InstrAct: Towards Action-Centric Understanding in Instructional Videos

Zhuoyi Yang[1], Jiapeng Yu[1], Reuben Tan[2], Boyang Albert Li[3], and Huijuan Xu[1]*

[1] Pennsylvania State University, University Park, PA, USA
{zzy5267,jfy5396,hkx5063}@psu.edu
[2] Microsoft Research
tanreuben@microsoft.com
[3] Nanyang Technological University, Singapore
boyang.li@ntu.edu.sg

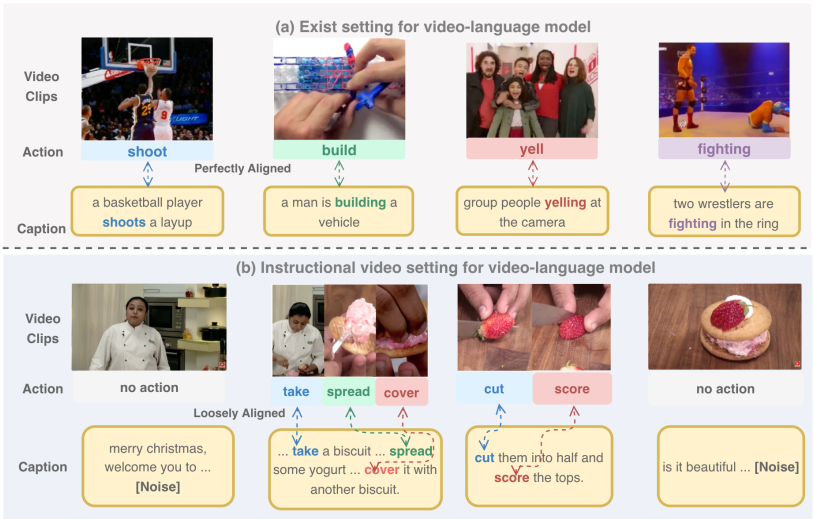


Fig. 1: Video-language modeling paradigms. (a) Existing paradigms use trimmed clips with single atomic actions and perfectly aligned captions. (b) Instructional video setting involves sequential actions, loosely aligned descriptions, and non-instructional noise.

Abstract. Understanding instructional videos requires recognizing fine-grained actions and modeling their temporal relations, which remains challenging for current Video Foundation Models (VFMs). This difficulty stems from noisy web supervision and a pervasive "static bias", where models rely on objects rather than motion cues. To address this, we propose **InstrAct**, a pretraining framework for instructional videos' action-centric representations. We first introduce a data-driven strategy,

Project page: https://zyyangzy.github.io/InstrAct/
* Corresponding author.

which filters noisy captions and generates action-centric hard negatives to disentangle actions from objects during contrastive learning. At the visual feature level, an Action Perceiver extracts motion-relevant tokens from redundant video encodings. Beyond contrastive learning, we introduce two auxiliary objectives: Dynamic Time Warping alignment (DTW-Align) for modeling sequential temporal structure, and Masked Action Modeling (MAM) for strengthening cross-modal grounding. Finally, we introduce the InstrAct Bench to evaluate action-centric understanding, where our method consistently outperforms state-of-the-art VFMs on semantic reasoning, procedural logic, and fine-grained retrieval tasks.

Keywords: Video-language model · Instructional video understanding · Action-centric pretraining

1 Introduction

Early research in video understanding primarily focused on trimmed clips, typically formulated as supervised action recognition on standard benchmarks [14, 15, 27, 28]. With the emergence of large-scale instructional video datasets [22, 30] and the success of cross-modal contrastive learning in image–text models [26], research has increasingly shifted toward video–language pretraining using long, untrimmed instructional videos [2, 21, 29, 35]. Unlike trimmed clips with a single action, instructional videos depict procedures composed of sequential actions. Consequently, understanding instructional videos requires not only recognizing individual actions, but also modeling their temporal organization as coherent procedures.

However, recent studies show that current video–language models exhibit limited understanding of verbs and actions. Instead, these models rely heavily on static visual cues (e.g., objects and backgrounds) and textual cues (e.g., nouns), failing to learn true action semantics [12,25,39]. To mitigate this issue, prior work has explored strategies such as generating action-focused hard negatives [17, 34] and designing verb-aware contrastive objectives [23].

Nevertheless, these approaches are primarily developed for trimmed videos [9, 10, 24], which usually capture one short event with clean captions rather than a temporally ordered sequence of actions. In contrast, instructional videos rely on Automatic Speech Recognition (ASR) transcripts that describe continuous sequences of actions. Such transcripts are often noisy and loosely aligned with visual content [22, 30]. As a result, the supervision signals derived from these transcripts are substantially weaker, making it difficult to disentangle multiple actions and ground the action semantics in videos. These challenges indicate that simply extending trimmed videos' techniques is insufficient for instructional video understanding. To extend action-centric modeling beyond the trimmed-video paradigm, we propose **InstrAct**, a pretraining framework for instructional videos' action-centric representations. Our framework combines a data-driven strategy with an action-centric model design.

On the data side, to address the noisy supervision of ASR transcripts, we develop an LLM-assisted data-curation pipeline that filters non-instructional captions and extracts verb phrases for fine-grained temporal supervision. We further construct action-centric hard negatives (HNs) to disentangle actions from static objects: verb-altered HNs penalize incorrect action identification, while order-swapped HNs enforce procedural temporal reasoning. This curated data provides strong action-centric training signals for contrastive learning.

