Open Set Video HOI detection from Action-centric Chain-of-Look Prompting

Nan Xi[1] Jingjing Meng[2] Junsong Yuan[1]

[1]State University of New York at Buffalo [2]Amazon

{nanxi, jsyuan@buffalo.edu}, jingjing.meng1@gmail.com

Abstract

Human-Object Interaction (HOI) detection is essential for understanding and modeling real-world events. Existing works on HOI detection mainly focus on static images and a closed setting, where all HOI classes are provided in the training set. In comparison, detecting HOIs in videos in open set scenarios is more challenging. First, under open set circumstances, HOI detectors are expected to hold strong generalizability to recognize unseen HOIs not included in the training data. Second, accurately capturing temporal contextual information from videos is difficult, but it is crucial for detecting temporal-related actions such as `open`, `close`, `pull`, `push`. *To this end, we propose* ***ACoLP****, a model of* ***A****ction-centric* ***C****hain-****o****f-****L****ook* ***P****rompting for open set video HOI detection. ACoLP regards* ***actions*** *as the carrier of semantics in videos, which captures the essential semantic information across frames. To make the model generalizable on unseen classes, inspired by the chain-of-thought prompting in natural language processing, we introduce the* ***chain-of-look*** *prompting scheme that decomposes prompt generation from large-scale vision-language model into a series of intermediate visual reasoning steps. Consequently, our model captures complex visual reasoning processes underlying the HOI events in videos, providing essential guidance for detecting unseen classes. Extensive experiments on two video HOI datasets, VidHOI and CAD120, demonstrate that ACoLP achieves competitive performance compared with the state-of-the-art methods in the conventional closed setting, and outperforms existing methods by a large margin in the open set setting. Our code is available at* `https://github.com/southnx/ACoLP`.

1. Introduction

Human-object interaction (HOI) detection generates a set of meaningful <*person, interaction, object*> triplets for a given scene. It has attracted much attention in recent years, due to its importance for scene understanding, and applications in healthcare, autonomous driving, etc. [47, 42, 15, 6, 12, 16, 24]. Most current works investigate HOI detection in static images in a closed setting[45, 15, 6, 29, 32], where all HOI classes are predefined with examples provided in the training set. However, real-world scenarios are often open-set, where HOIs with novel actions are not present in the training set. Detecting open set HOIs in videos is challenging due to the following reasons. First, static images barely contain temporal information of human and objects, thus directly transferring HOI detection methods for static images to videos fails to model the temporal dynamics. Second, the possi-

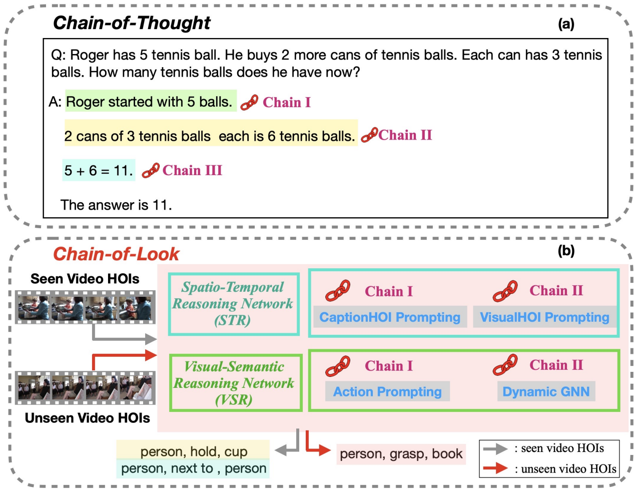


Figure 1. **Illustration on chain-of-thought prompting and chain-of-look prompting.** (a) Chain-of-thought prompting divides a problem into a series of intermediate reasoning steps, enabling large language models on symbolic reasoning tasks. (b) We propose chain-of-look prompting scheme to captures complex visual reasoning processes with two reasoning networks from large-scale vision-language models. Each reasoning network consists of two visual reasoning chains. The learned visual reasoning processes can be expanded to unseen classes with strong generalization ability.

ble HOIs occurred in videos can hardly be exhausted since the compositional space resulting from the <*person, interaction, object*> triplets is tremendous. Moreover, semantic ambiguity of HOIs widely exists in videos, especially for temporal-related interactions. For example, <*person, hold, cup*> and <*person, lift, cup*> contain similar and overlapping semantic information, but it is challenging to distinguish these interdependent HOIs. To this end, we need to build a robust HOI detector for videos with strong generalizability that is capable of handling unseen and ambiguous HOI classes in the open set setting.

Several existing works have investigated HOI detection in videos. A number of methods are based on spatio-temporal graph, including structured RNN [13], LIGHTEN [41], STIGPN [46] and weakly-supervised video HOI detection [21]. Another line of works model the inherent properties of video HOI to help HOI detection. For instance, ASSIGN [28] models asynchronous and sparse properties of HOIs in videos to detect the structure of interaction events in a video scene. 2G-GCN [31] extends ASSIGN by considering geometric features while modeling human and object dependencies. Recently, Transformer-based methods [43] have been proposed for video HOI detection by structurizing a video into a few tubelet tokens. Although these works have made significant progresses in detecting HOIs in videos, they largely focus on the *nouns*, i.e., human and objects, in the video, and infer interactions based on human and objects features. This strategy may result in the loss of valuable information on the *verbs*, i.e., actions, inherent in videos. In addition, previous efforts assume that all testing HOI classes are known in training, leading to unsatisfactory results and poor generalizability when applied to open set circumstances.

