Open-Vocabulary Fine-Grained Hand Action Detection

Ting Zhe[1], **Mengya Han**[1*], **Xiaoshuai Hao**[2], **Yong Luo**[1*], **Zheng He**[1*], **Xiantao Cai**[1] and **Jing Zhang**[1]
[1]School of Computer Science, National Engineering Research Center for Multimedia Software and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University
[2]Beijing Academy of Artificial Intelligence
{zheting, myhan1996, luoyong, hezheng, caixiantao}@whu.edu.cn, xshao@baai.ac.cn, jingzhang.cv@gmail.com

Abstract

In this work, we address the new challenge of open-vocabulary fine-grained hand action detection, which aims to recognize hand actions from both known and novel categories using textual descriptions. Traditional hand action detection methods are limited to closed-set detection, making it difficult for them to generalize to new, unseen hand action categories. While current open-vocabulary detection (OVD) methods are effective at detecting novel objects, they face challenges with fine-grained action recognition, particularly when data is limited and heterogeneous. This often leads to poor generalization and performance bias between base and novel categories. To address these issues, we propose a novel approach, **Open-FGHA** (**O**pen-vocabulary **F**ine-**G**rained **H**and **A**ction), which learns to distinguish fine-grained features across multiple modalities from limited heterogeneous data. It then identifies optimal matching relationships among these features, enabling accurate open-vocabulary fine-grained hand action detection. Specifically, we introduce three key components: Hierarchical Heterogeneous Low-Rank Adaptation, Bidirectional Selection and Fusion Mechanism, and Cross-Modality Query Generator. These components work in unison to enhance the alignment and fusion of multimodal fine-grained features. Extensive experiments demonstrate that Open-FGHA outperforms existing OVD methods, showing its strong potential for open-vocabulary hand action detection. The source code is available at OV-FGHAD.

1 Introduction

Hand action detection (HAD) focuses on accurately recognizing and localizing a broader range of specific human hand actions, which is crucial for applications in various fields, such as Virtual Reality (VR) [Villegas *et al.*, 2020], Augmented Reality (AR) [Skovsen *et al.*, 2020], Human-Computer Interaction (HCI) [Hu *et al.*, 2022], Design and Control of Robot Hands [Palli *et al.*, 2013], and Healthcare [Ye *et al.*, 2022].

*Corresponding authors: Mengya Han, Yong Luo, Zheng He

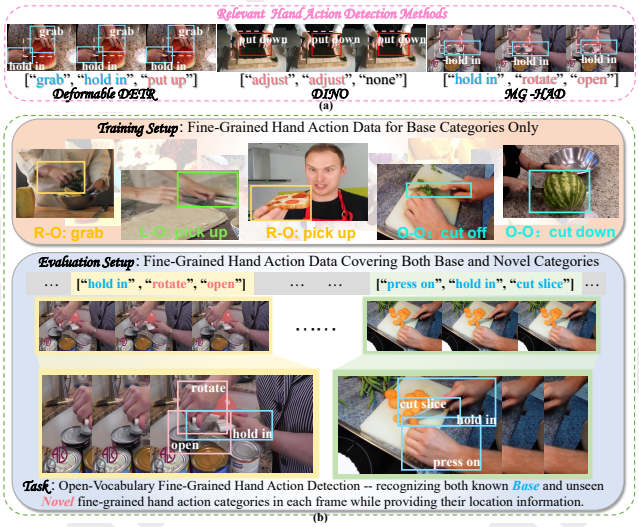

Figure 1: **Overview of the OV-FGHAD.** (a) Detection performance of existing representative fine-grained hand action detection methods. (b) Basic setup of the new task (OV-FGHAD). Training and evaluation configurations of our method Open-FGHA, which primarily follows a generalized setting. Base categories are represented by light blue boxes and text, while novel categories are shown by pink boxes and text. Misclassified categories are highlighted in red. Missed detections for novel categories are highlighted in pink dashed boxes. Each hand action event is denoted as ["*left hand-object interaction* (L-O)", "*right hand-object interaction* (R-O)", "*object-object interaction* (O-O)"], indicating the corresponding fine-grained hand action.

Existing models often struggle to generalize to novel actions not seen during training [Zhe *et al.*, 2024; Zhang *et al.*, 2022; Zhu *et al.*, 2020]. As shown in Figure 1(a), models detect and localize known hand action categories (*e.g.*, *"hold in"*) accurately, but novel actions (*e.g.*, *"put up"*) may be undetected or misclassified as similar categories, such as *"adjust"* → *"put down"*. This limits the exploration of diverse fine-grained actions and hinders the model's generalization capabilities. Therefore, developing methods capable of detecting and localizing unseen categories is essential for advancing this field.

To address this, we propose a new task setting called Open-Vocabulary Fine-Grained Hand Action Detection (OV-FGHAD), which aims to develop HAD models capable of recognizing and localizing base and novel fine-grained hand

action categories. While recent advancements in large-scale vision-language models, such as GDINO [Liu *et al.*, 2024], have shown significant promise in open-vocabulary visual tasks, two key issues hinder their effective application to fine-grained hand action detection.

How can we maintain exceptional performance and balance between base and novel categories when faced with limited and heterogeneous data? A straightforward solution adopted in previous work [Wang *et al.*, 2024] is to freeze the parameters of the vision or language model during the training of vision-language models (VLMs). However, freezing the parameters of the model limits its ability to adapt to new tasks, often resulting in unsatisfactory performance. To overcome this, Parameter Efficient Fine-Tuning (PEFT) methods, such as LoRA [Hu *et al.*, 2021], have been proposed. Nevertheless, while PEFT methods may mitigate the adaptation issue, they still do not effectively resolve the balance issue between base and novel categories in the OV-FGHAD task.

