Orchestrating the Symphony of Prompt Distribution Learning for Human-Object Interaction Detection

Mingda Jia[1], **Liming Zhao**[2], **Ge Li**[1*], **Yun Zheng**[2],

[1]Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology, Shenzhen Graduate School, Peking University

[2]Alibaba Group

mingdaj@stu.pku.edu.cn, geli@ece.pku.edu.cn, {lingchen.zlm, zhengyun.zy}@alibaba-inc.com

Abstract

Human-object interaction (HOI) detectors with popular query-transformer architecture have achieved promising performance. However, accurately identifying uncommon visual patterns and distinguishing between ambiguous HOIs continue to be difficult for them. We observe that these difficulties may arise from the limited capacity of traditional detector queries to represent diverse intra-category patterns and inter-category dependencies. To address this, we introduce **Inter**action **Pro**mpt **D**istribution Le**a**rning (InterProDa) approach. InterProDa learns multiple sets of soft prompts and estimates category distributions from various prompts. It then incorporates HOI queries with category distributions, making them capable of representing near-infinite intra-category dynamics and universal cross-category relationships. Our InterProDa detector demonstrates competitive performance on HICO-DET and vcoco benchmarks. Additionally, our method can be integrated into most transformer-based HOI detectors, significantly enhancing their performance with minimal additional parameters.

1 Introduction

Human-object interaction (HOI) detection (Chao et al. 2018; Kim et al. 2021) aims to identify the interactions between humans and objects within a static image and generate predictions in the format $<$ human, interaction, object$>$. Human interactions typically exhibit more complex and diverse visual content than objects. Arguably, the primary difficulty of detecting HOI is understanding the diversity of action patterns.

Recent advancements in querying-transformer-based HOI methods (Tamura et al. 2021; Ning et al. 2023; Lei et al. 2023; Wang et al. 2024; Liao et al. 2022) have achieved promising performance. Despite this progress, recent HOI detectors still face two difficulties. Firstly, they often struggle to recognize interactions involving uncommon visual content. For instance, interactions like $jumping$ a $motorcycle$ suggest a range of visual patterns, such as driving over a ramp or performing mid-air stunts. While humans can easily categorize various patterns, intelligent models tend to misclassify rare visual content into common patterns in the training set, as illustrated in Figure 1. Secondly, HOI detectors sometimes confuse visually similar HOIs.

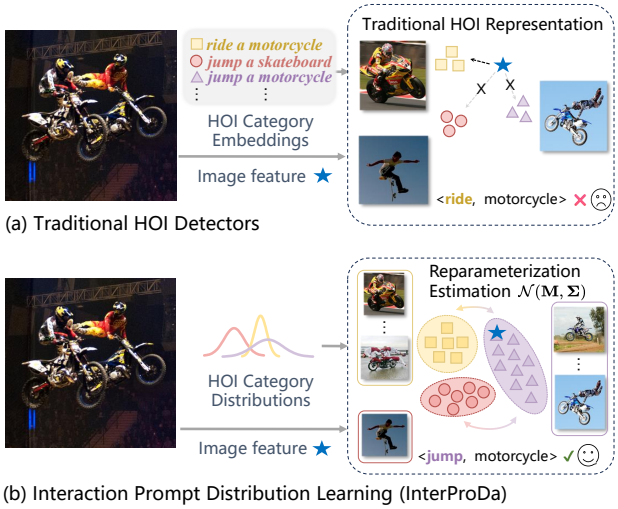

Figure 1: Traditional HOI detectors memorize limited and common visual content, which makes them fragile in recognizing uncommon visual patterns. InterProDa models each HOI category as a distribution to represent unlimited intra-category patterns. We also learn cross-category dependencies with constraints between every distribution. Zoom in for details.