On the model side, to mitigate the static bias of video encoders, we introduce an Action Perceiver network to distill action-related representations from redundant video features. The module compresses dense video embeddings into a compact set of latent action tokens using a Perceiver-based resampler [1, 13]. To ensure that these tokens capture true action semantics, we further introduce a verb-guided distillation mechanism [18]. Verb phrase embeddings attend to video features to produce teacher tokens that capture action semantics, while the Perceiver's output latents act as students. This guidance encourages the latent space to focus on action-relevant cues, enabling robust action representations even when verb phrases are unavailable during downstream inference.

Beyond global contrastive video–text alignment, we introduce two auxiliary objectives to better model the temporal and cross-modal structure of instructional videos. To capture sequential action structure, we propose the DTW-Align loss, a Dynamic Time Warping objective that models temporal alignment between action tokens and extracted verb phrases, enabling flexible, non-linear correspondence between visual action dynamics and textual descriptions [4]. To further strengthen cross-modal grounding under noisy transcripts, we introduce the Masked Action Modeling (MAM) loss, similar to masked language modeling [6]. MAM masks action tokens in captions and predicts them from both visual and textual contexts, which encourages deep integration between visual motion features and language features.

Contributions. In summary, we propose InstrAct, a comprehensive pretraining framework for instructional videos that integrates action-centric data curation, an action-distilling module, and dedicated self-supervised objectives:

- **Data-Driven Training & Curation:** We develop an LLM-assisted data curation pipeline to filter non-instructional noise and integrate action-centric hard negatives. Based on this curated data, we also introduce the InstrAct Bench to diagnose action-centric understanding.
- **Motion-Distilling Architecture:** We propose the Action Perceiver to aggregate motion-essential cues from redundant video embeddings, further refined by a verb-guided distillation mechanism to ensure the latent space is action-centric.
- **Multi-Action Learning Objectives:** We introduce the DTW-Align loss for flexible action-verb alignment and the MAM loss to strengthen cross-modal action grounding.

2 Related Work

2.1 Action Understanding in Vision–Language Models

Static cues like background scenery and object appearances are well-known shortcut features in action and video understanding [8, 16, 17, 40], and vision-language models are not immune to them. Hendricks and Nematzadeh [12] found that image-text models tend to understand verbs using object occurrence rather than relational and temporal cues. In the video domain, Park et al. [25] showed that even with multiple frames, video-text models rely heavily on static cues rather than action dynamics and interactions. More recently, Varma et al. [31] showed that temporal VLMs can learn error-inducing static feature biases, and introduced TROVE to automatically surface such failures.

To mitigate this issue, action-sensitive datasets are built to provide strong supervision on actions. The Action Dynamics Benchmark [34] created hard negatives by reversing individual actions and entire video clips. VFC [23] generate negative captions with a different verb for each video clip. These methods are designed for short videos that contain single atomic action. In contrast, instructional videos consist of diverse and compositional actions with complex causal and temporal dependencies. To our best knowledge, this is the first paper tackling the static shortcut issue in instructional video understanding.

2.2 Instructional Video Datasets

Instructional videos feature multi-step, compositional activities with strong temporal dependencies. Early datasets such as YouCook2 [42] and CrossTask [43] provide high-quality annotations for procedure segmentation and step localization. EPIC-KITCHENS [5] and ProceL [7] focus on fine-grained actions and hand-object interactions. To improve generalization, larger corpora like COIN [30] and HowTo100M [22] offer diverse domains for large-scale pretraining. However, their automated transcripts often contain noise or off-topic content, leading to weak semantic alignment between text and visual actions.

3 Data Curation

To construct action-centric supervision for instructional video pretraining, we develop an LLM-assisted data curation pipeline to refine the cooking subset of HowTo100M [11, 22]. As illustrated in Figure 2, the pipeline consists of three stages: (1) filtering non-instructional captions, (2) extracting structured verb phrases, and (3) generating action-centric hard negatives. Detailed prompts and implementation details are provided in the released code and Appendix.

Filtering Non-Instructional Captions: Raw subtitles in HowTo100M contain conversational filler and other non-instructional content that lacks visual grounding. To remove such noise, we employ an LLM-based binary classifier that distinguishes instructional steps from conversational text. Only captions describing concrete food-related actions are retained.

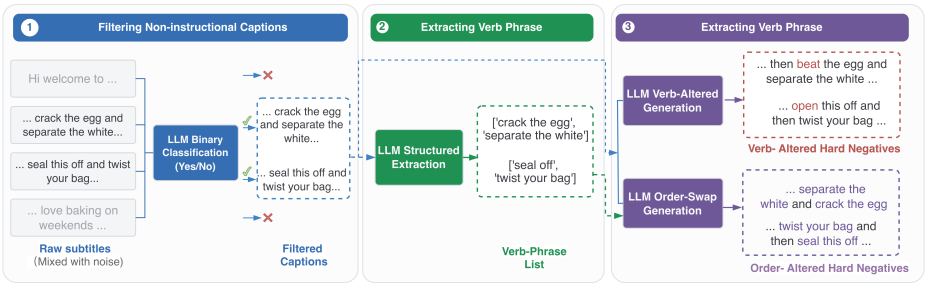


Fig. 2: Overview of the LLM-assisted data curation pipeline. Our data-driven strategy follows three stages: subtitle filtering, structured action-unit parsing, and action-centric hard-negative generation via verb alteration and order swapping.

Verb Phrase Extraction: To obtain structured supervision for temporal action modeling, we employ an LLM to parse captions and extract discrete verb phrases. We design prompts that structure actions into predefined syntactic patterns: "V + N", "V-ing", or "V + Prep + N". For example, a caption such as *"Crack the egg using the side of this ceramic mug and separate the white into the small container."* is converted into a standardized action list: *['crack egg', 'separate white']*. This structured decomposition isolates motion cues from redundant objects and provides supervision anchors for temporal alignment and hard negative generation.

