aligned to preserve accurate grounding. BiScopeCap is constructed solely from research-licensed datasets (EPIC-KITCHENS, Ego4D, HD-EPIC, HOI-QA). To address concerns about annotation hallucination, we conducted a manual audit of 200 randomly sampled training clips balanced across datasets. We observed that $\approx 5\%$ contain minor descriptive inaccuracies (e.g. imprecise verbs), while only $\approx 8\%$ and $\approx 10\%$ exhibit left and right hand hallucination, respectively. We consider this level of noise acceptable, as the model is intended to learn generalized grounding patterns rather than memorize individual annotations. Similar annotation noise has been reported in other large-scale egocentric datasets [19], where large-scale manual or semi-automatic labelling inevitably introduces errors and biases. All derived annotations and curation scripts will be released, while the raw videos remain under their original licences.

Dataset Statistics: We summarise key statistics highlighting BiScopeCap 's suitability for sub-action detection (see Table 1). It contains 12,814 action instances and 5,727 active objects (including hands), covering diverse manipulation scenarios. Analysing left/right-hand and bimanual interactions confirms balanced supervision for hand-role disambiguation. Object frequency further demonstrates broad coverage across contexts. These statistics show that BiScopeCap provides diverse, grounded supervision for fine-grained sub-action understanding beyond high-level action classification. Additional details are in the supplementary.

Table 3: Comparison of **BiScopeNet** with state-of-the-art VLMs on BiScopeCap. **Blue** denote the best results.

Method	Vision Encoder	Backbone	Trained	% ActRec ↑	% RH ↑	% LH ↑	% ObjRec ↑	# Unstruct ↓
SmolVLM [33]	SigLIP [58]	-	✗	10.4	11.8	14.6	27.6	4505
Qwen 2.5-VL [3]	-	-	✗	25.6	29.9	37.0	54.0	1
Qwen 3.5 [48]	-	-	✗	26.0	32.2	40.6	53.3	6
Intern VL 2.5 [8]	InternViT [8]	-	✗	24.8	27.4	37.0	51.5	5
NVILA [25]	SigLIP	-	✗	14.1	17.2	28.7	39.7	0
NVILA [25]	SigLIP	-	✓	44.0	43.0	39.1	58.0	0
Qwen 3.5 [48]	-	-	✓	41.8	42.0	40.0	54.3	40
VideoChatGPT [31]	CLIP [40]	-	✓	14.3	13.7	17.7	30.1	488
BiScopeNet (Ours)	CLIP	VideoChatGPT	✓	30.7	19.4	29.6	47.9	7
BiScopeNet (Ours)	DinoV3 [42]	VideoChatGPT	✓	33.4	31.9	31.4	53.5	5
BiScopeNet (Ours)	SigLIP	NVILA	✓	**45.5**	**44.9**	**41.2**	**58.9**	0

4.2 Metric Scoring System

To evaluate performance on sub-action detection, we assess model outputs along four complementary dimensions: (1) high-level action recognition (**ActRec**), (2) right-hand role recognition (**RH**), (3) left-hand role recognition (**LH**), and (4) object recognition (**ObjRec**). Together, these metrics assess whether a model correctly and explicitly captures the high-level action, along with the left- and right-hand roles, and their interactions with objects that define sub-actions. Accuracy is used for all dimensions.

We adopt an LLM-as-a-judge evaluation protocol using GPT-5 Mini [36] to score generated captions across these categories. Importantly, caption evaluation relies on a separate model from BiScopeCap generation, ensuring independence between data construction and assessment. The full evaluation prompt is provided in the supplementary material. To verify reliability, we manually scored 100 randomly sampled predicted captions from the EPIC-KITCHENS subset and compared these with human judgments and GPT-5-based scores. The automatic evaluation closely matches human assessment, with overall accuracy differing by less than one percentage point and hand- and object-specific metrics varying by only 2–3 points on average (see Table 2). These results indicate that the LLM judge provides a consistent and reliable assessment of the correctness of fine-grained sub-actions.