How can we mitigate the impact of the "inter-categories similarity and intra-categories variation" in fine-grained categories for the OV-FGHAD task? In the OV-FGHAD task, inter-category differences are minimal, as observed in both text descriptions (*e.g.*, *"cut off"* vs. *"cut down"*) and visual appearances (*e.g.*, *"grab"* vs. *"pick up"*). In contrast, intra-category variability is significant. For instance, hand actions such as *"pick up brush"* and *"pick up piece-pizza"* exhibit substantial visual differences despite belonging to the same hand action category, as illustrated in Figure 1(b). Current Open-Vocabulary Detection (OVD) models [Liu *et al.*, 2024] primarily rely on aligning and fusing global visual features with textual features. Such approaches often introduce irrelevant representations and increase the risk of incorrect alignment between multimodal fine-grained features, leading to confusion among fine-grained categories. Consequently, they fail to effectively address these challenges.

In this work, we propose a new baseline for the OV-FGHAD task, **Open-FGHA**, based on a vision-language model (VLM). By extracting high-quality multimodal fine-grained features from limited and heterogeneous data and locally enhancing cross-modality features, our method emphasizes well-matched fine-grained features, enabling effective generalization to unseen hand actions while maintaining balance across categories. Specifically, we devise three key components: (1) **Hierarchical Heterogeneous Low-Rank Adaptation (HiH-LoRA)**: To effectively extract multimodal fine-grained features, HiH-LoRA employs low-rank adaptation with distinct parameter settings for each weight matrix within the visual backbone based on their characteristics. It generates heterogeneous low-rank adaptations for each block, enabling the model to learn diverse fine-grained features. By freezing the vision-language model to preserve global knowledge, HiH-LoRA facilitates knowledge exchange at both global and local levels, thereby improving the balance between base and novel categories. (2) **Bidirectional Selection and Fusion (BSF) mechanism**: To identify the optimal matching relationships between multimodal fine-grained features, we propose the BSF mechanism, which includes multimodal bidirectional selective cross-attention (**Bi-SCA**) and multiple fusion processes. By adopting a "multiple-selection multiple-fusion" strategy, the mechanism extracts and refines relevant local cross-modality features from global representations, effectively reducing confusion caused by fine-grained categories. (3) **Cross-modality Query Generator (CQG)**: Building on the high-quality multimodal features produced by the first two components, we propose a novel cross-modality query generator that generates enhanced cross-modality queries for the cross-modality decoder. The CQG consists of two parts: *the content part*, which concatenates the Bi-selective text features from the Bi-SCA module with randomly initialized static content queries to generate dual content queries, and the *positional part*, which generates dynamic anchors from the updated fused text and image features produced by the BSF mechanism using a GDINO-inspired strategy [Liu *et al.*, 2024]. This approach improves content representation and preserves multimodal correlations.

We evaluate the effectiveness of Open-FGHA for fine-grained hand action detection in two task settings: OVD and closed-set Action Detection (AD). The first setting is supported by FHA-Kitchens OVD, a new benchmark established in this study built upon the FHA-Kitchens benchmark [Zhe *et al.*, 2024], while the latter setting is conducted on a subset of the FHA-Kitchens benchmark.

The contributions of this work are summarized as follows:

- We introduce **Open-FGHA**, a simple yet strong baseline approach tailored specifically for the OV-FGHAD task. This method excels at capturing the distinctions and relationships among multimodal fine-grained features from limited and heterogeneous data. As a result, it enables the model to generalize effectively to unseen fine-grained hand actions while maintaining balanced performance across both base and novel categories.
- We propose three novel components to tackle the challenges of fine-grained category confusion and performance imbalance: HiH-LoRA, BSF, and CQG. These components enhance the learning of multimodal fine-grained features and significantly boost the prediction accuracy for unseen fine-grained categories.
- We evaluate the proposed method on both the OV-FGHAD and closed-set FG-HAD tasks. The experimental results demonstrate that Open-FGHA achieves state-of-the-art (SOTA) performance on both the OVD setting and closed-set AD setting. For instance, on the established FHA-Kitchens OVD benchmark, our method improves the AP50 for novel categories by an impressive **+4.17** with the Swin-T backbone, surpassing the performance of SOTA OVD methods.

2 Related Work

2.1 Hand Action Detection

Most existing hand action detection methods [Zhe *et al.*, 2024; Chen *et al.*, 2024] rely on video- or image-based trained detectors [Zhang *et al.*, 2022; Li *et al.*, 2022]. Early action detection methods [Feichtenhofer *et al.*, 2019; Pan *et al.*, 2021; Tang *et al.*, 2020] followed a two-stage pipeline, using separate 2D and 3D backbones for localization and feature extraction. With the advent of transformers [Vaswani *et al.*, 2017], they became a dominant backbone for visual tasks [Zhang *et al.*, 2022; Li *et al.*, 2019; Wang *et al.*, 2023], and recent methods

shifted towards using unified backbones for action detection. For instance, MG-HAD [Zhe *et al.*, 2024], a multi-layer transformer-based model, leveraged multidimensional action queries to predict fine-grained hand actions. Meanwhile, methods such as EVAD [Chen *et al.*, 2023], built on the ViT framework, provided efficient end-to-end solutions for video action detection. Query-based action detectors like WOO [Chen *et al.*, 2021] and TubeR [Zhao *et al.*, 2021] followed the detection frameworks in [Sun *et al.*, 2021; Carion *et al.*, 2020] to predict bounding boxes and action categories. Additionally, STMixer [Wu *et al.*, 2023a], another query-based one-stage detector, adaptively sampled discriminative features. Despite these advances, HAD remained more challenging than human full-body action detection due to limited labeled hand action data and the larger number and finer granularity of hand actions [Zhe *et al.*, 2024]. To address these challenges, we propose a novel method for recognizing and localizing unseen fine-grained hand actions.