Generating Action-Centric Hard Negatives: To discourage reliance on static visual cues, we generate verb-altered negatives by replacing the main action verb with a semantically different one while keeping the rest of the caption unchanged. This forces the model to distinguish actions based on motion rather than static context. To enforce procedural temporal reasoning, we further generate order-swapped negatives by permuting the chronological order of actions while preserving the original vocabulary, e.g., transforming *"crack the egg and then whisk it"* into *"whisk it and then crack the egg"*. This penalizes bag-of-words shortcuts and encourages the model to capture sequential action dependencies.

4 Action Centric Model

We next introduce the action-centric model architecture of InstrAct. The model builds on a video–text backbone to extract initial cross-modal representations, which are further refined by specialized modules for action understanding.

Specifically, we introduce an Action Perceiver network with verb-guided distillation (Sec. 4.1) to extract action-relevant cues from redundant video features. To model the temporal dependencies of sequential actions, we incorporate a DTW-based Frame–Action Alignment objective (Sec. 4.2). To further strengthen cross-modal grounding, we adopt a Masked Action Modeling objective (Sec. 4.3) that requires the model to recover masked action tokens conditioned on visual evidence.

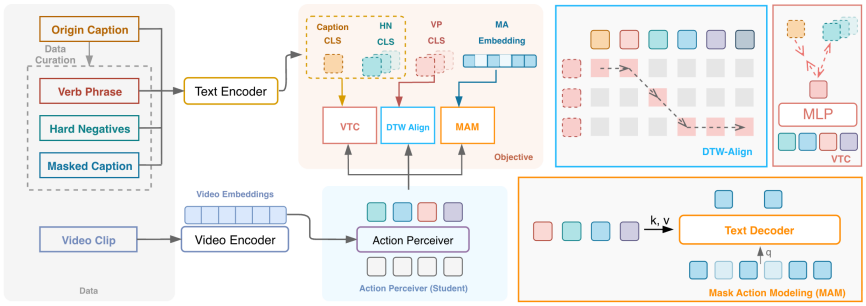


Fig. 3: Overview of the InstrAct framework. Our model builds on a video–text backbone and an Action Perceiver that extracts action-centric representations. The model is trained with three objectives: video–text contrastive learning, DTW-Align for temporal grounding, and Masked Action Modeling (MAM) for cross-modal action understanding.

4.1 Action Perceiver with Knowledge Distillation

To mitigate the static bias of video encoders in instructional videos, we introduce an Action Perceiver network optimized via verb-guided distillation. The module compresses redundant visual features into compact motion-aware Action Tokens.

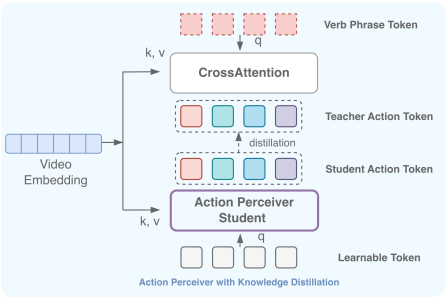


Fig. 4: The Action Perceiver module.

Action Token Extraction We employ a Perceiver [13] network to compress the visual features into a compact set of latent action representations. Formally, the visual embeddings are denoted as $\mathbf{V} \in \mathbb{R}^{T\times N\times C}$, where T, N, and C denote the number of frames, patches per frame, and feature dimension. A set of K learnable latent queries $\mathbf{L} \in \mathbb{R}^{K\times C}$ interacts with the flattened visual features via cross-attention to produce Student Action Tokens $\mathbf{S} \in \mathbb{R}^{K\times C}$:

$$\mathbf{S} = \text{Perceiver}(q = \mathbf{L}, k = \mathbf{V}, v = \mathbf{V}). \tag{1}$$

To preserve temporal structure, we ensure each latent query aggregates localized motion dynamics by restricting each latent query to attend to a local temporal window rather than the entire sequence. This is implemented by injecting temporal embeddings and applying an attention mask to the similarity matrix.

Verb-guided Distillation To guide the latent tokens toward action semantics, we introduce a verb-guided distillation stream acting as a teacher. As illustrated in

Figure 4, extracted verb phrase embeddings $\mathbf{P} \in \mathbb{R}^{M \times C}$ attend to visual features to produce Teacher Action Tokens $\mathbf{T}$:

$$\mathbf{T} = \text{Perceiver}(q = \mathbf{P}, k = \mathbf{V}, v = \mathbf{V}). \tag{2}$$

Since $\mathbf{T}$ is conditioned on explicit action verbs, it provides a semantic reference for motion. The student tokens $\mathbf{S}$ are then optimized to match the teacher through a soft contrastive distillation objective.

Let $\mathbf{s}$ denote similarity scores predicted from the student tokens and $\mathbf{s}'$ the teacher targets. The unidirectional distillation loss is defined as:

$$\mathcal{L}_{\text{single}}(\mathbf{s}, \mathbf{s}') = -(1-\alpha) \sum \mathbf{y} \log(\sigma(\mathbf{s})) - \alpha \sum \sigma(\mathbf{s}') \log(\sigma(\mathbf{s})), \tag{3}$$

where $\sigma(\cdot)$ is the Softmax function and α controls the distillation intensity. The final bidirectional objective sums the losses over both video-to-text ($v2t$) and text-to-video ($t2v$) directions:

$$\mathcal{L}_{distill} = \mathcal{L}_{\text{single}}(\mathbf{s}_{v2t}, \mathbf{s}'_{v2t}) + \mathcal{L}_{\text{single}}(\mathbf{s}_{t2v}, \mathbf{s}'_{t2v}). \tag{4}$$

This distillation guides the latent tokens toward verb-centric motion cues, enabling robust action representations even when verb phrases are unavailable during inference.