To address the challenges mentioned above, we propose **ACoLP** in this work, which is an open set HOI detection model for videos. The key idea of ACoLP model is to abstract each frame into action prompts and model the prompt generating processes as a series of intermediate visual reasoning steps. The motivation underlying this idea is that the *verbs* (***actions***) convey central information of the events happening in a video. In light of this, videos can be regarded as a sequence of actions. Modeling the temporal dynamics of those action sequences captures the core semantic information of events in videos. Meanwhile, to empower the model with strong generalization ability for video HOI detection, ACoLP adopts the chain-of-thought prompting [48] strategy in natural language processing (NLP) to "prompt" the model with input-output visual reasoning steps, which we call the ***chain-of-look***. Those visual reasoning steps, equipped with the few/zero-shot learning ability from large-scale visual-language (VL) models, confer the capability of reasoning visual events on unseen classes. As shown in Figure 1, we introduce visual-semantic reasoning network (VSR) and spatio-temporal reasoning network (STR) for HOI prompting and action prompting, respectively. Each network contains two steps of chains of reasoning for prompt generation, which are akin to the chain-of-thought prompting in NLP. Specifically, VSR includes CaptionHOI Prompting (CHP) and VisualHOI Prompting (VHP). CHP is designed to incorporate global semantic information into individual HOI prompts, serving as the first reasoning chain. VHP follows CHP as the other reasoning chain with visual information for HOI prompting. Similarly, STR also contains two reasoning chains: Action Prompting (AP) and Dynamic GNN (D-GNN). AP is introduced to abstract visual information of each frame into a fixed number of action prompt representations. D-GNN is then employed to model the temporal dynamics across frames, which also benefits open set HOI detection by propagating semantic information to neighboring frames, thus enabling action prompt representations to be more semantic-aware and discriminative in open set settings. Finally, HOI classification and bounding box regression are conducted under the guidance of action prompt representations, which is less noisy than only utilizing HOI prompts. With this open set HOI detection model, we can better infer unseen HOIs in videos.

We validate the effectiveness of ACoLP model via comprehensive experiments on two video HOI detection benchmark datasets: VidHOI [4] and CAD-120 [18]. Comparing with current state-of-the-art (SOTA) methods for both video HOI detection and image HOI detection, our model outperforms these SOTA methods in the open set setting and achieves comparable results in the closed setting.

Our main contributions on the open set HOI detection in videos are summarized as follows:

- We present an action-centric video HOI detection model, which focuses on modeling the HOIs via the *verbs* (actions) instead of *nouns* (humans and objects). It helps more reliably capture the most central semantic information in understanding video HOIs.
- We introduce the chain-of-look prompting scheme to capture the underlying visual reasoning processes in videos, and generate visual-semantic aware and spatio-temporal aware prompts from VL models. Visual reasoning processes are further expanded to unseen classes for better generalization ability.
- Our model achieves substantial improvements in terms of open set video HOI detection and is on par with or better than SOTA methods in the closed setting.

2. Related Work

Human-Object Interactions in Videos. Early works on video HOI detection took spatio-temporal context of videos into consideration for HOI modeling [18, 9] by utilizing

Markov Random Filed (MRF). Follow-up studies extended MRF to Conditional Random Field (CRF) by incorporating features from frame-level nodes [19, 40]. Further attempts incorporate spatio-temporal graphs into Recurrent Neural Networks (RNN) to model high-level structures in videos [13]. Graph Parsing Neural Network (GPNN) [30] was designed to adaptively capture the spatial graph structure. Recent progress on video HOI detection further advanced the spatial temporal graph at multiple granularities [46, 41] and with explicit temporal information [4]. ASSIGN [28] proposed to model the asynchronous and sparse properties of HOIs in videos, which was further extended into 2G-GCN [31] by combing geometric feature and visual feature. TUTOR [43] is introduced recently to abstract a video into several tubelet tokens and further progressively emerge and represent high-level visual semantics with Transformer.

Open Set Recognition. Open set recognition (OSR) tackles the incomplete knowledge of the world during training, aiming to correctly recognize both seen and unseen classes during testing. Existing OSR methods can be classified as discriminative models (DM) and generative models (GM). For DM, traditional methods employ Support Vector Machines (SVM) [38, 14], Sparse Representation [49, 36] or Nearest Neighbor [37, 27]. Further studies extended SVM by adding another constraint on positive samples [3, 2]. Besides, probabilistic open set SVM (POS-SVM) classifier [39] was proposed to determine unique rejection threshold. OpenMax [1] model was among one of the pioneer solutions towards open set Deep Networks by replacing the SoftMax layer with an OpenMax layer. OpenMax effectively handled fooling open set images but failed to recognize visully similar adversarial images. To solve this problem, a neural network based representation learning scheme [10] was introduced for open set recognition.

Large Scale Visual-Language (VL) Models. Recent pretrained large-scale VL models with a representative work of CLIP [33] bridge visual and language information by jointly learning two encoders. Follow-up studies employing the pretrained VL models on downstream tasks have achieved remarkable progress, including CLIP-Adapter [7] and PointCLIP [53]. However, how to apply VL models efficiently on downstream video tasks is an open problem, due to the expensive computation cost of replacing image-text pretraining to video-text pretraining proposed in VideoCLIP [51]. We tackle this challenge by directly employing pretrained VL models for video HOI detection, without expensive video-text pretraining.

Prompt Learning. Prompt learning was first introduced in NLP area [25, 8], aiming to produce a task-specific template for language models. Common prompt learning scheme involves hard prompt learning [7] and soft prompt learning [22]. Hard prompt learning searches for a specific word for the predesigned template, such as "`I [MASK] running.`" in sentiment analysis, where the mask placeholder will be replaced with either "love" or "hate". Different from hard prompt learning, soft prompt learning is designed to tune masked tokens into learnable vectors. We employ the idea of soft prompting, proposing chain-of-look prompting modules for actions and HOIs in video HOI detection.

3. Open Set HOI Detection in Videos

3.1. Problem Formulation

Given a video $V \in \mathcal{V}$ containing T frames $\{I_1, \cdots, I_T\}$, HOI detection model aims to recognize the interactions among the N_e entities (N_h humans and $(N_e - N_h)$ objects, $N_h < N_e$) in the video. Namely, by taking inputs the frames $\{I_t\}_{t=1,\cdots,T}$ and labels $\{g_t\}_{t=1,\cdots,T}$ of entities in the frames, the model outputs the triplet <*person, interaction, object*> in each frame. The output of an HOI is represented by four components: $[b^{(h)}, b^{(o)}, c^{(o)}, a]$, which indicate person bounding box, object bounding box, object class and interaction class, respectively. For person $m \in \{1, \cdots, N_h\}$ and object $l \in \{1, \cdots, N_e - N_h\}$, the HOI detection model detects pair-wise human-object interactions $\{r_{t,c}\}_{t=1}^{T}, c \in \{0,1\}^C$, indicating the existence or not of the interaction class c, where C is the number of all possible HOI classes $\{h_i\}_{i=1}^{C}$. In open set settings, only a portion of HOIs containing M_a actions are utilized for training, while the remaining HOIs containing $M - M_a$ actions are not seen during training (where M indicates the number of all possible actions and $M > M_a$). Note that the open set settings in our work requires all possible actions and HOIs are predefined. It is different from the open world scenarios where we do not know the maximum possible number of classes we may encounter during inference.