2.2 Open-Vocabulary Detection

Open-vocabulary detection is a task in object detection that aims to detect objects from novel categories not seen during training. OVR-CNN [Zareian *et al.*, 2021] first introduced this concept, aligning region features with nouns from image-caption pairs as a baseline solution. OV-DETR [Zang *et al.*, 2022], the first DETR-style open-vocabulary detector, addressed the issue of missing novel category assignments using conditional matching, albeit at the cost of inefficient inference. CORA [Wu *et al.*, 2023b] developed a DETR-based framework that leveraged region prompts and anchor pre-matching to adapt CLIP for open-vocabulary detection. RegionCLIP [Zhong *et al.*, 2022] introduced a two-stage pre-training mechanism to adapt CLIP [Radford *et al.*, 2021] for encoding region features, demonstrating its capability in OVD and zero-shot transfer settings. GLIP [Li *et al.*, 2022] unified object detection and phrase grounding tasks, incorporating a grounded vision-language pretraining model to detect unseen categories. GDINO [Liu *et al.*, 2024], a recent open-set object detector, combined the Transformer-based DINO detector with grounded pretraining to detect arbitrary objects using human-provided category names or referring expressions. However, we found that existing OVD methods still performed suboptimally in the context of fine-grained hand action detection. To address this issue, we propose a new method for open-vocabulary fine-grained hand action detection. By extracting high-quality multimodal fine-grained features, our method effectively learns the distinctions and relationships among these features, allowing the model to generalize to unseen hand action categories while maintaining balanced performance across both base and novel categories.

3 Open-FGHA: A Simple Yet Strong Baseline

3.1 Overview

We propose **Open-FGHA**, a simple yet strong baseline, tailored specifically for open-vocabulary fine-grained hand action detection tasks. The overall framework is illustrated in Figure 2. Open-FGHA follows a dual-encoder-single-decoder architecture, consisting of a text backbone, an image backbone enhanced with HiH-LoRA, and a global feature enhancer used to fuse global image and text features. Additionally, a BSF mechanism is introduced to locally enhance cross-modality fine-grained features, followed by a CQG to initialize cross-modality queries. Finally, the multi-layer cross-modality decoder processes the queries to generate predictions.

To extract high-quality multimodal fine-grained features while maintaining balanced performance, we introduce hierarchical heterogeneous LoRA into the image backbone. Meanwhile, the text backbone extracts rich textual features from the corresponding text descriptions. These image and text features are then processed by a global feature enhancer to generate global cross-modality text and image features. However, due to the similarity among fine-grained categories, relying solely on global multimodal features can lead to misalignment during multimodal feature matching. To address this, we propose a novel BSF mechanism, which locally enhances cross-modality text and image features, reducing the similarity between features and emphasizing the most relevant ones. Furthermore, we introduce the CQG, consisting of two parts: the content part and the positional part. CQG takes the previously extracted high-quality cross-modality features as input to generate enhanced queries. These queries are then processed by the cross-modality decoder, which probes the relevant features from both the visual and textual modalities and updates the queries. Finally, the output queries from the last decoder layer are used to predict bounding boxes and assign the corresponding fine-grained categories to the detected regions.

3.2 HiH-LoRA: Hierarchical Heterogeneous Low-Rank Adaptation

Existing methods for a range of multimodal tasks [Xiao *et al.*, 2024; Tian *et al.*, 2024] have shown that freezing the pre-trained VLMs while integrating LoRA [Hu *et al.*, 2021] effectively preserves its original capabilities, achieving superior performance on new task-specific data. The LoRA method updates two low-rank matrices, A and B, and uses BA as the change to the frozen pre-trained weight matrix W_0 in a linear layer, as described in Eq. (1).

$$h = W_0 x + \Delta W x = W_0 x + BAx, \tag{1}$$

where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, and $r \ll \min(d, k)$, meaning the low-rank r is significantly smaller than the dimensions (d, k) of the original model. During training, W_0 remains frozen, while A and B are trainable.

Existing methods typically adopt a standard LoRA configuration, applying uniform parameter updates across the model without considering the varying influence of pre-trained weight matrices on the new task. This approach limits the model's ability to effectively learn heterogeneous knowledge. Additionally, while existing OVD methods freeze VLMs to retain original knowledge, they often fail to generalize well with limited data, leading to overfitting on base categories and poor performance on novel ones, thereby causing imbalanced results between the two. To overcome these challenges, we propose HiH-LoRA, a hierarchical heterogeneous LoRA fine-tuning paradigm designed for multimodal open-vocabulary visual tasks (see the bottom left of Figure 2). HiH-LoRA applies distinct LoRA settings to specific weight matrices within