4.2 DTW-based Action–Verb Alignment

To model temporal correspondence between visual actions and verb phrases, we design a DTW-based alignment objective that aligns the action token sequence with the extracted verb sequence.

Alignment Formulation Let $\mathbf{S} \in \mathbb{R}^{K \times C}$ denote the sequence of Action Tokens extracted by the Action Perceiver, forming a temporally ordered representation of video dynamics. The curated verb phrases are encoded to a sequence of embeddings $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_M\}$, where $\mathbf{v}_j \in \mathbb{R}^C$.

We compute the alignment cost between $\mathbf{s}_i$ and $\mathbf{v}_j$ using cosine distance

$$d(\mathbf{s}_i, \mathbf{v}_j) = 1 - \frac{\mathbf{s}_i \cdot \mathbf{v}_j}{\|\mathbf{s}_i\| \|\mathbf{v}_j\|}$$

which yields a cost matrix $\mathbf{C} \in \mathbb{R}^{K \times M}$. The goal of DTW is to find the monotonic path through $\mathbf{C}$ with minimal cumulative cost. To enable end-to-end optimization, we adopt the differentiable Soft-DTW [4] and define the alignment loss as

$$\mathcal{L}_{align}(\mathbf{S}, \mathcal{V}) = \text{Soft-DTW}(\mathbf{C}). \tag{5}$$

Order-Aware Regularization Optimizing $\mathcal{L}_{align}$ alone may lead to degenerate alignments. To enforce temporal sensitivity, we introduce a contrastive regularization term inspired by [37]. The intuition is that aligning the correct action order should yield lower cost than aligning a temporally disrupted sequence.

We construct a negative action sequence $\tilde{\mathbf{S}}$ by reversing the token order (i.e., $\tilde{\mathbf{S}} = [\mathbf{s}_K, \ldots, \mathbf{s}_1]$). A hinge loss enforces a margin between the positive and negative alignments:

$$\mathcal{L}_{order} = \max\left(0, \mathcal{L}_{align}(\mathbf{S}, \mathcal{V}) - \mathcal{L}_{align}(\tilde{\mathbf{S}}, \mathcal{V}) + \beta\right). \tag{6}$$

where β is a margin hyperparameter. This objective explicitly penalizes the model if it fails to distinguish the correct procedural flow from a reversed sequence.

Total Objective The final DTW objective is computed over the batch:

$$\mathcal{L}_{dtw} = \frac{1}{B}\sum_{i=1}^{B}\left(\mathcal{L}_{align}^{(i)} + \gamma\mathcal{L}_{order}^{(i)}\right). \tag{7}$$

4.3 Masked Action Modeling

While DTW-Align captures temporal correspondence between actions and verbs, it operates under a dual-encoder contrastive paradigm and provides limited token-level cross-modal interaction. To address this limitation, we introduce the Masked Action Modeling (MAM) loss, which requires the model to reconstruct masked action verbs by attending to visual action tokens $\mathbf{S}$.

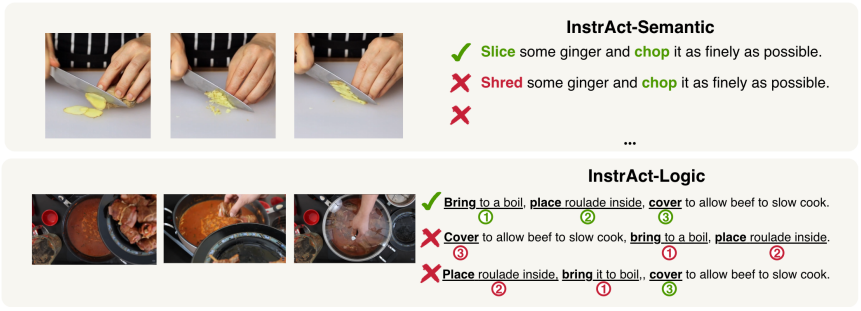
Multimodal Decoding: We employ a multimodal text decoder [38] to predict masked caption tokens. For each layer l, textual hidden states $\mathbf{h}_t^{(l)}$ are first processed by causal self-attention to capture linguistic context, followed by cross-attention that integrates visual action tokens $\mathbf{S}$:

$$\mathbf{h}_t^{(l)} = \text{Cross-Attn}(\text{Self-Attn}(\mathbf{h}_t^{(l-1)}), \mathbf{S}) \tag{8}$$

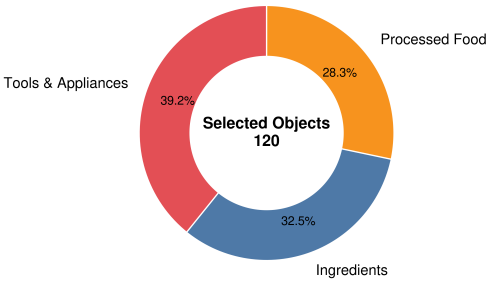
Learning Objective: The decoder is optimized using an auto-regressive objective. Let $\mathbf{h}_t$ denote the textual hidden state at step t. The MAM loss is defined as the cross-entropy between predicted probabilities and ground-truth masked tokens $\mathbf{Y}$ conditioned on the visual action tokens $\mathbf{S}$:

$$\mathcal{L}_{mam} = -\sum_{t=1}^{N-1}\log P(y_{t+1} \mid \mathbf{h}_{1:t}, \mathbf{S}) \tag{9}$$

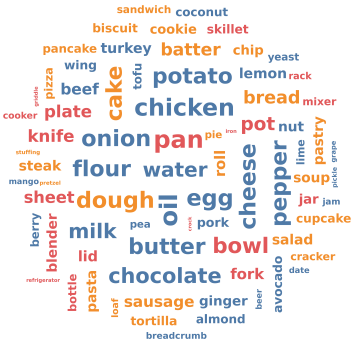
where N is the sequence length. This objective encourages the model to infer missing verbs and action sequences by attending to motion cues, thereby mitigating the ambiguity often found in noisy instructional transcripts.