Table 2 shows that the GPT-based evaluation closely aligns with human judgements across all considered metrics. The differences are minimal, typically within a few percentage points, confirming that automated scoring reliably reflects fine-grained performance in hand-role and object recognition. As sub-action detection requires structurally coherent captions that explicitly separate high-level actions and hand-specific roles, invalid or unstructured outputs can directly compromise evaluation. Vision–language models occasionally generate captions that are grammatically inconsistent, improperly formatted, or contain extraneous tags (see Figure 4). We therefore apply an additional filtering prompt using GPT-5 Mini [36] (provided in the supplementary material) to identify such outputs. The frequency of these cases is recorded as an indicator of model stability under structured sub-action supervision. These unstructured captions are treated as incorrect predictions, as they fail to provide evaluable hand-role or object descriptions. For completeness, we also report scores without penalising such samples, illustrating how structured-caption compliance influences overall sub-action detection performance (see the supplementary material).

4.3 BiScopeNet Performance Analysis

For fair comparison, we evaluate across both zero-shot and supervised configurations. We compare against off-the-shelf VLMs, SmolVLM [33], Qwen [3, 48], InternVL [8], and

Table 4: Ablation study of **BiScopeNet** components using NVILA backbone [29] on BiScopeCap. Results compare projection variants and loss configurations. The combination of the proposed FocusAlign and MetaLoss yields consistent improvement across all metrics, demonstrating their complementary roles in sub-action understanding. **Blue** highlights the best results.

Method	Component	Vision Encoder	Backbone	% ActRec ↑	% RH ↑	% LH ↑	% ObjRec ↑	# Unstruct ↓
NVILA	MLP + Cross Entropy	SigLIP	-	44.0	43.0	39.1	58.0	0
BiScopeNet	MLP + MetaLoss	SigLIP	NVILA	44.0	44.1	39.1	58.7	0
BiScopeNet	FocusAlign + MetaLoss	SigLIP	NVILA	**45.5**	**44.9**	**41.2**	**58.9**	0

NVILA [29], in zero-shot mode without task-specific adaptation on BiScopeCap. For VideoChatGPT, zero-shot performance was consistently poor across all metrics, so we exclude this configuration from the main comparison. We further include supervised baselines, where models such as NVILA [29], Qwen 3.5 [48], and VideoChatGPT [31] are fine-tuned on BiScopeCap using their original architectures without integrating BiScopeNet. Finally, we evaluate BiScopeNet integrated into both VideoChatGPT and NVILA backbones, where the proposed modules are introduced while keeping the core language and vision backbones frozen and optimizing only the projection-related components. This evaluation protocol allows us to distinguish improvements arising from general zero-shot capability, supervised task adaptation, and the additional gains introduced by BiScopeNet.

Table 3 reports results on the full BiScopeCap, which combines the Ego, Epic, and HD-Epic subsets for comprehensive evaluation across diverse egocentric settings. Overall, BiScopeNet consistently outperforms competing VLM baselines across all evaluation metrics, including action recognition (ActRec), left/right hand understanding (LH, RH), object recognition (ObjRec), and structured caption generation (Unstrct).

Among the baseline architectures, the supervised version of NVILA [29] achieves a substantial improvement over its zero-shot counterpart, increasing ActRec from 14.1% to 44.0%, which highlights the effectiveness of BiScopeCap for fine-grained action and sub-action learning. Integrating BiScopeNet into the NVILA backbone yields further gains across all metrics, achieving the best overall performance while maintaining fully structured outputs with zero malformed generations. These improvements demonstrate that the proposed FocusAlign and MetaLoss provide complementary benefits beyond standard supervised adaptation and generalize effectively across different VLM architectures.

To further evaluate backbone generalizability, we additionally integrate BiScopeNet into VideoChatGPT [31]. Using the same CLIP vision encoder as the original VideoChatGPT model, BiScopeNet improves ActRec from 14.3% to 30.7% and ObjRec from 30.1% to 47.9%, while simultaneously reducing unstructured outputs from 488 to only 7. Replacing CLIP with DINOv3 further improves performance. These results indicate that BiScopeNet not only improves reasoning and grounding within existing VLM backbones, but also benefits from stronger visual representations provided by more advanced vision encoders.

4.4 Ablation Studies

4.4.1 Model Architectures.

Table 4 presents an ablation study of BiScopeNet components on BiScopeCap using NVILA as the backbone. Applying MetaLoss alone with the standard MLP projection already yields modest improvements over the original NVILA configuration, particularly in RH and ObjRec performance. Incorporating the proposed FocusAlign together with MetaLoss produces the

Table 5: Ablation study of **BiScopeNet** components using VideoChatGPT backbone [31] on the EPIC subset of BiScopeCap. Results compare projection variants and loss configurations. The combination of the proposed FocusAlign and MetaLoss yields the most consistent improvement across all metrics, demonstrating their complementary roles in sub-action understanding. **Blue** highlights the best results.