3.2. Visual-Semantic Reasoning (VSR) Network

In this section, we delve into VSR network, which is designed as a two-step reasoning process for HOI prompt representation generation from pretrained large-scale VL model (CLIP [33]) to video HOI detection task. Concretely, the first chain-of-look reasoning process CHP employs global semantic information of each frame for HOI prompt representation generation, while the follow-up reasoning process VHP employs visual information from frames. These two complementary reasoning steps enable the final HOI prompt representations to be visual-semantic aware of the events happening in videos.

For a given dataset, there are N possible HOIs in the form of <*person, interaction, object*> . The template t_i="A person is `[interaction]`ing the `[object]`." is pre-defined, where "`[interaction]`" and "`[object]`" are replaced with their corresponding class names in the triplets. Then each template is applied with a pretrained

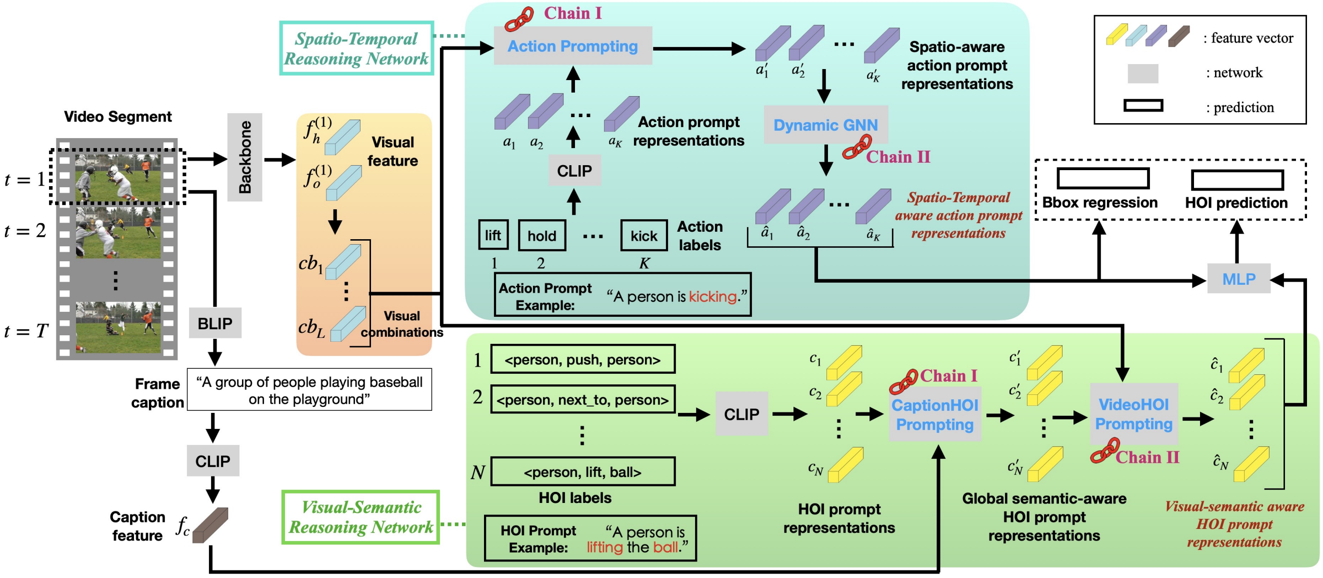

Spatio-Temporal Reasoning Network
Chain I
Action Prompting
a'_1 a'_2 ... a'_K
Spatio-aware action prompt representations
: feature vector
: network
: prediction
Video Segment
$t=1$
$t=2$
$t=T$
Backbone
$f_h^{(1)}$
$f_o^{(1)}$
Visual feature
cb_1
cb_L
Visual combinations
a_1 a_2 ... a_K
Action prompt representations
Dynamic GNN
Chain II
CLIP
lift
hold
kick
1 2 K
Action labels
Action Prompt Example: "A person is kicking."
$\hat{a}_1$ $\hat{a}_2$ ... $\hat{a}_K$
Spatio-Temporal aware action prompt representations
Bbox regression
HOI prediction
MLP
BLIP
Frame caption
"A group of people playing baseball on the playground"
CLIP
Caption feature f_c
1 <person, push, person>
2 <person, next_to, person>
N <person, lift, ball>
HOI labels
HOI Prompt Example: "A person is lifting the ball."
CLIP
c_1 c_2 ... c_N
HOI prompt representations
Chain I
CaptionHOI Prompting
c'_1 c'_2 ... c'_N
Global semantic-aware HOI prompt representations
VideoHOI Prompting
Chain II
$\hat{c}_1$ $\hat{c}_2$ ... $\hat{c}_N$
Visual-semantic aware HOI prompt representations
Visual-Semantic Reasoning Network

Figure 2. **Overview of the ACoLP model.** The model is constructed with the spatio-temporal reasoning (STR) network and the visual-temporal reasoning (VSR) network. Each of the two networks consists of two chain-of-look reasoning steps (STR: Action Prompting + Dynamic GNN; VSR: CaptionHOI Prompting + VideoHOI Prompting). Only the networks highlighted in blue are optimized during training. BLIP model produces image captions of each frame. CLIP model produces caption features and initial action prompt representations or HOI prompt representations. Details on the model are in Section 3.2 and Section 3.3.

large-scale VL model (noted as CLIP shown in Fig. 2) to generate original HOI prompts $c_i = \mathtt{CLIP}(t_i) \in \mathbb{R}^d$, $i \in \{1, \cdots, N\}$, where T_i is the i-th HOI triplet and d denotes the feature dimension. For each frame in the video clip, in order to incorporate global semantic information, the caption of that frame is produced from image caption model BLIP [20]. The generated image caption is further applied with CLIP [33] text encoder to produce caption feature $f_c = \mathtt{CLIP}(\mathtt{BLIP}(I_i)) \in \mathbb{R}^d$, which will be utilized in the following sections.