(a) Example Multiple-Choice Questions (MCQ) for InstrAct-Semantic and InstrAct-Logic.



(b) Category distribution of *InstrAct-Dynamics*.



(c) Word cloud visualization of the evaluated objects.

Fig. 5: Data Visualization. Top: Illustrative examples from our semantic and logic benchmarks. Bottom: Statistical breakdown of the dynamics benchmark, showcasing the diverse object pool and category proportions.

5 InstrAct Bench

To evaluate whether video-language models truly understand instructional procedures rather than relying on static visual shortcuts, we introduce the **InstrAct Bench**, a benchmark constructed from the LLM-refined dataset (Sec. 3). The benchmark consists of three complementary tasks: InstrAct-Semantic for fine-grained verb discrimination, InstrAct-Logic for procedural reasoning, and InstrAct-Dynamics for action-centric cross-modal retrieval. The formulation and evaluation metrics are described below.

5.1 InstrAct-Semantic

InstrAct-Semantic diagnoses whether models rely on static visual shortcuts instead of motion cues. We formulate this task as a 10-choice multiple-choice question (MCQ) task with 2,000 video instances (Fig. 5a, top). Given a video, the candidate answers consist of the ground-truth caption and nine action-altered

hard negatives that preserve identical objects and context (Sec. 3). By controlling object semantics, this task forces models to distinguish fine-grained actions purely through motion dynamics (e.g., whisking eggs vs. cracking eggs). Performance is evaluated using Recall@1 (R@1), Recall@5 (R@5), and Mean Rank (MR).

5.2 InstrAct-Logic

InstrAct-Logic evaluates whether models capture procedural dependencies in multi-step actions. This task is also formulated as a multiple-choice question (MCQ) task with 2,000 video instances. Each example contains the ground-truth caption and 2–3 order-altered negatives that permute the chronological order of actions while preserving the same content (Fig. 5a, bottom). This design isolates temporal reasoning: models that rely on bag-of-actions representations may succeed on InstrAct-Semantic but fail here without modeling correct action order. Performance is evaluated using Accuracy (ACC).

5.3 InstrAct-Dynamics

To complement synthetic evaluation with real-world data, we introduce InstrAct-Dynamics, a fine-grained cross-modal retrieval benchmark consisting of 16,326 video-text pairs grouped into 120 object-centric pools. Within each pool, all videos share the same primary object, forcing models to distinguish subtle procedural variations rather than relying on object recognition.

To capture diverse interaction types, the pools are categorized into Ingredients (raw state changes, e.g., cutting), Processed Food (complex transformations, e.g., baking), and Tools & Appliances (functional manipulation). Retrieval is evaluated in both Video-to-Text and Text-to-Video settings within each pool. We report the clip-count-weighted average Recall@K (K = 1, 5, 10).

6 Experiments

6.1 Experiment Setting

We evaluate a diverse set of video-language models, including early contrastive approaches (e.g., MIL-NCE [21]) and recent Video Foundation Models (e.g., PerceptionLM [3]). Baselines are categorized as In-Domain (ID) or Out-of-Domain (OOD) depending on whether their pretraining data includes cooking instructional videos (e.g., YouCook2 [42]). ID models are evaluated in a zero-shot setting, while OOD models are first fine-tuned on a filtered cooking subset before evaluation. Unless otherwise specified, InstrAct uses the ViCLIP-B branch of InternVideo as the default backbone.

Existing baselines are trained only with matched video–text pairs and do not utilize hard negatives (HNs). Directly evaluating our model using the same HN types seen during training would introduce an unfair advantage. To ensure a

Table 1: Performance on InstrAct-Semantic and InstrAct-Logic. Methods are grouped as In-domain (ID) and Out-of-domain (OOD). ↓ indicates lower is better.

	InstrAct-Semantic			InstrAct-Logic
Method	**R@1 (%)**	**R@5 (%)**	**MR (↓)**	**ACC (%)**
ID Models (Zero-shot)				
MIL-NCE [21]	7.2	49.4	6.0	5.1
UNiVL [19]	13.2	60.3	5.0	33.6
VideoPrism [41]	18.2	74.4	4.0	31.3
PerceptionLM [3]	16.5	63.2	4.0	31.4
ViCLIP [32]	14.7	61.2	4.0	30.0
OOD Models (Finetuned on our Filtered Cooking Subset of 175,690 Clips)				
CLIP-ViP [36]	18.4	63.2	4.0	24.5
CLIP4Clip [20]	14.8	58.0	5.0	23.7
Our Methods				
InstrAct (Verb-Altered HNs)	-	-	-	**40.0**
InstrAct (Order-Swapped HNs)	**28.1**	**78.0**	**3.0**	-

fair comparison, we adopt a cross-evaluation strategy. Specifically, for InstrAct-Semantic we train with Order-Swapped HNs, while for InstrAct-Logic we train with Verb-Altered HNs. For InstrAct-Dynamics, which contains only authentic video–text pairs without synthetic negatives, all HN types are used during training.