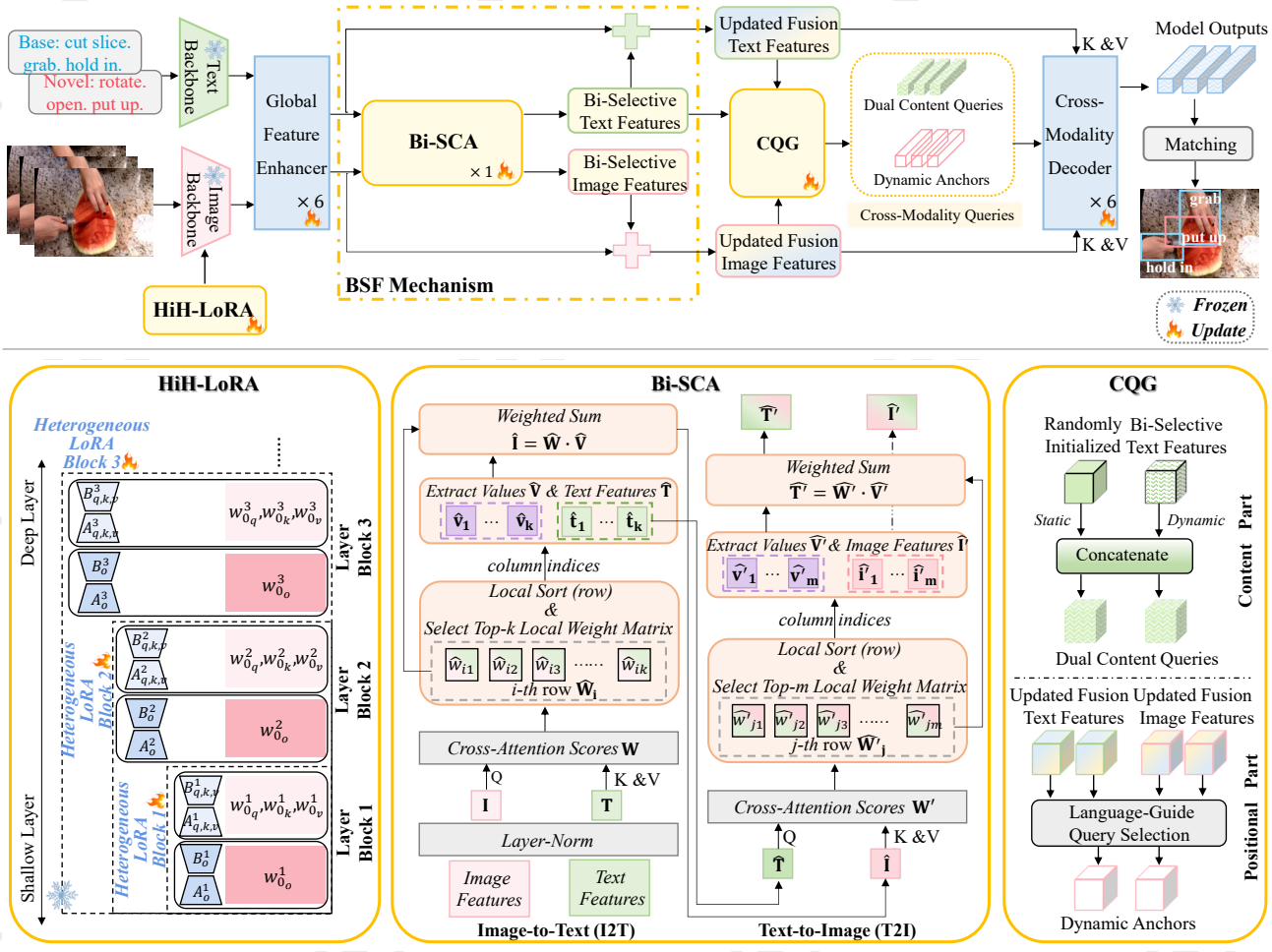

Figure 2: **The overall framework of the proposed Open-FGHA.** Open-FGHA is a simple yet strong open-vocabulary fine-grained hand action detection model. Open-FGHA consists of three novel components: (1) **HiH-LoRA**: A module specifically designed for limited heterogeneous data, enabling effective extraction of multimodal fine-grained features while maintaining model balance (see Section 3.2); (2) **BSF** Mechanism: Comprising a bidirectional selective cross-attention module and multiple fusion processes, this mechanism identifies the optimal matching between multimodal fine-grained features, reducing confusion between fine-grained categories (see Section 3.3); (3) **CQG**: A cross-modality decoder module that utilizes high-quality multimodal fine-grained features to generate enhanced cross-modal queries, further improving content representation and inter-modal correlation (see Section 3.4). The top indicates the whole pipeline, and the bottom describes each module.

each block of the image backbone, focusing on QKV (Query, Key, Value) and O (Output) matrices in the attention layers. Let the image backbone consist of L blocks, and $W^l_{0_i} \in \mathbb{R}^{d\times k}$ represent the i^{th} pre-trained weight matrix in the l^{th} block, where $i \in \{q, k, v, o\}$ and $l \in [1, L]$. For HiH-LoRA, the low-rank matrices at the i^{th} weight matrix in the l^{th} block are denoted as A^l_i and B^l_i. Let h^l and x^l denote the hidden state and input at the l^{th} block, respectively. The forward pass for each hidden state h^l ($l \in [1, L]$) is given by:

$$h^l = \begin{cases} (W^l_{0_q} + S^l_q B^l_q A^l_q)x^l = (W^l_{0_q} + \frac{\alpha^l_q}{r^l_q} B^l_q A^l_q)x^l & i = q, \\ (W^l_{0_k} + S^l_k B^l_k A^l_k)x^l = (W^l_{0_k} + \frac{\alpha^l_k}{r^l_k} B^l_k A^l_k)x^l & i = k, \\ (W^l_{0_v} + S^l_v B^l_v A^l_v)x^l = (W^l_{0_v} + \frac{\alpha^l_v}{r^l_v} B^l_v A^l_v)x^l & i = v, \\ (W^l_{0_o} + S^l_o B^l_o A^l_o)x^l = (W^l_{0_o} + \frac{\alpha^l_o}{r^l_o} B^l_o A^l_o)x^l & i = o, \end{cases} \quad (2)$$

where we scale the $B^l_i A^l_i$ by S^l_i, and $S^l_i = \frac{\alpha^l_i}{r^l_i}$. Here, r^l_i represents the rank of the heterogeneous low-rank matrix corresponding to the i^{th} weight matrix in the l^{th} block, and α^l_i denotes the scaling factor for i^{th} weight matrix in the l^{th} block. For detailed information, please refer to the *Appendix*.

We observe that different pre-trained weights contribute differently to the task. To account for this, we introduce distinct scaling factors, S^l_i, to adjust their importance throughout the model. Specifically, we use a scaling of $S^l = \frac{1}{2}$ for QKV weights and $S^l = \frac{1}{4}$ for O weights, maintaining consistent scaling settings across different backbones.