CaptionHOI Prompting. The **first Chain-of-Look promptig** in VSR network is CHP. CHP module takes original HOI prompt representations c_i and frame caption feature f_c as inputs, thus incorporating global semantic information of the entire image into HOI features. The generated HOI prompt representation $c'_i \in \mathbb{R}^d$ is thus semantic-aware to the video frame, making it applicable for video HOI detection. Concretely, CHP consists of a multi-head attention (MHA), taking input both HOI prompt representation c_i and caption feature f_c at time t. In the MHA module, the query is HOI prompt representation c_i, while the key and value are both caption feature f_c. The output is further applied with a feed-forward network (FFN) to learn video-specific prompts c'_i,

$$\bar{c}_i = \mathtt{MHA}(c_i, f_c) + c_i, \tag{1}$$

$$c'_i = \mathtt{FFN}(\bar{c}_i) + \bar{c}_i. \tag{2}$$

VideoHOI Prompting. The above CHP module generates semantic-aware HOI prompt representations by employing global semantic information from image caption. VHP, which is the **second Chain-of-Look prompting** in VSR network, on the other hand, extends the chain-of-look reasoning process by incorporating visual information to further enhance HOI prompt representations for video HOI detection task. The structure of VHP is the same as CHP, with only the inputs changed to be the updated HOI features c'_i and visual feature $f_v \in \mathbb{R}^d$.

As shown in Fig. 2, for a given frame I_t at time t, pre-trained object detection model FasterRCNN [34] is utilized as the backbone to detect human and object instances in each frame. We select top Q instances based on the prediction scores generated from FasterRCNN. Each human instance is generated with its normalized bounding box $\hat{b}_i^{(h)} \in [0,1]^4, i \in \{1, \cdots, Q_h\}$, where Q_h is the number of human instances. Each object instance is represented with its normalized bounding box $\hat{b}_i^{(o)} \in [0,1]^4, i \in \{1, \cdots, Q_o\}$ and object category e_i, where Q_o is the number of object instances. If the number of detected instances are less than Q, we take all the predicted instances. The instance feature map $f_h^{(t)} \in \mathbb{R}^d$, $f_o^{(t)} \in \mathbb{R}^d$ (h represents human, o represents object) are then generated with ROIAlign [11]. The maximum number of potential combinations of human-object and human-human is $N_c = Q_h Q_o + \binom{Q_h}{2}$. The feature of each combination $f_{comb}^{(n)} \in \mathbb{R}^{2d} (n \in \{1, ..., N_c\})$ is generated by concatenating $f_h^{(t)}$ and $f_o^{(t)}$. The visual feature f_v of frame I_t is taken by averaging all the combination

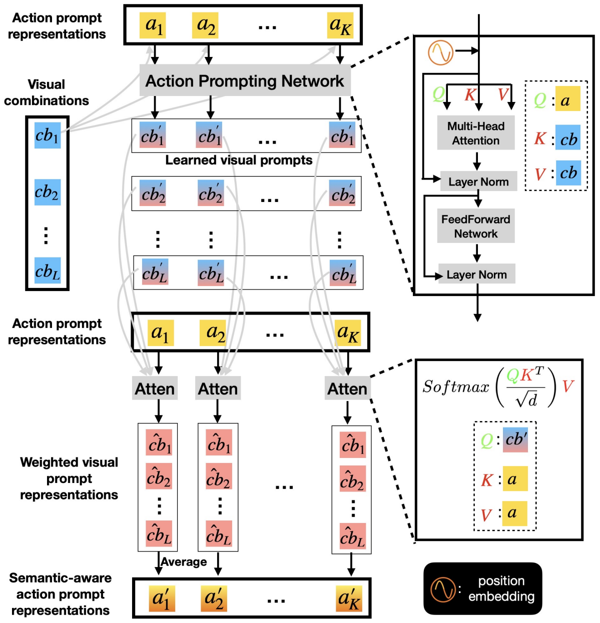


Figure 3. **Action Prompting.** The structure of action prompting module.

features, followed by a linear layer (Proj) to project feature dimension from $2d$ to d: $f_v = \texttt{Proj}(\frac{1}{N_c}\sum_{n=1}^{N_c} f_{comb}^{(n)})$. Similar to Eq. 1 and Eq. 2, the visual-semantic aware (VSA) HOI prompt representation outputted from VHP is formulated as

$$\bar{c}'_i = \texttt{MHA}(c'_i, f_v) + c'_i, \tag{3}$$

$$\hat{c}_i = \texttt{FFN}(\bar{c}'_i) + \bar{c}'_i. \tag{4}$$

3.3. Spatio-temporal Reasoning (STR) Network

The essential semantic information in a video is determined by the **actions** happened in that video, while human and objects serve as participants to accomplish actions. Therefore, modeling temporal dynamics of actions in videos provides fundamental semantic information. To this end, we structurize the visual feature combinations generated from Sec. 3.2 into a fixed number of action features, where each action feature represents a specific action.

For all the K action labels $\{u_k\}_{k=1}^{K}$ in the dataset, the template $t(u_k)$="A person is [action]ing ..." is predefined for each action, where [action] represents each action name. Then we generate action prompt representations $\{a_k\}_{k=1}^{K}$ of each action with CLIP [33] text encoder: $a_k = \texttt{CLIP}(t(u_k))$. To endow the action prompt representations with spatial-aware reasoning capabilities, we design a novel Action Prompting (AP) module as the **first Chain-of-Look prompting** shown in Fig. 3. AP takes visual combinations cb_l and action prompt representation a_k as inputs and outputs semantic-aware action prompt representation a'_k. The AP module is divided into two stages: (I) In the first stage, to align the action prompt representations and visual combinations to the same embedding space, each combination cb_l is applied with an AP module against all the K action prompt representations $\{a_k\}_{k=1}^{K}$. The AP module first takes the positional embedding of cb_l as input, where the position of cb_l is determined by the average center coordinates of human and object (or human and human). Then a Multi-Head Attention (MHA) is applied, where action prompt representation a_k serves as query, while visual combination cb_l serves as key and value. MHA is followed by a Layer Norm (LN) module, Feed-Forward Network (FFN) and another LN. In this way, the generated learned visual prompt $cb'_l = \texttt{AP}(a_k, cb_l)$ is aligned to the same embedding space with respect to each action prompt representation a_k. (II) The second stage of AP module employs a lightweight attention module ($\texttt{Atten}$) to compute the relative importance of all the learned visual prompt representations $\{cb'_l\}_{l=1}^{L}$ to a specific action prompt representation a_k. This attention module takes the learned visual prompt representation cb'_l as query, while a_k as key and value, outputting the weighted visual prompt representation $\hat{cb}_l = \texttt{Atten}(cb'_l, a_k)$. By averaging all the L weighted visual prompt representations $\{\hat{cb}_l\}_{l=1}^{L}$, we generate the semantic-aware action prompt representation $a'_k = \frac{1}{L}\sum_{l=1}^{L}\hat{cb}_l$, which is essentially the weighted sum of all information from visual combinations. Thus the combined formulation of a'_k can be expressed as:

$$a'_k = \frac{1}{L}\sum_{l=1}^{L}\texttt{Atten}(\texttt{AP}(a_k, cb_l), a_k). \tag{5}$$