To further validate generalizability, we integrate InstrAct with four different VFM backbones and observe consistent improvements. Additional implementation details are provided in the Appendix.

6.2 Main Results

On the InstrAct-Semantic benchmark, VideoPrism [41] and PerceptionLM [3] achieve the strongest performance among ID models, suggesting that large-scale pretraining combined with action-aware modeling improves sensitivity to motion dynamics. In contrast, MIL-NCE [21] performs poorly when confronted with our hard negatives, despite being pretrained on HowTo100M, indicating that simple video–text matching is insufficient for distinguishing fine-grained actions. Among OOD methods, CLIP-ViP slightly outperforms CLIP4Clip, reflecting modest gains from large-scale video–text pretraining. Nevertheless, our proposed method achieves the best performance across all metrics, demonstrating its effectiveness in disentangling actions from static object cues.

On the InstrAct-Logic benchmark, most ID models (except MIL-NCE) achieve similar performance levels. This suggests that scaling training data or adopting specialized architectures alone is insufficient under the prevalent holistic video–text matching paradigm. MIL-NCE performs substantially worse on this task, indicating that simple video–text matching models struggle with order-sensitive action sequences. Since the order-altered negatives share nearly identical vocabularies with the ground-truth captions, distinguishing them requires modeling fine-grained temporal dependencies. OOD methods exhibit a similar

Table 2: Overall performance on the InstrAct-Dynamics benchmark. Weighted average Recall@K across 120 objects (16,326 clips).

Method	R@1			R@5			R@10		
	T2V	V2T	Mean	T2V	V2T	Mean	T2V	V2T	Mean
MIL-NCE [21]	12.63	15.04	13.84	30.77	34.42	32.59	42.10	46.69	44.40
UniVL [19]	13.77	15.91	14.84	32.67	36.80	34.73	43.89	49.14	46.51
PerceptionLM [3]	16.33	23.10	19.72	32.11	42.05	37.08	41.15	51.48	46.32
VideoPrism [41]	27.23	34.13	30.68	48.49	59.08	53.79	58.36	**70.00**	64.18
Ours	**33.86**	**35.19**	**34.53**	**54.73**	**59.52**	**57.13**	**63.13**	69.71	**66.42**

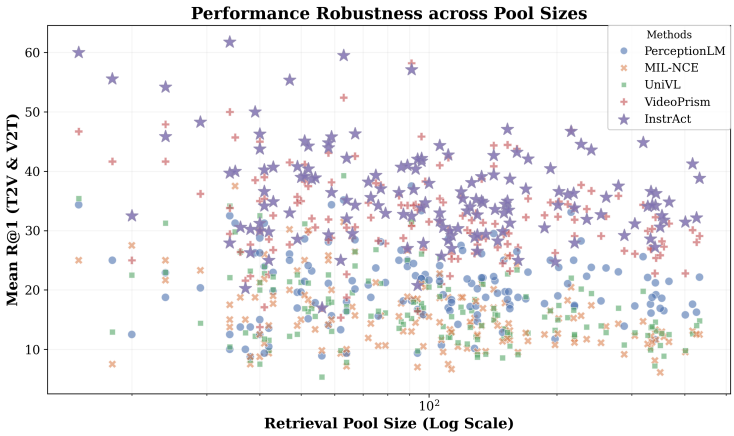


Fig. 6: Performance robustness across retrieval pool sizes on InstrAct-Dynamics.

trend even after domain-specific fine-tuning, further highlighting the limitations of standard matching objectives for modeling procedural logic. In contrast, our method achieves a clear performance improvement, demonstrating its effectiveness in capturing sequential action dependencies.

Table 2 reports performance on authentic videos (detailed breakdowns in the Appendix). Our model achieves the best results on nearly all metrics, improving mean R@1 from 30.68 to 34.53 and mean R@5 from 53.79 to 57.13 over the strongest baseline (VideoPrism). Because each pool contains videos with the same primary object, this benchmark requires models to distinguish subtle procedural variations rather than rely on static object cues. VideoPrism also performs strongly, especially in V2T retrieval, consistent with its motion-aware design. Category-level analysis in the Appendix further shows that our model gains the most on Tools & Appliances scenarios, where fine-grained functional manipulation is critical.

Figure 6 further analyzes retrieval robustness under different pool sizes. As the number of candidates increases, most baselines degrade noticeably due to the growing number of in-domain distractors. In contrast, our model maintains con-

Table 3: Generalization across different Video Foundation Models. We evaluate the performance gains of applying solely our data-driven augmentation (+ HNs only) versus deploying our complete InstrAct framework which incorporates both data- and model-driven approaches (+ InstrAct) across four distinct backbones.

Backbone	Method	InstrAct-Semantic			InstrAct-Logic
		R@1 (%)	R@5 (%)	MR (↓)	ACC (%)
MIL-NCE [21]	Backbone	7.2	49.4	6.0	5.1
	+ HNs only	32.4	82.2	3.0	6.1
	+ InstrAct	32.9	83.3	2.0	4.7
CLIP-ViP [36]	Backbone	18.4	63.2	4.0	24.5
	+ HNs only	38.5	86.7	2.0	30.2
	+ InstrAct	42.3	88.7	2.0	37.1
CLIP4Clip [20]	Backbone	14.8	58.0	5.0	23.7
	+ HNs only	44.7	90.4	2.0	35.6
	+ InstrAct	43.3	88.8	2.0	40.2
ViCLIP [33]	Backbone	14.7	61.2	4.0	30.0
	+ HNs only	40.7	89.4	2.0	37.2
	+ InstrAct	45.1	90.8	2.0	44.7

sistently higher accuracy across pool scales, demonstrating stronger robustness in fine-grained action discrimination.