3.3 BSF: Bidirectional Selection and Fusion

We propose a bidirectional selection and fusion mechanism to address the issue of multimodal feature alignment confusion caused by the low intra-category similarity and high inter-

category similarity among fine-grained categories, leading to increased alignment errors during the process of multimodal fine-grained feature alignment. The BSF mechanism adopts a "multiple-selection multiple-fusion" strategy to extract the most relevant cross-modality text and image fine-grained features from global representations, significantly enhancing local multimodal fine-grained features, expanding the distinction between similar features, and highlighting the most relevant ones, as shown in the first row of Figure 2. The bidirectional selection and fusion mechanism consists of two key subcomponents: the multimodal bidirectional selective cross-attention and the multiple fusion processes.

Bi-SCA. To extract the most relevant text fine-grained features for the given image and the most relevant visual fine-grained features for the given text, Bi-SCA performs two rounds of cross-attention: Image-to-Text (I2T) and Text-to-Image (T2I), as shown in the bottom middle of Figure 2. In I2T process, the enhanced global cross-modality text and image features undergo layer normalization, producing $\mathbf{I} \in \mathbb{R}^{n\times d}$ and $\mathbf{T} \in \mathbb{R}^{s\times d}$, where n represents the number of image tokens, s represents the number of text tokens, and d represents the dimension of these tokens. Subsequently, $\mathbf{I}$ is utilized as $\mathbf{Q}$, and $\mathbf{T}$ is employed as both the Key $\mathbf{K}$ and Value $\mathbf{V}$ to compute the cross-attention scores, resulting in the cross-attention matrix $\mathbf{W} \in \mathbb{R}^{n\times s}$. Then, $\mathbf{W}$ is locally sorted by row, retaining the top-k weights in each row, resulting the top-k local weight matrix $\hat{\mathbf{W}} \in \mathbb{R}^{n\times k}$. For example, for the i-th row, $\hat{\mathbf{W}}_i = [\hat{w}_{i1}, \hat{w}_{i2}, ..., \hat{w}_{ik}]$. Subsequently, according to the column indices corresponding to the top-k weights, $\hat{\mathbf{V}} \in \mathbb{R}^{k\times d}$ and $\hat{\mathbf{T}} \in \mathbb{R}^{k\times d}$ are selected from $\mathbf{V}$ and $\mathbf{T}$, respectively. Finally, the fused visual features $\hat{\mathbf{I}}$ are obtain by calculating the weighted sum of $\hat{\mathbf{V}}$ and $\hat{\mathbf{W}}$. The outputs $\hat{\mathbf{T}}$ and $\hat{\mathbf{I}}$ from the I2T process serve as inputs for the T2I process, where $\hat{\mathbf{T}}$ as the query, and $\hat{\mathbf{I}}$ as the key and value. The subsequent computation in the T2I process follows a similar approach to I2T. The last output of Bi-SCA includes the top-k text features (**Bi-Selective Text features** $\hat{\mathbf{T}}'$) and the top-m image features (**Bi-Selective Image features** $\hat{\mathbf{I}}'$).

Multiple Fusion Processes. To align the most pertinent text and visual features at the local level, we employ the cross-attention mechanism within the Bi-SCA module for local multimodal fusion. Subsequently, to mitigate the challenge of distinguishing similar features within the global cross-modality text and image features, we integrate the bidirectional selective features obtained from Bi-SCA with the global text and image features, while preserving their original extraction positions. This approach effectively reduces the potential misalignment during the process of multimodal feature fusion.

3.4 CQG: Cross-modality Query Generator

We observed that solely relying on feature extraction and alignment fusion is insufficient for effectively handling fine-grained heterogeneous data. To maintain the optimal matching between multimodal features, we introduce a new cross-modality query generator that generates enhanced cross-modality queries for the decoder. By leveraging the previous obtained high-quality cross-modality features, the CQG significantly strengthens the content part of the decoder input, as shown in the first row and bottom right of Figure 2.

Specifically, the cross-modality query generator consists of two components: the content part and the positional part. In the **content part**, we concatenate the Bi-Selective text features generated by the Bi-SCA module with randomly initialized static content queries to produce *dual content queries*. In the **positional part**, we utilize the updated fused text and image features from the BSF mechanism and follow the dynamic anchor generation strategy of GDINO [Liu *et al.*, 2024] to derive positional information. These resulting enhanced cross-modality queries are then used as the input to the decoder. Note that if the content part does not include the Bi-Selective text features from our BSF mechanism, the cross-modality query generation process is identical to that of GDINO.

4 Experiments

4.1 Experiments Settings

Datasets and Metrics. To facilitate fair comparisons with existing open-vocabulary detection methods, we propose the FHA-Kitchens OVD benchmark. Following the convention of the COCO OVD benchmark [Lin *et al.*, 2014], we have restructured the publicly available FHA-Kitchens benchmark [Zhe *et al.*, 2024], focusing on multi-granularity hand actions. The FHA-Kitchens benchmark includes 130 fine-grained action categories, divided into 46 base categories and 15 novel categories. We have re-split the original train and validation sets of the FHA-Kitchens benchmark to create new train and validation sets suitable for the OV-FGHAD task. The model is trained on the 46 base categories, containing 35,351 instances, and evaluated on a validation set containing 9,361 instances, which includes both the 46 base and 15 novel categories. Finally, we evaluate Open-FGHA and other representative detection models in both the open-vocabulary detection setting on the FHA-Kitchens OVD benchmark and the closed-set action detection setting on the hand action subset of the FHA-Kitchens benchmark.