To address these difficulties and improve zero-shot capabilities, many recent HOI detectors incorporate prior knowledge from pre-trained visual-linguistic models (VLM) into their detection queries by learning discrete (Liao et al. 2022; Ning et al. 2023) or soft (continuous) (Lei et al. 2023; Wang et al. 2024) prompt embeddings of HOI categories. These category-specific priors shape detector queries to be category queries (Xie et al. 2023), which learn cross-category dependencies for classification. However, these models still face challenges in capturing diverse intra-category patterns since they do not incorporate any explicit scheme to represent the variations within each query or category.

It motivates us to explore a novel HOI category query structure incorporating an intra-category pattern dimension.

*Corresponding author

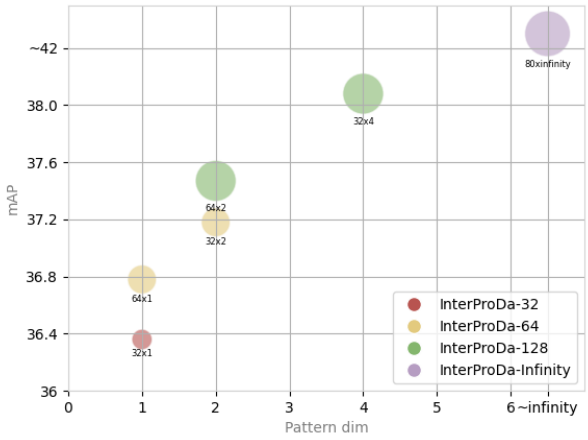

Figure 2: **Performance comparison w.r.t different settings of decoder queries** on HICO-DET. We expand the query of our model with an additional intra-category pattern dimension. Each model is labeled with a format like 32×2, which indicates a model with a category (query) dimension of 32 and pattern dimension of 2. Models with the same color share identical parameters and overall query dimensions (for example, 32×2 equals 64×1). Comparisons between models of the same color demonstrate that a higher pattern dimension improves performance. More experiment details in Section 4.7.

While exploring the query pattern design, we observed a noticeable phenomenon that, as shown in Figure 2, increasing the query pattern dimension enhances model performance without changing the total number of model parameters. It inspires us to further extend the query pattern dimension, ideally towards infinity, to improve performance even more.

Hence, we propose **Inter**action **Pro**mpt **D**istribution Le**a**rning (InterProDa), which learns category distribution queries representing approximately infinite intraclass patterns and universal inter-category dependencies in the continuous distribution space.

In InterProDa, we learn category prompt embeddings as HOI knowledge representation. We first split HOI categories into three groups, including implicit human classes, explicit object classes, and explicit interaction classes. For each group, we design specialized prompt structures. To capture diverse intra-category visual patterns, we learn a collection of soft prompts (Li and Liang 2021) for each HOI category. Subsequently, we estimate a unique continuous distribution for each group of prompts. A crucial aspect of distribution learning is efficiency. Discriminating between HOI categories involves calculating the marginal likelihood across the multivariate distribution of all categories, which demands substantial computations (Williams and Barber 1998). To optimize this, we assume HOI category representations follow Gaussian distributions, and then we can estimate these distributions efficiently through their means and standard variations.

We capture global inter-category dependencies in the shared feature space of all HOI category distributions. We introduce a novel dynamic orthogonal constraint to enhance space learning with explicit supervision. This constraint increases the distance between ambiguous interaction categories and expands the separation among semantically distinct categories.

In the feed-forward process, we perform sampling on the distribution spaces using a reparameterization trick similar to generative variational auto-encoders (VAE) (Kingma and Welling 2022) to make the distribution spaces differentiable. We observe that, though the VAE style reparameterization is suitable for generative models, their sampling process with noises makes deterministic models unstable. To mitigate this, we introduce a learnable noise factor during sampling to achieve a balance between learning diverse representations and maintaining stability in HOI detection.

We sample from the learned distribution space and obtain online distribution guidance. By integrating this guidance with learnable queries, we obtain category distribution queries enhancing HOI prediction. InterProDa with our best practice achieves competitive performance on HICO-DET and vcoco. Additionally, our distribution learning pipeline can easily incorporate and boost most HOI detectors with lightweight additional parameters.