Table 3 demonstrates the generalizability of InstrAct across four VFM backbones. Applying only our data-driven augmentation (+HNs) consistently improves the base models; for example, CLIP-ViP improves from 18.4 to 38.5 and CLIP4Clip from 14.8 to 44.7 in R@1 on InstrAct-Semantic. Incorporating the full framework (+InstrAct) further brings improvements for most backbones, especially on the more challenging InstrAct-Logic task, indicating that both the data-driven augmentation and the action-centric modeling contribute to stronger action understanding. However, MIL-NCE shows negligible gains and even slight drops on InstrAct-Logic, which aligns with our earlier observations on InstrAct-Dynamics: its relatively simple text encoder struggles to capture order-sensitive dependencies when captions share nearly identical vocabularies.

6.3 Ablation Study

To validate the Action Perceiver, we compare it with a baseline using raw frame embeddings on a subset of videos containing more than three sequential verb phrases. Alignment quality is measured using the normalized Dynamic Time Warping (DTW) cost:

$$\mathcal{C}_{DTW} = \frac{1}{L} \sum_{(i,j)\in P^*} (1 - S_{i,j}) \tag{10}$$

where $S \in \mathbb{R}^{T\times V}$ denotes the similarity matrix between T visual tokens and V verb phrases, and P^* is the alignment path. As shown in Table 4, Perceiver

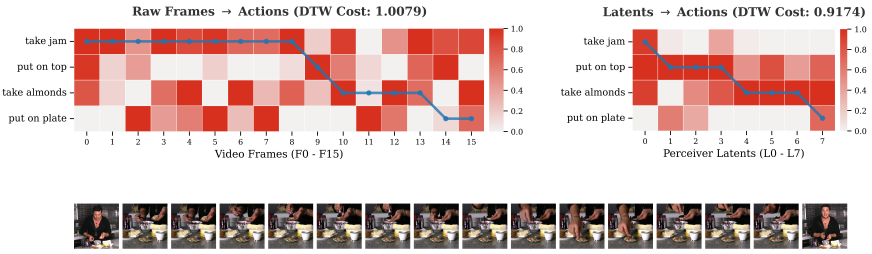


Fig. 7: Qualitative comparison of DTW alignment heatmaps. Left subplots show Baseline (Raw Frames → Actions), and Right subplots show (Latents → Actions).

Table 5: Ablation Study on InstrAct Bench. We ablate hard-negative (HN) construction, the Action Perceiver, and alignment objectives. Verb-altered HNs target action–object disentanglement, while order-swapped HNs target temporal-order reasoning. ↓ indicates lower is better.

	HN Type		Action	Alignment		InstrAct-Semantic			Logic
Setting	**Verb**	**Order**	**Perceiver**	**MAM**	**DTW**	**R@1**	**R@5**	**MR (↓)**	**ACC**
HN ablation									
Verb-altered HNs	✓	–	✓	✓	✓	46.7	91.2	2.0	40.0
Order-swapped HNs	–	✓	✓	✓	✓	28.1	78.0	3.0	44.2
Action Perceiver ablation									
MAM w/o AP	✓	✓	–	✓	–	41.0	89.4	2.0	37.6
DTW w/o AP	✓	✓	–	–	✓	37.3	87.2	2.0	37.4
AP + MAM	✓	✓	✓	✓	–	43.7	90.0	2.0	39.4
AP + DTW	✓	✓	✓	–	✓	38.7	86.5	2.0	40.2
Alignment objective ablation									
AP + MAM	✓	✓	✓	✓	–	43.7	90.0	2.0	39.4
AP + DTW	✓	✓	✓	–	✓	38.7	86.5	2.0	40.2
Full model									
AP + MAM + DTW	✓	✓	✓	✓	✓	**45.1**	**90.8**	**2.0**	**44.7**

latents yield a lower alignment cost than raw embeddings, indicating reduced temporal noise and better alignment with sequential actions.

Qualitatively (Figure 7), the baseline exhibits scattered activations caused by static object bias, often correlating common objects such as "jam" with all frames. In contrast, the Action Perceiver produces a cleaner staircase-like diagonal pattern that isolates sequential steps (e.g., "put on"). These results suggest that the Action Perceiver yields temporally cleaner representations; we thus keep it fixed and ablate the training objectives.

Visual Representation	**Cost ↓**
Raw Embeddings	0.97
Action Latents	**0.91**

Table 4: Quantitative comparison of alignment cost.

Table 5 analyzes the effects of the proposed training objectives. Starting from the baseline trained with hard negatives only, adding MAM improves InstrAct-

Semantic (R@1: 40.7→43.7) by encouraging fine-grained cross-modal reconstruction. In contrast, DTW-Align mainly benefits InstrAct-Logic (ACC: 37.2→40.2) by enforcing order-sensitive temporal alignment, although it slightly degrades Semantic performance where local verb discrimination is more critical. When both objectives are jointly optimized, the model achieves the best overall results (R@1: 45.1, ACC: 44.7), indicating that temporal grounding and masked semantic reconstruction provide complementary supervision for robust action understanding.

7 Conclusion

We address the object shortcut bias in instructional videos by shifting from static matching to action-centric modeling. By combining a data-driven strategy for constructing action-centric supervision with our Action Perceiver and DTW-based alignment, we decouple fine-grained action semantics from redundant background context. Our framework and the new InstrAct Bench demonstrate that while data scale is helpful, explicit architectural priors and high-quality action supervision are essential for understanding how procedural activities structurally unfold.