Our next goal is to enhance spatial-aware action prompt representations into spatio-temporal aware action prompt representations by virtue of the **second Chain-of-Look prompting** module. Now that we have semantic-aware action prompt representations $\{a'_k\}_{k=1}^{K}$ for each frame, we construct a fully-connected graph $\mathcal{G}$ whose nodes are $\{a'_k\}_{k=1}^{K}$. Motivated by the ROLAND model [52] of dynamic GNN, we capture the temporal dynamics of semantic-aware action prompt representations by recurrently updating node features over time. To this end, we design the second Chain-of-Look prompting module with dynamic GNN as shown in Fig. 4. At time t, dynamic GNN takes into $a'_{(t)}$, followed by GNN Layer 1 to generate updated level 1 node state $H_t^{(1)}$:

$$H_t^{(l)} = \texttt{Update}^{(l)}(H_{t-1}^{(l)}, \tilde{H}_t^{(l)}), \tag{6}$$

$$\tilde{H}_t^{(l)} = \texttt{GNN}^{(l)}(H_t^{(l-1)}), \tag{7}$$

where $l = \{1, 2\}$ indicates the number of GNN layer. Then node embedding update is employed by taking $\tilde{H}_t^{(l)}$ and historical node state $H_{t-1}^{(l)}$. Following ROLAND [52], we take

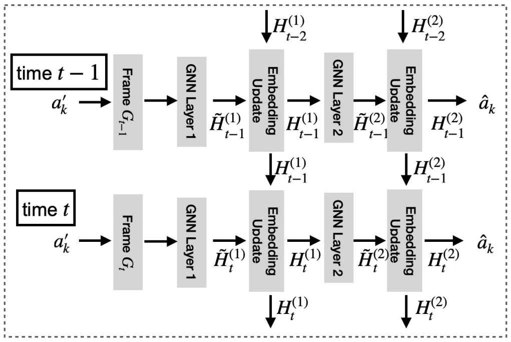


Figure 4. **Dynamic GNN architecture.** The structure of dynamic GNN for temporal modeling of action features across frames.

GRU (Gated Recurrent Unit) cell [5] for node embedding updating:

$$H_t^{(l)} = \texttt{GRU}(H_{t-1}^{(l)}, \tilde{H}_t^{(l)}). \quad (8)$$

With generated $H_t^{(l)}$, the other stacked GNN Layer and Embedding Update layer is applied to generate final node embedding H_t^L, where $L = 2$ in our architecture. With the second Chain-of-Look prompting module of dynamic GNN, we produce spatio-temporal aware (STA) action prompt representations $\{\hat{a}_i\}_{i=1}^K$.

3.4. HOI Prediction & Bounding Box Regression

HOI prediction is performed by employing the above computed STR action prompt representations $\{\hat{a}_k\}_{k=1}^K$ and VSA HOI prompt representations $\{\hat{c}_n\}_{n=1}^N$. For each $\hat{a}_k \in \mathbb{R}^d$ of frame I_t, we concatenate it with $\hat{c}_n \in \mathbb{R}^d$ that contains the same action class with it. Then we apply a multilayer perceptron (MLP) followed by a sigmoid function to generate predicted logits $p_n^{hoi} \in [0, 1]$ of the HOI relating to $\hat{c}_n$: $p_n^{hoi} = \texttt{Sigmoid}(\texttt{MLP}([\hat{a}_k, \hat{c}_n]))$, where $[\cdot, \cdot]$ indicates concatenating operation. Thus the HOI prediction loss $\mathcal{L}_{hoi}$ for each frame can be generated by computing the binary cross-entropy (BCE) between total HOI prediction logits $\{p_n^{hoi}\}_{i=1}^N$ and HOI ground truth $\{y_n^{hoi}\}_{i=1}^N$, where $y_n = \{0, 1\}$:

$$\mathcal{L}_{hoi} = \frac{1}{N}\sum_{n=1}^{N} \texttt{BCE}(p_n^{hoi}, y_n^{hoi}). \quad (9)$$

To compute the loss of bounding box regression for frame I_t, the STR action prompt representations $\{\hat{a}_g\}_{g=1}^G$ whose action classes occur in the ground-truth of that frame are first selected, where G indicates the number of actions in the ground truth of that frame. For each $\hat{a}_g \in \mathbb{R}^d$, an MLP followed by a sigmoid function is applied on $\hat{a}_g$ and every single visual combination $cb_l \in \mathbb{R}^d$ to compute the possibility $p_g^{bbox} \in [0, 1]$ of the action accompanied by that visual combination: $p_g^{bbox} = \texttt{Sigmoid}(\texttt{MLP}([\hat{a}_g, cb_l]))$. At the same time, we learn a threshold $th_k \in [0, 1]$ for each action, selecting those visual combinations cb_l whose p_g^{bbox} are no less than th_k as the predicted human-object or human-human pairs with respect to the action. In open set settings, for actions $\{a_i\}_{i=1}^{K_{test}}$ only in testing set (K_{test} is the number of actions only in testing set), these actions will not be selected and their thresholds will not be optimized during training. To solve this problem, the thresholds of $\{a_i\}_{i=1}^{K_{test}}$ are computed from weighted sum of the thresholds of all the actions in training set. The weights are the similarities between the STR of every two actions. For the G actions in ground truth, we compute BCE loss of predicted bounding boxes for each action:

$$\mathcal{L}_{bbox\_cls} = \frac{1}{G}\sum_{g=1}^{G} \texttt{BCE}(p_g^{bbox}, y_g^{bbox}), \quad (10)$$

where $p_g^{bbox} \in \{0, 1\}^{N_{bbox}}$ and $y_g^{bbox} \in \{0, 1\}^{N_{bbox}}$ are binary representations of predicted and ground-truth bounding box pairs, $N_{bbox} = max\{N_{pred}, N_{gt}\}$, N_{pred} indicates the number of predicted combinations of taht action, N_{gt} indicates the number of ground-truth combinations of that action. If $N_{pred} \neq N_{gt}$, we choose the larger value as the binary vector dimension and pad the shorter binary vector with zeros.