Contributions. Our contributions are threefold:

- To the best of our knowledge, we are the first to expand query learning to include category distributions, enhancing the recognition of HOIs through uncommon visual content and distinguishing between ambiguous HOIs.
- We propose InterProDa, which represents near-infinite intra-category visual patterns and universal inter-category dependencies by category distributions.
- InterProDa achieves competitive performance on HOI benchmarks. It can also incorporate and enhance most HOI detectors with minimal additional parameters.

2 Related Works

2.1 HOI Detection

HOI detectors with detection transformers have achieved significant success in recognizing human actions toward objects. Both one-stage HOI detectors (Tamura et al. 2021; Liao et al. 2022; Ning et al. 2023; Li et al. 2023) and two-stage detectors (Zhang et al. 2022; Lei et al. 2023; Wang et al. 2024) show their strengths in different detection scenarios. However, recent HOI detectors equipped with traditional decoder queries (Ning et al. 2023) and category-wise priors (Xie et al. 2023) from VLMs still lack the capacity to represent intra-category dynamics sufficiently.

2.2 Prompt Learning

Prompt learning is an impressive knowledge distillation approach in natural language processing (NLP). It transfers the knowledge of pre-trained models to downstream tasks by learning proper prompt templates (Brown et al. 2020; Jiang et al. 2020; Shin et al. 2020). Adopting prompt learning for computer vision, (Radford et al. 2021; Jia et al. 2021) aligns large-scale visual features with hand-crafted category prompts to train large visual linguistic models (VLM).

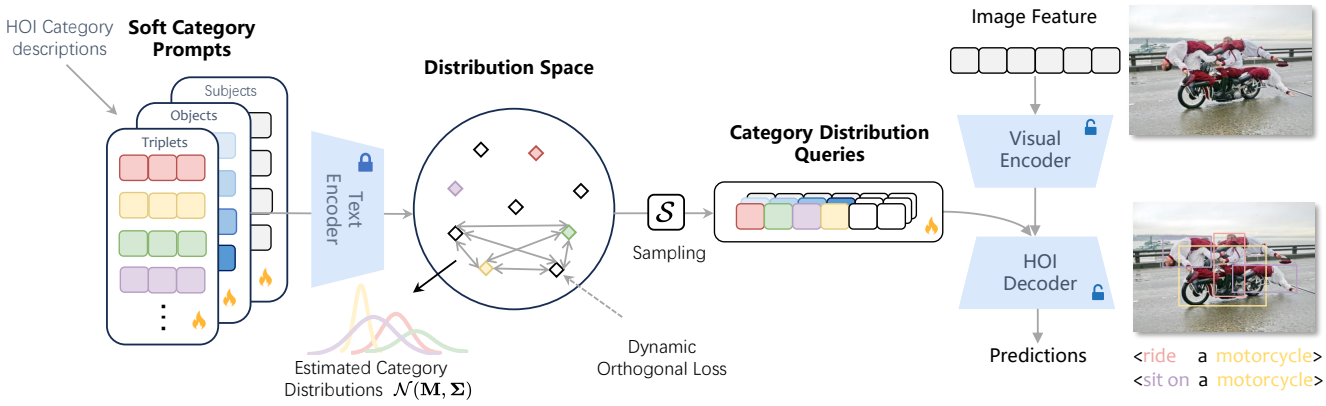

Figure 3: **The pipeline of InterProDa.** We learn multiple groups of soft prompts for subject, object, and interaction categories. Then, we estimate the distributions of these category prompt embeddings and constrain them in a continuous feature space. Such an approach learns diverse intra-category patterns in each category distribution and captures the universal inter-category dependencies. We sampled from the learned distribution space to obtain a category distribution query to enhance HOI prediction.

Leveraging *prompt tuning*, CoOp (Zhou et al. 2022) learns a soft prompt for visual classification.