Following previous OVD works [Zhong *et al.*, 2022; Wu *et al.*, 2023b], we evaluate our model under the "generalized" setting, which requires the model to predict objects from both base and novel categories, and then evaluate novel objects. In the **FHA-Kitchens OVD benchmark**, we use AP50 as our primary evaluation metric, which calculates the average precision for each category at an intersection-over-union (IoU) threshold of 50%, and then averages across all categories. In the **FHA-Kitchens subset**, we adopt the mean Average Precision (mAP) as the primary evaluation metric. Additionally, to evaluate the Open-FGHA's ability to balance performance between base and novel categories, we introduce the Harmonic Mean (HM) as a balance metric, defined as:

$$HM = 2 \times \frac{(P_{base} \times P_{novel})}{(P_{base} + P_{novel})}, \tag{3}$$

where P denotes the AP50 for base or novel categories.

Implementation Details. We trained the Open-FGHA model on the FHA-Kitchens OVD benchmark using the MMDetection codebase [Chen *et al.*, 2019]. The image

Method	Image Backbone	Text Backbone	Pre-Training Data	FHA-Kitchens OVD val AP50(%)			HM
				Novel_15	Base_45	All	
mm-Grounding DINO [Zhao *et al.*, 2024]	Swin-T	BERT-B	O365,GoldG,GRIT,V3Det	24.33	51.53	44.80	33.05
	Swin-B		O365,GoldG,V3Det	23.22	50.71	44.00	31.85
	Swin-L		O365V2,OpenImageV6,GoldG	25.16	51.98	45.40	33.91
Grounding DINO [Liu *et al.*, 2024]	Swin-T	BERT-B	O365,GoldG,Cap4M	23.63	51.91	45.00	32.48
	Swin-B		COCO,O365,GoldG,Cap4M, OpenImage,ODinW-35,RefCOCO	24.12	52.27	45.40	33.01
GLIP [Li *et al.*, 2022]	Swin-T	BERT-B	O365,GoldG,CC3M,SBU	24.14	51.45	44.70	32.86
	Swin-L		FourODs,GoldG,CC3M+12M,SBU	23.81	52.00	45.10	32.66
Open-FGHA (Ours)	Swin-T	BERT-B	O365,GoldG,Cap4M	**28.50** ↑ *4.17*	50.11	44.80	**36.33** ↑ *3.28*
	Swin-B		COCO,O365,GoldG,Cap4M, OpenImage,ODinW-35,RefCOCO	**29.16** ↑ *5.04*	52.53	45.60	**37.50** ↑ *4.49*
	Swin-L		O365V2,OpenImageV6,GoldG	**29.68** ↑ *4.52*	52.24	45.80	**37.85** ↑ *3.94*

Table 1: **Comparison with SOTA open-vocabulary fine-grained hand action detection models on the FHA-Kitchens OVD validation set.** It is worth noting that Open-FGHA-T and Open-FGHA-B are fine-tuned from the Grounding DINO-T and Grounding DINO-B pre-trained models, respectively, while Open-FGHA-L is fine-tuned from the mm-Grounding Dino-L pre-trained model. HM: Harmonic Mean score.

backbone was based on the Swin Transformer [Liu *et al.*, 2021], while the text backbone utilized a pre-trained BERT-based uncased model [Devlin *et al.*, 2019]. We fine-tuned three variants of the model for the OV-FGHAD task: **Open-FGHA-T**, **Open-FGHA-B**, and **Open-FGHA-L**, using pre-training data from the GDINO series [Liu *et al.*, 2024; Zhao *et al.*, 2024]. Fine-tuning was performed with the Adam optimizer [Kingma and Ba, 2015], using an initial learning rate of 5×10^{-5} for the tiny variant, 1×10^{-4} for the base and large variants, and weight decay set to 10^{-4}. The experiments were conducted using 4 NVIDIA GeForce RTX 4090 GPUs with the total batch size set to 16 for Open-FGHA-T and Open-FGHA-B, and 4 for Open-FGHA-L. The model was trained for 12 epochs by default. More details on datasets and implementation are available in the *Appendix*.

4.2 Comparisons with SOTA Methods

Results on open-vocabulary fine-grained hand action detection. Table 1 summarizes our main results. Since the pre-trained model is crucial for the open-vocabulary capability of the detector, we compare our method with baseline methods that are trained using the same vision-language models. We compare Open-FGHA with these SOTA OVD methods on the FHA-Kitchens OVD benchmark, utilizing their optimal settings (referencing the MMDetection [Chen *et al.*, 2019] codebase), where mm-GDINO [Zhao *et al.*, 2024] is an enhanced version of GDINO [Liu *et al.*, 2024].

Existing OVD methods in the FG-HAD task show limited performance, particularly in balancing base and novel categories, as well as producing unstable results for novel categories across different image backbones. We adopted the pre-trained models with stable performance under different image backbones and fine-tuned three model variants over 12 epochs. The results in Table 1 indicate the following key insights: **(1)** Our method significantly improves performance on the novel category compared to the SOTA methods using the same backbone, greatly enhancing the model's generalization ability. This improvement is mainly attributed to the design of our BSF mechanism, which effectively addresses the challenges posed by fine-grained categories; **(2)** By utilizing a more powerful image backbone and larger datasets, our method consistently outperforms in the novel category, resulting in a substantial improvement in the harmonic mean score, while also demonstrating robust capability in balancing base and novel categories. This performance is primarily attributed to the design of our HiH-LoRA, which effectively addresses the challenges imposed by limited and heterogeneous data. Note that in OVD [Wu *et al.*, 2023b], performance is primarily assessed based on generalization to novel categories, rather than on the base categories upon which the models are trained.