In terms of bounding box locations, each visual combination cb_s^{gt} ($s \in [1, N_{gt}]$) in ground-truth needs to find a corresponding visual combination cb_v^{pred} ($v \in [1, N_{bbox}]$) in prediction to compute loss. Therefore, we calculate the intersection size of human bounding boxes between each cb_s^{gt} against all $\{cb_v^{\text{pred}}\}_{v=1}^{N_{bbox}}$ and assign the cb_v^{pred} with largest intersection size to cb_s^{gt} as a prediction to ground-truth match. Thus, bounding box localization loss between a ground-truth visual combination cb_i^{gt} ($i \in [1, N_{gt}]$) and its corresponding predicted visual combination $cb_{\omega_{(i)}}^{\text{pred}}$ is

$$\mathcal{L}_{bbox\_loc} = \frac{1}{N_{\text{GT}}}\sum_{i=1}^{N_{\text{GT}}}\{\mu_1[||\hat{b}_i^{(h)} - \hat{b}_{\omega_{(i)}}^{(h)}|| + ||\hat{b}_i^{(o)} - \hat{b}_{\omega_{(i)}}^{(o)}||] - \mu_2[\text{GIoU}(\hat{b}_i^{(h)}, \hat{b}_{\omega_{(i)}}^{(h)}) + \text{GIoU}(\hat{b}_i^{(o)}, \hat{b}_{\omega_{(i)}}^{(o)})]\}, \quad (11)$$

where N_{GT} is the total number of ground-truth visual combinations in that frame; GIoU is the generalized IoU [35]; μ_1 and μ_2 are the hyper-parameters for adjusting the weights.

The overall loss for a frame to be minimized in the training phase is

$$\mathcal{L} = \alpha_1\mathcal{L}_{hoi} + \alpha_2\mathcal{L}_{bbox\_cls} + \alpha_3\mathcal{L}_{bbox\_loc} \quad (12)$$

where α_1, α_2 and α_3 are weights for adjusting different loss components.

3.5. Inference

During inference, for each frame in a given video segment, we generate HOIs consisting of four components:

<person bounding box, object bounding box, object class, interaction class >. The object class and interaction class are predicted from HOI logits p_n^{hoi} in Eq. 9, with threshold 0.5 to determine whether the HOI exists or not. Bounding boxes of human and object pairs are selected from visual combinations cb_L with the learned thresholds $\{th_k\}_{k=1}^{K}$ for each action to determine the existence or not of a visual combination.

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate our ACoLP model on two video Human-Object Interaction datasets: VidHOI [4] and CAD120 [18]. VidHOI is a large-scale dataset for detecting video HOIs. Following ST-HOI[4], we take 6,366 videos for training and 756 videos for validation. In VidHOI, there are 50 annotated relation categories and half of them are temporal-related actions. CAD-120 is a relatively smaller dataset, consisiting of 120 RGB-D videos of 4 subjects ad 10 different activities. In this work, only the RGB images and bounding box annotations are utilized.

For VidHOI dataset, we follow ST-HOI [4] to use mean average precision (mAP) as the metric for VidHOI dataset, which is also the same as the widely used image HOI detection dataset HICO-DET [50]. An HOI prediction is considered to be true positive if it satisfies the following two criteria: (1) both the object class and the interaction class is the same with ground truth; (2) both the bounding boxes of human and object overlap with the ground truth boxes with interest-over-union (IoU) more than 0.5. We follow standard scheme of sub-activity F1 score for CAD-120 dataset evaluation.

4.2. Implementation Details

Human and object bounding boxes in video frames are extracted from parameter frozen backbone model with a FasterRCNN [34] pretrained on MS-COCO [23]. We set Q in Sec. 3.2 as 20 to select top score entities. Instance features are then extracted by ROIAlign [11], followed by a linear layer to project instance features to 1024-dim. Node feature dimension d in Dynamic GNN module is 1024. GNN layers in Dynamic GNN is implemented as Graph Convolutional Networks (GCN) [17]. MLPs in the model consists of 3 layers, with ReLU activation function and LayerNorm at the end of each layer except the last one. We employ the pretrained CLIP [33] model as text encoder for extracting text features of 768-dim, which are further projected to 1024-dim. The parameters of CLIP are frozen during training. The number of heads in MHA module in Sec. 3.3 is set to be 8. α_1, α_2 and α_3 in Eq. 12 are set to be 2, 1.5 and 1, respectively. AdamW [26] optimizer is used for training 100 epochs on 4 GPUs with a batch size of 128. The initial learning rate is 0.0001 and decays by 0.9 every 20 epochs.

Method	VidHOI	CAD120
THID, *CVPR 2022* [47]	19.05	87.6
STIGPN, *MM 2021* [46]	-	91.9
ST-HOI, *ICDAR 2021* [4]	17.60	-
ASSIGN *CVPR 2021* [28]	21.43	89.9
2G-GCN, *ECCV 2022* [31]	-	89.5
TUTOR, *NurIPS 2022* [43]	26.92	94.7
ACoLP (Δ *VHP*)	19.45	86.1
ACoLP (Δ *CHP*)	22.06	88.6
ACoLP (Δ *AP*)	21.94	87.7
ACoLP (Δ *D-GNN*)	20.67	85.4
ACoLP, *Ours*	**28.27**	**94.9**

Table 1. HOI detection results compared with SOTA methods on VideoHOI and CAD120 datasets. Δ *VHP*, Δ *CHP*, Δ *AP* and Δ *D-GNN* indicate removing VideoHOI Prompting module, CaptionHOI Prompting module, Action Prompting and Dynamic GNN, respetively. '-' indicates no results are reported in the original paper. THID method and ASSIGN method performance on VidHOI dataset are evaluated by the authors of this submission.