In HOI detection, several works utilise discrete category prompt (Liao et al. 2022; Ning et al. 2023; Yuan et al. 2023), soft category prompt (Lei et al. 2023; Wang et al. 2024) and tailored multimodal prompts (Yang et al. 2024; Lei et al. 2024) to improve HOI detection. ADA-CM (Lei et al. 2023) is most related to us, as they tune soft text prompts for HOI detection. Different from them, we learn the distributions of various prompts.

2.3 Prompt Distribution Learning

The varying dynamics of visual content are more complex than language semantics (He et al. 2021). To address this, ProDA (Lu et al. 2022) estimates a category distribution from diverse prompts to achieve generalized adoption of downstream recognition tasks. DreamDistribution (Zhao et al. 2023) learns a prompt distribution from user input images, allowing the personalized text-to-image (T2I) generation of novel images. Inspired by their advancement, we leverage prompt distribution to represent theoretically infinite inter-class HOI visual patterns and universal cross-category HOI dependencies.

3 Method

In this section, we term our InterProDa, which can approximately represent infinite intra-category patterns and efficiently capture universal inter-category dependencies. The overall pipeline is illustrated in Figure 3.

3.1 HOI Detection

Our HOI detector is shown on the right of Figure 3. Given an input image feature map summarized by a visual backbone, a trainable visual encoder will take the feature map and summarize visual content memory $\mathbf{Z}$. Then, a set of learnable query embeddings $\boldsymbol{Q}$ will be leveraged by the HOI decoder to ground informative features from $\mathbf{Z}$ and generate decoder features $\mathbf{Z}_{dec}$. This process can be expressed by:

$$\mathbf{Z}_{dec} = \text{Decoder}\left(\text{Encoder}(\mathbf{Z}), \boldsymbol{Q}\right), \tag{1}$$

Several prediction heads will take the decoder features $\mathbf{Z}_{dec}$ and generate HOI predictions (b_s, b_o, c_o, c_{hoi}), where b_s is the bounding boxes of subjects in the image, b_o is the object boxes of objects, c_o is the object categories, and c_{hoi} is the categories of HOI triplet combinations.

3.2 Interaction Prompt Distribution Learning

Prompt learning. InterProDa aims to learn multiple groups of soft prompts describing the category characteristics for HOI detection. For each HOI category, we learn multiple prompts to represent it. Here, we first present how to learn a single prompt for a category. We feed a randomly initialized natural language prompt to a pre-trained text encoder $\mathcal{F}$ to obtain initial prompt embedding. This initial prompt can be expressed by $\boldsymbol{p} \in \mathbb{R}^{L \times e}$, where L is the length of tokens in the prompt and e is the embedding dimension. Our best practice is using the openAI CLIP (Radford et al. 2021) text encoder.

Then, the category prompt $\boldsymbol{p}_{c_i}$ of a single category c_i in all the HOI category descriptions $\{c_i\}_{i=1}^{N}$ can be obtained by concatenating the initial prompt $\boldsymbol{p}$ with the CLIP text embeddings of c_i, where N is the number of all categories. Specifically, the category prompt $\boldsymbol{p}_{c_i}$ follows the format [PREFIX]$\mathbf{V}$[SUFFIX], with [PREFIX] and [SUFFIX] representing the category-specific prefix and suffix embeddings, respectively. $\mathbf{V} \in \mathbb{R}^{M \times e}$ signifies the learnable token embedding of the corresponding category, where M represents its token length. The sum of M and the token length of each c_i should be L. During prompt learning, $\mathbf{V}$ will be updated by the gradient backpropagated through the HOI decoder.