Method	Epoch	BB	FHA-Kitchens subset val	
			mAP(%)	mAP50(%)
Single-modal detection				
DERT [Carion *et al.*, 2020]	150	R-50	54.10	67.90
DeformDETR [Zhu *et al.*, 2020]	50	R-50	55.80	71.10
DINO [Zhang *et al.*, 2022]	24	R-50	58.70	72.80
MG-HAD [Zhe *et al.*, 2024]	24	R-50	59.20	72.80
Multi-modal detection				
GLIP-T [Li *et al.*, 2022]	12	Swin-T	57.40	72.60
GDINO-T [Liu *et al.*, 2024]	12	Swin-T	58.60	73.40
Open-FGHA-T(Ours)	12	Swin-T	***59.80***	***75.00***
MG-HAD[Zhe *et al.*, 2024]	12	Swin-L	59.90	73.30
Open-FGHA-L(Ours)	12	Swin-L	***61.50***	***75.60***

Table 2: **Comparison with SOTA closed-set fine-grained hand action detection models on the FHA-Kitchens hand action subset.** R-50 is short for ResNet-50, while mAP50 represents the average precision of the model at an IoU threshold of 50%, BB: Backbone.

Results on closed-set fine-grained hand action detection. To comprehensively evaluate the effectiveness of our method in the fine-grained hand action detection task, we compare it against representative single-modal and multimodal detection methods [Zhang *et al.*, 2022; Zhe *et al.*, 2024; Li *et al.*, 2022; Liu *et al.*, 2024] on the closed-set detection task using the FHA-Kitchens hand action subset. We adopt the mAP as the

primary metric to assess the performance of fine-grained hand action detection across different backbone settings. As shown in Table 2, our method demonstrates superior performance in the closed-set fine-grained hand detection task, validating the robust generalization and adaptability of Open-FGHA. Specifically, using the tiny backbone, our method achieves a ***+1.2%*** improvement in AP over the current SOTA multimodal method [Liu *et al.*, 2024]. For the large backbone, our method attains ***61.5%*** AP, surpassing the most advanced single-modal hand action detection method [Zhe *et al.*, 2024].

Visualization. A visualization of the results for open-vocabulary fine-grained hand action detection is presented in Figure 3. Utilizing the Swin-T backbone, qualitative comparison results against the baseline [Liu *et al.*, 2024] on the FHA-Kitchens OVD benchmark are shown. Our model successfully detects all fine-grained hand actions in the current frame. Compared to the baseline, the proposed method demonstrates superior performance on novel fine-grained categories while effectively maintaining balanced performance across base categories. These findings highlight the effectiveness of the proposed model tailored for the OV-FGHAD task.

4.3 Ablation Studies

Effectiveness of Individual Components. Our method incorporates the HiH-LoRA module to address the challenge of extracting effective multimodal fine-grained representations from limited, heterogeneous data while simultaneously maintaining model balance, as discussed in Section 3.2; the BSF mechanism to alleviate confusion in multimodal feature alignment caused by fine-grained categories, as outlined in Section 3.3; and the CQG module, which further enhances cross-modality representations, as detailed in Section 3.4. We conducted an ablation study to assess the effectiveness of each individual module within our method.

By incrementally integrating these modules into the full model, we aim to understand their individual impact on performance under the "generalized" setting. As shown in Table 3, the baseline represents the GDINO-T [Liu *et al.*, 2024] model, and the strong baseline refers to GDINO-T with frozen vision and language model. Although the strong baseline shows some improvement, the proposed Open-FGHA-T with the three novel components substantially enhance performance on novel categories, while also ensuring balanced performance across base and novel categories. Each component contributes meaningfully to improving the baseline, and their combined effect further advances overall performance, highlighting their complementary roles in the OV-FGHAD task.

In the design of the HiH-LoRA and BSF modules, comparative experiments were conducted across various backbone networks to determine the optimal parameter settings. The detailed results are presented in the *Appendix*.

5 Conclusion

In this paper, we introduce a novel task: open-vocabulary fine-grained hand action detection, which aims to detect fine-grained hand actions on both base and novel categories by using fine-grained action textual descriptions. Through a comprehensive evaluation, we observe that existing OVD methods

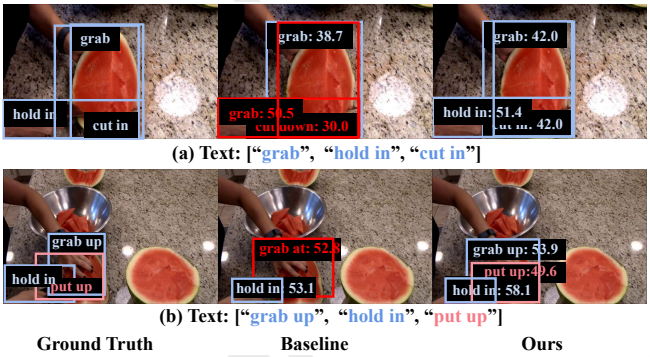

Figure 3: **Qualitative comparison on the FHA-Kitchens OVD benchmark.** (a) Base Category: Our model outperforms the baseline, effectively distinguishing similar actions within the base categories. (b) Base and Novel Categories: Our model successfully detects and localizes unseen fine-grained categories while maintaining strong performance on the base categories. This highlights the effectiveness of the three components designed for open-vocabulary fine-grained hand action detection. Base categories are marked with light blue boxes and text, while novel categories are indicated with pink boxes and text. Misclassified categories are highlighted in red.