Method	Component	Vision Encoder	% ActRec ↑	% RH ↑	% LH ↑	% ObjRec ↑	# Unstruct ↓
VideoChatGPT	Linear + Cross Entropy	CLIP	12.8	10.4	5.9	27.3	374
VideoChatGPT	MLP + Cross Entropy	CLIP	16.8	15.5	12.8	26.4	437
BiScopeNet	FocusAlign + Cross Entropy	CLIP	13.9	10.6	12.9	18.8	196
BiScopeNet	Linear + MetaLoss	CLIP	19.6	19.0	25.8	40.4	0
BiScopeNet	FocusAlign + MetaLoss	CLIP	**32.5**	18.4	25.8	46.4	5
BiScopeNet	FocusAlign + MetaLoss + CLS	CLIP	30.7	**19.4**	**29.6**	**47.9**	7

strongest and most consistent gains across all evaluation metrics. These results demonstrate the complementary roles of FocusAlign and MetaLoss in improving fine-grained sub-action understanding and visually grounded reasoning. To investigate if our approach generalizes across different baselines, we tested our approach on the weakest baseline, VideoChatGPT [31], on the EPIC subset of BiScopeCap; the results are reported in Table 5. Replacing the linear projection in VideoChatGPT [31] with an MLP improves the overall accuracy, indicating that a deeper mapping better aligns modalities. However, using FocusAlign alone yields inconsistent gains, improving hand-related metrics but reducing object accuracy, suggesting that FocusAlign without proper guidance may overfocus on local cues. When coupled with the proposed MetaLoss, performance increases substantially across all dimensions, while the unstructured error drops sharply, confirming a strong correlation between feature selection and loss supervision. Adding a CLS token slightly stabilises hand and object recognition, but marginally lowers global accuracy, implying a trade-off between summarization and fine-grained diversity. Overall, the FocusAlign –MetaLoss combination proves essential for robust video–text alignment and neither component alone yields the same improvement.

Table 6: Comparison of **BiScopeNet** using VideoChatGPT [31] as backbone with state-of-the-art VLMs across **Ego**, **Epic+Ego**, and **HD-Epic** subsets of BiScopeCap. **Blue** denote the best results.

	Method	Trained	Vision Encoder	% ActRec↑	% RH↑	% LH ↑	% ObjRec↑	# Unstruct↓
Ego4D	SmolVLM [33]	✗	SigLIP	8.9	12.0	16.5	25.8	1256
	NVILA [29]	✗	SigLIP	13.0	16.7	30.9	39.7	0
	VideoChatGPT [31]	✓	CLIP	7.2	10.1	12.3	24.0	563
	BiScopeNet (Ours)	✓	DinoV3	**22.1**	**38.7**	**39.6**	**42.1**	46
Epic+Ego	SmolVLM [33]	✗	SigLIP	13.1	15.6	18.9	29.6	2785
	NVILA [29]	✗	SigLIP	17.7	21.5	34.9	43.1	0
	VideoChatGPT [31]	✓	CLIP	14.5	20.7	22.6	31.0	380
	BiScopeNet (Ours)	✓	DinoV3	**33.5**	**33.4**	**34.9**	**53.6**	4
HD-Epic	SmolVLM [33]	✗	SigLIP	6.4	5.9	8.2	24.6	1720
	NVILA [29]	✗	SigLIP	8.8	10.7	**19.3**	34.5	0
	VideoChatGPT [31]	✓	CLIP	7.1	8.2	8.5	28.8	0
	BiScopeNet (Ours)	✗	DinoV3	10.1	10.7	15.8	24.0	0
	BiScopeNet (Ours)	✓	DinoV3	**28.5**	**21.4**	15.3	**57.2**	10

4.4.2 Ablation on Dataset Subsets.

To analyse the robustness of sub-action detection under different supervision sources, we evaluate BiScopeNet on individual subsets using the VideoChatGPT backbone of BiScopeCap:

Ego, Epic+Ego, and HD-Epic (see Table 6). Across all splits, BiScopeNet consistently achieves the strongest performance in nearly every metric. On the Ego and Epic+Ego subsets, the model demonstrates particularly strong left- and right-hand recognition accuracy, indicating effective learning of fine-grained hand-role distinctions. On HD-Epic, improvements are especially pronounced in object recognition, reflecting reliable grounding in visually dense and high-resolution scenarios. The consistency of these gains across subsets with different characteristics suggests that the improvements stem from task-aligned cross-modal alignment and selective supervision, rather than overfitting to a particular data source. Overall, the results confirm that BiScopeNet generalises sub-action detection across varied egocentric environments and interaction contexts. We also study the cross-subset generalisation capability of our approach by training BiScopeNet on Epic+Ego and evaluating it on the HD-Epic subset. Despite not being trained on the HD-Epic distribution, our model achieves better results than the original VideoChatGPT, even when the latter is trained on HD-Epic. This demonstrates the generalisation ability of the proposed approach.

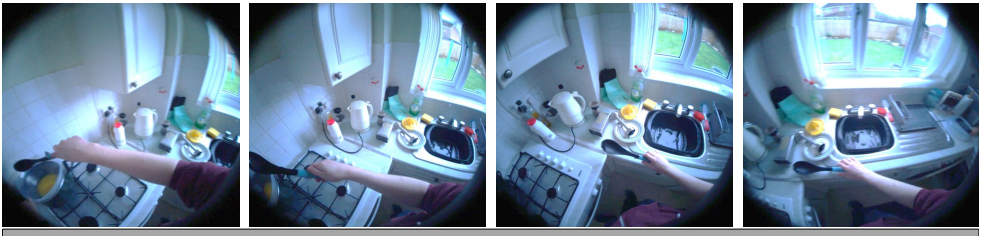
Table 7: Ablation study of **BiScopeNet** integrated into the VideoChatGPT [31] backbone with different vision encoders (CLIP, DINOv3, and SigLIP2) and the presence or absence of the CLS token on the EPIC subset of BiScopeCap. The proposed framework consistently improves over the baseline across all encoders. SigLIP2 with CLS achieves the highest overall and object-level accuracy, while DINOv3 shows superior sensitivity to hand-specific cues, demonstrating the adaptability of BiScopeNet to diverse vision backbones. **Blue** highlights the best results.

Method	Vision Encoder	CLS	% ActRec ↑	% RH ↑	% LH ↑	% ObjRec ↑	# Unstruct ↓
VideoChatGPT	CLIP	✗	12.8	10.4	5.9	27.3	374
BiScopeNet	CLIP	✗	32.5	18.4	25.8	46.4	5
BiScopeNet	CLIP	✓	30.7	19.4	29.6	47.9	7
BiScopeNet	DinoV3	✗	33.1	17.6	36.6	48.0	5
BiScopeNet	DinoV3	✓	28.2	**23.0**	**38.2**	49.4	13
BiScopeNet	SigLIP2	✗	32.4	21.6	31.8	47.0	5
BiScopeNet	SigLIP2	✓	**38.1**	21.9	15.5	**51.4**	3

4.4.3 Ablation on Vision Encoder and CLS Token

We further conduct an ablation study by evaluating BiScopeNet using VideoChatGPT as backbone with different vision encoders and with/without the CLS token to analyse the effect of visual feature representations on sub-action detection performance. Results on the EPIC subset of BiScopeCap are reported in Table 7. We consider three widely used encoders, CLIP [40], SigLIP2 [52], and DINOv3 [44], and compare against the baseline VideoChatGPT [31] under identical training settings. Across all encoders, BiScopeNet consistently outperforms the baseline, indicating that the gains arise from our task-conditioned alignment and selective supervision rather than vision-encoder-specific properties. While SigLIP2 with a CLS token achieves the strongest overall and object-level accuracies, DINOv3 shows competitive performance on hand-specific metrics, suggesting complementary strengths in semantic grounding and sensitivity to local motion.

The effect of the CLS token varies across encoders, modestly influencing high-level versus fine-grained metrics, but the relative improvement of BiScopeNet remains stable. Overall, these results demonstrate that BiScopeNet functions as a backbone- and vision-encoder-agnostic framework for sub-action detection, improving hand-role and object grounding without relying on a particular vision encoder.



GT: The person is placing a large plastic spoon onto the countertop. The right hand is positioning the spoon down onto the side of the countertop.