HOI prompt design. As illustrated in Figure 3, we design three HOI-aware structures for category prompts. Given that

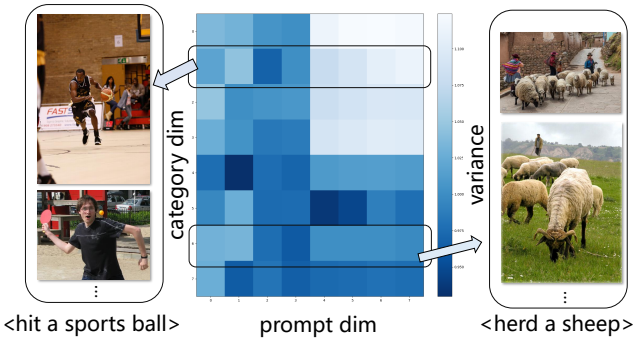

Figure 4: **Visualization of intra-category variances** of selected prompt distributions learned from HICO-DET. Each row refers to different categories, while each token refers to the variance of a single prompt. We select four distributions with the highest average variance and 4 with the lowest variance, respectively. We sort and list them from top to bottom. We also show HICO-DET images that correspond to two of these category distributions, the distributions with high variance always indicate HOI categories with more diverse visual patterns.

subjects, objects, and interactions highlight different types of visual content, we employ subject, object, and interaction category prompts to represent these respective concepts.

First, we initialize a set of interaction category prompts $\mathbf{P}_{int} \in \mathbb{R}^{N_{hoi} \times L \times e}$, where $\mathbf{P}_{int}$ is the combination of all $\{\boldsymbol{p}_c, c \in C_{hoi}\}$. C_{hoi} represents text descriptions of all HOI combinations in the format of <interaction> a/an <object>. N_{hoi} is the total number of HOI triplet combinations in the training dataset.

Second, we construct a set of object category prompts $\mathbf{P}_{obj} \in \mathbb{R}^{N_{obj} \times L \times e}$. Similar to the interaction prompts, $\mathbf{P}_{obj}$ is the collection of all $\{\boldsymbol{p}_c, c \in C_{obj}\}$, N_{obj} is the total number of object categories. Here, we set C_{obj} as the text labels of object categories.

Finally, for subject prompts, we do not use the text label $human$ as the category description. Humans in the real world inherit fertile semantics like drivers, farmers, athletes, etc. We construct a set of subject category prompts with anonymous categories, leaving all the tokens in these prompts learnable. Thus, subject prompts $\mathbf{P}_{sub} \in \mathbb{R}^{N_{sub} \times L \times e}$ are filled with N_{sub} randomly initialized soft prompts, where N_{sub} is set to the same as N_{obj}.

Learning prompt distribution. InterProDa further learns the distribution of each HOI category by estimating it from various category prompts. As the CLIP text embeddings with similar semantics remain close in the feature space (Lu et al. 2022), our prompt embeddings can be represented by a simple and general distribution. Recent studies found that Gaussian distributions are inherently suitable for feature representations learned by neural networks (Liu et al. 2023a,b; Zhang et al. 2022). Therefore, we assume that for each HOI category, all the prompt embeddings describing this category follow the same Gaussian distribution. In the following paragraphs, we use the same character to represent the shared process on prompts with different HOI-aware structures.

Specifically, we learn a soft prompt collection consisting of K learnable category prompts for each HOI category, resulting in $\mathcal{P}_{c_i} \in \mathbb{R}^{K \times L \times e}$. Each collection contains K learnable token embeddings $\{\mathbf{V}_k\}_{k=1}^{K}$. Leveraging the Gaussian distribution prior, we estimate the means $\boldsymbol{\mu}_{c_i} = \mu(\mathcal{P}_{c_i}) \in \mathbb{R}^{L \times e}$ and standard deviations $\boldsymbol{\sigma}_{c_i} = \sigma(\mathcal{P}_{c_i}) \in \mathbb{R}^{L \times e}$ of the normal distribution $\mathcal{N}(\boldsymbol{\mu}_{c_i}, \boldsymbol{\sigma}_{c_i}^2)$ for a given category c_i. It is worth noticing that K can be theoretically infinite, enabling a distribution estimation that closely approximates the target HOI distribution (Chen et al. 2023) in the real world, offering the potential ability to encapsulate real-world feature representations.