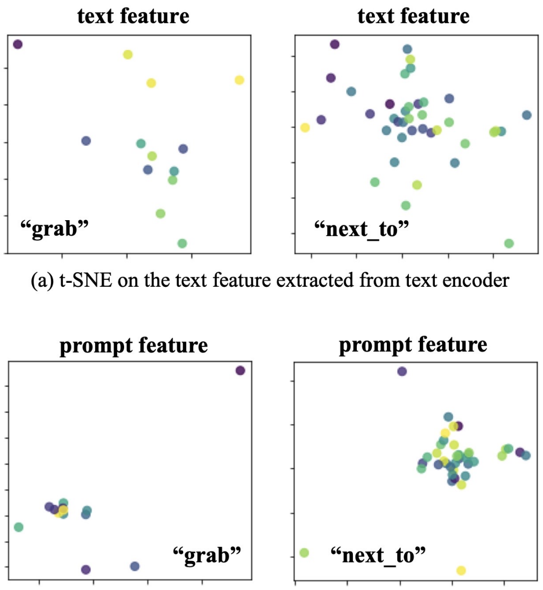
Method	20% unseen	50% unseen
THID, *CVPR 2022* [47]	15.86	10.49
ST-HOI, *ICDAR 2021* [4]	12.36	8.65
ASSIGN *CVPR 2021* [28]	14.63	10.98
2G-GCN, *ECCV 2022* [31]	14.23	10.16
ACoLP (Δ *VHP*)	14.38	9.67
ACoLP (Δ *CHP*)	17.94	11.57
ACoLP (Δ *AP*)	16.53	10.32
ACoLP (Δ *D-GNN*)	15.73	10.46
ACoLP, *Ours*	**19.23**	**12.78**

Table 2. Open set HOI detection performance comparisons on different ratios of unseen HOIs on VidHOI dataset. Notations are the same as those in Table. 1.

4.3. Ablation Studies

In this section, we conduct ablation studies to validate the function of each module.

We first remove CHP module and utilize HOI prompt representation c_i in Fig. 2 directly for HOI prediction. Results in Table 1 show that removing CHP module leads to more than 5% mAP drop on VidHOI dataset and more than 6 F1 points drop on CAD-120 dataset. If removing VHP module and only employ the global semantic-aware HOI prompt representation c_i', the results drops even further on both VidHOI dataset and CAD-120 dataset. In Fig. 5, we visualize the HOI prompt representation c_i and visual-semantic aware HOI prompt representation $\hat{c}_i$ with t-SNE [44]. We select HOIs that contain interaction classes of either `grab` or `next_to`. There are 13 `grab`-related HOIs and 40 `next_to`-related HOIs. Fig. 5 indicates that the visual-semantic aware HOI prompt representations (prompt



(a) t-SNE on the text feature extracted from text encoder

(b) t-SNE on the prompt feature learned by VideoHOI Prompting module

Figure 5. Visualization of original HOI prompt representations (text feature) and visual-semantic aware HOI prompt representations (prompt feature). Each dot represents an HOI that contains the action listed in the sub figure. Each dot is also assigned a different color for better visualization.

feature) exhibit cluster patterns, while original HOI prompt representations (text feature) tend to scatter without a cluster center. This indicates that HOI classes containing the same interaction class share more similar semantic information, especially for the temporal-related interaction class such as `grab`. Futhermore, we evaluate the function of AP module and D-GNN in spatio-temporal reasoning network shown in Fig. 2. Results in Table 1 indicate that removing AP and D-GNN will both harm HOI detection results by a large margin, demonstrate the effectiveness of AP and D-GNN.

4.4. Comparison with State-Of-The-Art (SOTA)

Open Set Video HOI detection Open set video HOI detection indicates we only have access to part of action classes in training set, while there are some actions in test set that we have never seen before. We select training data that contain 80% or 50% of the total action classes and the rest 20% or 50% action classes only exist in test set. Results in Table 3 indicate that our method leads a large margin on unseen HOI detection compared with current SOTA methods with both 20% unseen and 50% unseen. For the video-based methods of ST-HOI [4], ASSIGN [28] and 2G-GCN [31], the performance is worse than image-based method THID [47]. This is caused by the fact that THID [47] incorporates language semantic information into image HOI detection, thus increase the model generalizability on unseen classes. If we delete the VHP module in our model, the ability for open set HOI detection is greatly reduced more than 5% mAP as shown in Table 3. This result suggests that enhancing text information with visual content is essential for generating feasible prompt representations that generalize well on unseen classes.

Method	20% unseen	50% unseen
THID, *CVPR 2022* [47]	82.41	73.95
ST-HOI, *ICDAR 2021* [4]	80.20	73.62
ASSIGN *CVPR 2021* [28]	82.13	72.28
2G-GCN, *ECCV 2022* [31]	81.47	73.82
ACoLP (Δ *VHP*)	82.35	74.56
ACoLP (Δ *CHP*)	82.95	74.10
ACoLP (Δ *AP*)	80.65	72.76
ACoLP (Δ *D-GNN*)	81.32	72.97
ACoLP, *Ours*	**87.53**	**76.39**

Table 3. Open set HOI detection performance comparison on different ratios of unseen HOIs on CAD-120 dataset. Notations are the same as those in Table. 1.

Method	T	S
ST-HOI, *ICDAR 2021* [4]	14.4	25.0
ASSIGN *CVPR 2021* [28]	18.37	29.94
TUTOR, *NurIPS 2022* [43]	21.28	32.21
ACoLP, *Ours*	**24.77**	**32.96**

Table 4. HOI detection performance comparison on temporal-related (T) and static spatial-related (S) HOIs on VidHOI dataset.

	VidHOI		
Method	Full	None-rare	Rare
ST-HOI, *ICDAR 2021* [4]	17.6	27.2	17.3
ASSIGN *CVPR 2021* [28]	20.43	28.5	18.9
TUTOR, *NurIPS 2022* [43]	26.92	37.12	23.49
ACoLP, *Ours*	**28.27**	**39.66**	**26.63**

Table 5. HOI detection performance comparison on three different sets of HOI categories in VidHOI dataset: Full, Non-rare and Rare.