SmolVLM: The video shows a person cooking in a kitchen. The person is holding a spoon in their right hand and stirring a pot on the stove with their left hand. The stove has four burners, and the pot is on the front right burner. The person is wearing a long-sleeved purple shirt. The kitchen has white cabinets and a white countertop. There is a window on the right side of the kitchen, and there is a sink on the left side of the kitchen. The sink

NVILA: The overall action being performed in the video is washing dishes. The left hand is holding a sponge, which is used to scrub the dishes clean. The right hand is holding a dish, which is being cleaned by the sponge.

VideoChatGPT: The video shows a person washing dishes in a kitchen sink. The person is using both hands to wash the dishes, with one hand holding a sponge and the other hand holding a dish. The person is also using a spray bottle to clean the dishes. The video captures the ongoing action of the person washing the dishes and provides a clear view of their hand movements.

QWEN 3.5: The overall action being performed in the video is stirring a pot on the stove. \n\nThe left hand is holding a spoon and is actively stirring the contents of the pot, which appears to be a yellow liquid. The right hand is holding the handle of the pot, providing stability and control as the left hand stirs. The right hand remains steady, ensuring that the pot does not move excessively while the left hand performs the stirring motion.

Intern VL 2.5: The overall action being performed in the video is the washing of a spoon. The person is using a sponge to scrub the spoon, ensuring it is clean. The left hand is holding the spoon in place, while the right hand is using the sponge to scrub the spoon. This coordinated effort allows the person to effectively clean the spoon.

BiScopeNet (Ours): The person in the video is placing a plastic spoon down onto the countertop. The right hand is holding the spoon and positioning it carefully onto the surface.

Correct | Neither wrong nor relevant | Wrong

Figure 3: Qualitative comparison of visual understanding across methods.

Ours – Structured: In the video, a person is seen pushing a tree on the ground using a cart. The left hand and right hand are both involved in manipulating the cart to move the tree.

SmolVLM – Unstructured (Unfinished sentence): The image depicts a person holding a long-handled gardening tool, which appears to be a shovel or a pickaxe. The individual is standing on a grassy surface, which suggests that they are in a garden or a similar outdoor setting. The person is wearing a blue sweatshirt with white drawstrings. The sweatshirt has a hood, which is not visible in the image.\n\nThe individual is holding the long-handled gardening tool with both hands. The left hand is

Ours – Structured: In the video, a person is picking up dough from a table. The left hand is actively picking up the dough, while the right hand assists in the action.

VideoChatGPT – Unstructured (Inclusion of tags): A person is picking up dough from a bowl. The person is using both hands to pick up the dough. <video>\n
 <img src=\"https://www.invidio.com/static/images/medias/original/430x240/130121-04.jpg\"“

Figure 4: Qualitative comparison of output structure across methods.

4.5 Qualitative Results

We present qualitative comparisons to highlight differences in fine-grained visual grounding and caption structure. As shown in Figure 3, baseline models often fail to correctly localise hands, distinguish their roles, or associate them with manipulated objects. In contrast, our approach (on VideoChatGPT backbone) produces more coherent hand–object–action reasoning, leading to semantically grounded sub-action descriptions. Additional examples are provided in the supplementary material. We also illustrate unstructured captioning failures in Figure 4. Baseline outputs frequently contain incomplete sentences or extraneous artefacts (e.g., tags or URLs), which reduce clarity and structural consistency.

5 Conclusion

We presented BiScopeNet, a framework for sub-action recognition in egocentric video captioning that models left– and right-hand roles through a selective text-guided projection module, FocusAlign, coupled with an object-aware loss, MetaLoss. Together with the proposed

dataset, BiScopeCap, which provides structured video-level supervision for high-level actions, hand roles, and spatial metadata, our approach promotes more consistent and grounded hand–, object–, and action- reasoning. Across multiple video descriptors, we observe consistent improvements in hand-specific accuracy, action recognition, and caption structure, further supported by an LLM-as-a-judge protocol aligned with human ratings.

5.0.1 Limitations and future work

BiScopeNet samples a fixed number of frames per clip, which limits its ability to model sub-actions in long-term videos where hand roles evolve over extended temporal spans. Future work will investigate adaptive frame sampling and temporal segmentation strategies to extend sub-action detection to longer egocentric sequences.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, October 2021.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Zhou, Minghui Liu, Jianwei Wang, An Yang, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

[4] Siddhant Bansal, Michael Wray, and Dima Damen. Hoi-ref: Hand-object interaction referral in egocentric vision. *arXiv preprint arXiv:2404.09933*, 2024.