Acknowledgments

This research was funded by Huck Institutes of the Life Sciences at Penn State University through the 2026 Huck Seed Grant Program. We gratefully acknowledge the support from the National Research Foundation Fellowship (NRFF13-2021-0006), Singapore.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Molyneaux, D., Denil, M., Vinyals, O., Simonyan, K., Zisserman, A.: Flamingo: a visual language model for few-shot learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
3. Cho, J.H., et al.: Perceptionlm: Open-access data and models for detailed visual understanding. arXiv preprint arXiv:2504.13180 (2025)
4. Cuturi, M., Blondel, M.: Soft-dtw: a differentiable loss function for time-series. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. p. 894–903. ICML'17, JMLR.org (2017)
5. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: European Conference on Computer Vision (ECCV) (2018)

6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) 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. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423/
7. Elhamifar, E., Naing, Z.: Unsupervised procedure learning via joint dynamic summarization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
8. Fioresi, J., Dave, I.R., Shah, M.: Albar: Adversarial learning approach to mitigate biases in action recognition. In: The Thirteenth International Conference on Learning Representations (2025)
9. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "Something Something" Video Database for Learning and Evaluating Visual Common Sense . In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 5843–5851. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2017). https://doi.org/10.1109/ICCV.2017.622, https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.622
10. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., 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 (CVPR). pp. 18995–19012 (2022)
11. Han, T., Xie, W., Zisserman, A.: Temporal alignment network for long-term video. In: CVPR (2022)
12. Hendricks, L.A., Nematzadeh, A.: Probing image-language transformers for verb understanding. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 3635–3644. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.318, https://aclanthology.org/2021.findings-acl.318/
13. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 4651–4664. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/jaegle21a.html
14. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, A., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. In: arXiv preprint arXiv:1705.06950 (2017)
15. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: A large video database for human motion recognition. In: 2011 International Conference on Computer Vision. pp. 2556–2563 (2011). https://doi.org/10.1109/ICCV.2011.6126543
16. Lee, J., Lee, W., Park, G.M., Kim, S.T., Choi, J.: Disentangled concepts speak louder than words: Explainable video action recognition. In: NeurIPS (2025)
17. Li, H., Liu, Y., Zhang, H., Li, B.: Mitigating and evaluating static bias of action representations in the background and the foreground. In: International Conference on Computer Vision (ICCV) (2023)

18. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
19. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020)
20. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
21. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In: CVPR (2020)
22. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In: ICCV (2019)
23. Momeni, L., Caron, M., Nagrani, A., Zisserman, A., Schmid, C.: Verbs in action: Improving verb understanding in video-language models. In: International Conference on Computer Vision (ICCV) (2023)
24. Monfort, M., Jin, S., Liu, A., Harwath, D., Feris, R., Glass, J., Oliva, A.: Spoken Moments: Learning Joint Audio-Visual Representations from Video Descriptions . In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14866–14876. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2021). https://doi.org/10.1109/CVPR46437.2021.01463, https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01463
25. Park, J.S., Shen, S., Farhadi, A., Darrell, T., Choi, Y., Rohrbach, A.: Exposing the limits of video-text models through contrast sets. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3574–3586. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.261, https://aclanthology.org/2022.naacl-main.261/
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
27. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR). pp. 32–36 (2004)
28. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. In: arXiv preprint arXiv:1212.0402 (2012)
29. Sun, C., Baradel, F., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: ICCV (2019)
30. Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., Zhou, J.: Coin: A large-scale dataset for comprehensive instructional video analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
31. Varma, M., Delbrouck, J.B., Ostmeier, S., Chaudhari, A.S., Langlotz, C.: TRove: Discovering error-inducing static feature biases in temporal vision-language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=KOwhczyFpg

32. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Chen, X., Wang, Y., Luo, P., Liu, Z., Wang, Y., Wang, L., Qiao, Y.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942 (2023)
33. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., Qiao, Y.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
34. Wang, Z., Blume, A., Li, S., Liu, G., Cho, J., Tang, Z., Bansal, M., Ji, H.: Paxion: Patching action knowledge in video-language foundation models. In: Advances in Neural Information Processing Systems. vol. 36 (2023), https://openreview.net/forum?id=blm1pqiOXe
35. Xu, H., Yang, X., Tian, X., Wang, G., Deng, C., et al.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. In: EMNLP (2021)
36. Xue, H., Sun, Y., Fu, B., et al.: Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. In: The Eleventh International Conference on Learning Representations (ICLR) (2023)
37. Xue, Z., Grauman, K.: Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In: Advances in Neural Information Processing Systems. vol. 36, pp. 25725–25740 (2023)
38. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
39. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=KRLUvxh8uaX
40. Zhang, Y., Bai, Y., Wang, H., Wang, Y., Fu, Y.: Don't judge by the look: Towards motion coherent video representation. In: International Conference on Learning Representations (2024)
41. Zhao, L., Gundavarapu, N.B., Yuan, L., Zhou, H., Yan, S., Sun, J.J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., et al.: Videoprism: A foundational visual encoder for video understanding. In: International Conference on Machine Learning (ICML) (2024)
42. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: AAAI (2018)
43. Zhukov, D., Alayrac, J.B., Cinbis, R.G., Fouhey, D., Laptev, I., Sivic, J.: Cross-task weakly supervised learning from instructional videos. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3532–3540 (2019). https://doi.org/10.1109/CVPR.2019.00365