Closed Set Video HOI detection In closed setting, all the HOI classes are predefined and exist in training set. By using full training set, our model achieves about 1.5% mAP lead on VidHOI dataset as shown in Table 1. For CAD-120 dataset, our method is on par with most of the SOTA methods for HOI detection.

We further compare the detection performance of temporal-related and spatial-related interactions on VidHOI dataset. If the predicates need to be inferred from neighboring frames, they’re temporal HOIs. Otherwise, they’re spatial HOIs. Resutls in Table 4 show that our model outperforms SOTA method more than 2% mAP on temporal-related interactions. On spatial-related interaction detection, our method leads by about 0.5% mAP. Our model performs better on temporal-related interactions probably because Dynamic GNN module explicitly captures temporal dynamics in video frames, while ASSIGN [28], ST-HOI [4] and TUTOR [43] do not contain such scheme.

Following ST-HOI [4], we test our model performance on three HOI sets of VidHOI dataset: (1) Full: all the 557 HOI classes are evaluated; (2) Non-rare: 242 HOI categories with more than 25 instances; (3) Rare: 315 HOI instances with less than 25 instances. Table 5 shows that our model performs better than ASSIGN [28], ST-HOI [4] and TUTOR [43] in all three settings. Among all the three settings, our method obtain the largest advantages on the Rare setting. This is in accordance with the open set setting in the last section. The strong generalizability for unseen categories also helps in detecting rare categories.

5. Conclusion

In this work, we propose ACoLP model to tackle the challenging open set video HOI detection problem. We model video HOIs in a novel action-centric manner, which aims to capture the essential *verbs* in videos. Furthermore, we propose the chain-of-look prompting scheme to generate spatio-temporal-aware action prompt representations and visual-semantic-aware HOI prompt representations, in order to model the underlying visual reasoning process in videos. Extensive experimental analysis validates the effectiveness of our model, which outperforms state-of-the-art HOI detection methods in the open set setting, and is on par with or better than existing methods in the closed setting.

Acknowledgements This material is based upon work supported under the AI Research Institutes program by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award # 2229873 - AI Institute for Transforming Education for Children with Speech and Language Processing Challenges. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

References

[1] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

[2] Hakan Cevikalp. Best fitting hyperplanes for classification. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1076–1088, 2016.

[3] Hakan Cevikalp, Bill Triggs, and Vojtech Franc. Face and landmark detection by using cascade of classifiers. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–7. IEEE, 2013.

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

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[6] Leizhen Dong, Zhimin Li, Kunlun Xu, Zhijun Zhang, Luxin Yan, Sheng Zhong, and Xu Zou. Category-aware transformer network for better human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19538–19547, 2022.

[7] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[8] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

[9] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(10):1775–1789, 2009.

[10] Mehadi Hassen and Philip K Chan. Learning a neural-network-based representation for open set recognition. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 154–162. SIAM, 2020.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[12] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*, pages 584–600. Springer, 2020.

[13] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016.

[14] Lalit P Jain, Walter J Scheirer, and Terrance E Boult. Multi-class open set recognition using probability of inclusion. In *European Conference on Computer Vision*, pages 393–409. Springer, 2014.

[15] Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19056–19065, 2022.

[16] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision*, pages 718–736. Springer, 2020.

[17] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[18] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8):951–970, 2013.

[19] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015.

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[21] Shuang Li, Yilun Du, Antonio Torralba, Josef Sivic, and Bryan Russell. Weakly supervised human-object interaction detection in video via contrastive spatiotemporal regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1845–1855, 2021.

[22] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[23] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[24] Xue Lin, Qi Zou, and Xixia Xu. Action-guided attention mining and relation reasoning network for human-object interaction detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1104–1110, 2021.

[25] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[27] Pedro R Mendes Júnior, Roberto M De Souza, Rafael de O Werneck, Bernardo V Stein, Daniel V Pazinato, Waldir R de Almeida, Otávio AB Penatti, Ricardo da S Torres, and Anderson Rocha. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106(3):359–386, 2017.

[28] Romero Morais, Vuong Le, Svetha Venkatesh, and Truyen Tran. Learning asynchronous and sparse human-object interaction in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16041–16050, 2021.

[29] Jihwan Park, SeungJun Lee, Hwan Heo, Hyeong Kyu Choi, and Hyunwoo J Kim. Consistency learning via decoding path augmentation for transformers in human object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1019–1028, 2022.

[30] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 401–417, 2018.

[31] Tanqiu Qiao, Qianhui Men, Frederick WB Li, Yoshiki Kubotani, Shigeo Morishima, and Hubert PH Shum. Geometric features informed multi-person human-object interaction recognition in videos. *ECCV*, 2022.

[32] Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19558–19567, 2022.

[33] 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.

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[35] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[36] Ron Rubinstein, Alfred M Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[37] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.

[38] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2317–2324, 2014.

[39] Matthew D Scherreik and Brian D Rigling. Open set recognition for automatic target classification with rejection. *IEEE Transactions on Aerospace and Electronic Systems*, 52(2):632–642, 2016.

[40] Ozan Sener and Ashutosh Saxena. rcrf: Recursive belief estimation over crfs in rgb-d activity videos. In *Robotics: Science and systems*, 2015.

[41] Sai Praneeth Reddy Sunkesula, Rishabh Dabral, and Ganesh Ramakrishnan. Lighten: Learning interactions with graph and hierarchical temporal networks for hoi in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 691–699, 2020.

[42] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021.

[43] Danyang Tu, Wei Sun, Guangtao Zhai, Wei Shen, et al. Video-based human-object interaction detection from tubelet

tokens. In *Advances in Neural Information Processing Systems*, 2022.

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[45] Guangzhi Wang, Yangyang Guo, Yongkang Wong, and Mohan Kankanhalli. Chairs can be stood on: Overcoming object bias in human-object interaction detection. *arXiv preprint arXiv:2207.02400*, 2022.

[46] Ning Wang, Guangming Zhu, Liang Zhang, Peiyi Shen, Hongsheng Li, and Cong Hua. Spatio-temporal interaction graph parsing networks for human-object interaction recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4985–4993, 2021.

[47] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022.

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

[49] John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[50] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[51] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

[52] Jiaxuan You, Tianyu Du, and Jure Leskovec. Roland: Graph learning framework for dynamic graphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2358–2366, 2022.

[53] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5678–5686, 2017.