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. URL https://arxiv.org/abs/2412.05271.

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. URL https://arxiv.org/abs/2312.14238.

[9] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. URL https://arxiv.org/abs/2406.07476.

[10] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43 (11):4125–4141, 2021. doi: 10.1109/TPAMI.2020.2991965.

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. URL https://doi.org/10.1007/s11263-021-01531-2.

[14] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022.

[15] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020.

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[17] Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. Soda: Story oriented dense video captioning evaluation framework. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, page 517–531, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58538-9. doi: 10.1007/978-3-030-58539-6_31. URL https://doi.org/10.1007/978-3-030-58539-6_31.

[18] Google. Gemini (april, 2025 version), 2025. https://gemini.google.com/app. Accessed: 21 December 2025.

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990, 2022. doi: 10.1109/CVPR52688.2022.01842.

[20] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina González, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbeláez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg,

Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shout, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19383–19400, 2024. doi: 10.1109/CVPR52733.2024.01834.

[21] Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3302–3310, 2025.

[22] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017.

[23] Alon Lavie and Abhaya Agarwal. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA, 2007. Association for Computational Linguistics.

[24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[26] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding, 2024. URL https://arxiv.org/abs/2305.06355.

[27] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6657–6678, 2024.

[28] Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Not all tokens are what you need for pretraining. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 29029–29063. Curran Associates, Inc., 2024. doi: 10.52202/079017-0914. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3322a9a72a1707de14badd5e552ff466-Paper-Conference.pdf.

[29] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Andriy Myronenko, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, , and Yao Lu. Nvila: Efficient frontier visual language models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

[30] Ziyang Luo, Haoning Wu, Dongxu Li, Jing Ma, Mohan Kankanhalli, and Junnan Li. VideoAutoArena: An Automated Arena for Evaluating Large Multimodal Models in Video Analysis through User Simulation . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8461–8474, Los Alamitos, CA, USA, June 2025. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.00792. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00792.

[31] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.

[32] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arxiv*, 2024. URL https://arxiv.org/abs/2406.09418.

[33] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

[34] Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15630–15649. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/mindermann22a.html.

[35] Takehiko Ohkawa, Takuma Yagi, Taichi Nishimura, Ryosuke Furuta, Atsushi Hashimoto, Yoshitaka Ushiku, and Yoichi Sato. Exo2EgoDVC: Dense video captioning of egocentric procedural activities using web instructional videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[36] OpenAI. ChatGPT (september, 2025 version). https://chat.openai.com/chat, 2025. Accessed: March 11, 2026.

[37] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, and Yu Qiao. Egovideo: Exploring egocentric foundation model and downstream adaptation, 2024. URL https://arxiv.org/abs/2406.18070.

[38] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

[39] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*, 2024.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[41] Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. Argus: Hallucination and omission evaluation in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20280–20290, October 2025.

[42] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024.

[43] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos, 2018. URL https://arxiv.org/abs/1804.09626.

[44] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[45] Guorui Song, Guocun Wang, Zhe Huang, Jing Lin, Xuefei Zhe, Jian Li, and Haoqian Wang. Towards fine-grained human motion video captioning. MM '25, page 846–855, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3754560. URL https://doi.org/10.1145/3746027.3754560.

[46] Yolo Y. Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey, 2025. URL https://arxiv.org/abs/2312.17432.

[47] Yunlong Tang, Daiki Shimada, Jing Bi, Mingqian Feng, Hang Hua, and Chenliang Xu. Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7293–7301, 2025.

[48] Qwen Team. Qwen3.5-9b. https://huggingface.co/Qwen/Qwen3.5-9B, 2025. Alibaba Cloud, Apache 2.0 License.

[49] Nandan Thakur, Ronak Pradeep, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. Support evaluation for the trec 2024 rag track: Comparing human versus llm judges, 2025. URL https://arxiv.org/abs/2504.15205.

[50] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann Lecun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings - 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 9568–9578, United States, 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.00914. Publisher Copyright: © 2024 IEEE.; 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 ; Conference date: 16-06-2024 Through 22-06-2024.

[51] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. doi: 10.1109/ICCV.2015.510.

[52] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

[53] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[54] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[55] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[56] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

[57] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10714–10726, 2023.

[58] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, 2023. doi: 10.1109/ICCV51070.2023.01100.

[59] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13577–13587, 2021.

[60] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models, 2024.

[61] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.