Model	Algorithm Components			Generalied_AP50(%)			HM
	HiH-LoRA	*BSF*	*CQG*	Novel_15	Base_45	All	
Baseline				23.63	51.91	45.00	32.48
StrBase				24.09	50.69	44.10	32.66
	✓			24.89 ↑ *1.26*	51.45	44.90	33.55
		✓		25.74 ↑ *2.11*	50.71	44.60	34.15
	✓	✓		26.31 ↑ *2.68*	51.34	45.20	34.79
Ours-T	✓	✓	✓	***28.50*** ↑ *4.87*	50.11	44.80	***36.33***

Table 3: **Ablation study of the key components in Open-FGHA-T.** StrBase: Strong Baseline, HiH-LoRA: Hierarchical Heterogeneous Low-Rank Adaptation, BSF: Bidirectional Selection and Fusion. CQG: Cross-modality Query Generator, HM: Harmonic Mean score.

exhibit a bias toward base categories, while struggling to generalize to novel categories. To address this, we propose a novel method, **Open-FGHA**, which integrates three novel components: HiH-LoRA, BSF, and CQG. By learning the distinctions and relationships among multimodal fine-grained features from limited heterogeneous data, Open-FGHA effectively generalizes to previously unseen fine-grained hand actions while maintaining balanced performance across base and novel categories. It outperforms existing fine-grained hand action detection methods and could serve as a valuable baseline for future research in OV-FGHAD.

6 Limitations

The Open-FGHA model effectively overcomes significant challenges faced by existing open-vocabulary detection models for fine-grained hand actions and reveals insightful observations. However, the sources of hand actions could be expanded to encompass more diverse scenarios. In the future, we hope to leverage richer descriptions of fine-grained hand actions to enhance the comprehension of the Open-FGHA model.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. U23A20318 and 62276195 and 62401407), the Science and Technology Major Project of Hubei Province (Grant No. 2024BAB046), the Yunnan provincial major science and technology special plan projects under Grant 202403AA080002, the Postdoctoral Fellowship Program of CPSF under Grant Number GZB20240571 and the China Postdoctoral Science Foundation under Grant Number 2024M762481. The numerical calculations in this paper were done using the supercomputing system at the Supercomputing Center of Wuhan University.

References

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[Chen *et al.*, 2019] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[Chen *et al.*, 2021] Shoufa Chen, Peize Sun, Enze Xie, Chongjian Ge, Jiannan Wu, Lan Ma, Jiajun Shen, and Ping Luo. Watch only once: An end-to-end video action detection framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8158–8167, 2021.

[Chen *et al.*, 2023] Lei Chen, Zhan Tong, Yibing Song, Gangshan Wu, and Limin Wang. Efficient video action detection with token dropout and context refinement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10388–10399, 2023.

[Chen *et al.*, 2024] Brian Chen, Nina Shvetsova, Andrew Rouditchenko, Daniel Kondermann, Samuel Thomas, Shih-Fu Chang, Rogerio Feris, James Glass, and Hilde Kuehne. What when and where? self-supervised spatio-temporal grounding in untrimmed multi-action videos from narrated instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18419–18429, 2024.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

[Feichtenhofer *et al.*, 2019] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6202–6211, 2019.

[Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[Hu *et al.*, 2022] Hezhen Hu, Weilun Wang, Wengang Zhou, and Houqiang Li. Hand-object interaction image generation. *Advances in Neural Information Processing Systems*, 35:23805–23817, 2022.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

[Li *et al.*, 2019] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8928–8937, 2019.

[Li *et al.*, 2022] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2024] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55, 2024.

[Palli *et al.*, 2013] Gianluca Palli, Salvatore Pirozzi, Ciro Natale, Giuseppe De Maria, and Claudio Melchiorri. Mechatronic design of innovative robot hands: Integration and control issues. In *2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, pages 1755–1760. IEEE, 2013.

[Pan *et al.*, 2021] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 464–474, 2021.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Skovsen *et al.*, 2020] Søren K Skovsen, Harald Haraldsson, Abe Davis, Henrik Karstoft, and Serge Belongie. Decoupled localization and sensing with hmd-based ar for interactive scene acquisition. In *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 167–171. IEEE, 2020.

[Sun *et al.*, 2021] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14454–14463, 2021.

[Tang *et al.*, 2020] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *Proceedings of the European Conference on Computer Vision: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 71–87. Springer, 2020.

[Tian *et al.*, 2024] Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *arXiv preprint arXiv:2404.19245*, 2024.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[Villegas *et al.*, 2020] Alvaro Villegas, Pablo Perez, Redouane Kachach, Francisco Pereira, and Ester Gonzalez-Sosa. Realistic training in vr using physical manipulation. In *2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*, pages 109–118. IEEE, 2020.

[Wang *et al.*, 2023] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.

[Wang *et al.*, 2024] Bin Wang, Chunyu Xie, Dawei Leng, and Yuhui Yin. Iaa: Inner-adaptor architecture empowers frozen large language model with multimodal capabilities. *arXiv preprint arXiv:2408.12902*, 2024.

[Wu *et al.*, 2023a] Tao Wu, Mengqi Cao, Ziteng Gao, Gangshan Wu, and Limin Wang. Stmixer: A one-stage sparse action detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14720–14729, 2023.

[Wu *et al.*, 2023b] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7031–7040, 2023.

[Xiao *et al.*, 2024] Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. Hivg: Hierarchical multimodal fine-grained modulation for visual grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5460–5469, 2024.

[Ye *et al.*, 2022] Ruolin Ye, Wenqiang Xu, Haoyuan Fu, Rajat Kumar Jenamani, Vy Nguyen, Cewu Lu, Katherine Dimitropoulou, and Tapomayukh Bhattacharjee. Rcare world: A human-centric simulation world for caregiving robots. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 33–40. IEEE, 2022.

[Zang *et al.*, 2022] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022.

[Zareian *et al.*, 2021] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.

[Zhang *et al.*, 2022] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[Zhao *et al.*, 2021] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Shuai Bing, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, et al. Tuber: Tubelet transformer for video action detection. *arXiv preprint arXiv:2104.00969*, 2021.

[Zhao *et al.*, 2024] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024.

[Zhe *et al.*, 2024] Ting Zhe, Jing Zhang, Yongqian Li, Yong Luo, Han Hu, and Dacheng Tao. Multi-granularity hand action detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5604–5613, 2024.

[Zhong *et al.*, 2022] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022.

[Zhu *et al.*, 2020] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.