3.3 Distribution Space Learning

To learn the distributions of all HOI categories simultaneously, we stack all the category distributions $\mathcal{P}_{c_i}$ together to form a universal distribution space $\boldsymbol{\mathcal{P}} \in \mathbb{R}^{N \times K \times L \times e}$. We utilize $\mathcal{N}(\mathbf{M}, \mathbf{\Sigma})$ to represent the multivariate normal distributions $\mathcal{N}(\boldsymbol{\mu}_{c_1:c_N}, \boldsymbol{\sigma}^2_{c_1:c_N})$ describing the distribution space $\boldsymbol{\mathcal{P}}$. This unified space enables our model to effectively capture cross-category dependencies. Building on our design of HOI-aware prompt structures in section 3.2, we independently learn three distribution spaces for subjects, objects, and interactions, represented by $\boldsymbol{\mathcal{P}}_{sub} \in \mathbb{R}^{N_{sub} \times K \times L \times e}$, $\boldsymbol{\mathcal{P}}_{obj} \in \mathbb{R}^{N_{obj} \times K \times L \times e}$ and $\boldsymbol{\mathcal{P}}_{int} \in \mathbb{R}^{N_{hoi} \times K \times L \times e}$.

Space learning constraint. We propose a novel orthogonal constraint with dynamic margins to explicitly guide the learning procedure of the distribution spaces. We aim to maintain that different distributions in the distribution spaces always focus on distinct HOI concepts. In contrast, the distributions of semantically close categories should be closer in the distribution space. The orthogonal constraint calculates a dynamic margin between different distributions according to their cosine similarity:

$$\cos(i,j) = \frac{\left|\bar{\mathcal{P}}_{c_i}, \bar{\mathcal{P}}_{c_j}\right|}{\|\bar{\mathcal{P}}_{c_i}\|_1 \|\bar{\mathcal{P}}_{c_j}\|_1 + \epsilon}, i \neq j, \tag{2}$$

where $\bar{\mathcal{P}}$ means the mean-pooled prompt embeddings of each category with shape $\mathbb{R}^{1 \times L \times e}$. Then, their dynamic margin Δ can be calculated by $\Delta(i,j) = \alpha(1 - \cos(i,j)$, where α is a learnable tuning parameter. Finally, the dynamic orthogonal loss can be expressed by:

$$\mathcal{L}_{do} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(\max\left(\epsilon, \Delta(i,j) - \cos(i,j)\right)\right)^2. \tag{3}$$

As the different semantic concepts of subjects do not vary too much as that of objects or interactions, we only calculate $\mathcal{L}_{do}$ in $\boldsymbol{\mathcal{P}}_{obj}$ and $\boldsymbol{\mathcal{P}}_{int}$.

Space aggregation. A significant challenge of HOI distribution space learning is parameter efficiency. To address this issue, we propose a tailor-made distribution aggregation module to compress the original space to a compressed representation $\hat{\boldsymbol{\mathcal{P}}} \in \mathbb{R}^{N' \times K \times L \times e}$, where $N' < N$ represents the number of compressed category distributions. The original N prompt distributions representing explicit HOI categories are summarized into high-level categories aggregating semantically similar categories.

Sampling from distribution space. Each sampling process on the distribution of high-level category c is represented by $\boldsymbol{\mathcal{S}}(\hat{\mathcal{P}}_c) = \hat{\boldsymbol{\mu}}_c + \boldsymbol{n}\hat{\boldsymbol{\sigma}}_c$, in which $\boldsymbol{n}$ is White Gaussian noise randomly sampled from $\mathcal{N}(0, 1)$.

To solve the noise in VAE-style reparameterization, we attempt to repeat $\hat{\boldsymbol{\mu}}$ as the estimation of the distribution. However, it will limit the ability to represent fertile HOI patterns. We aim to achieve a balance by adding a learnable scaling factor γ on $\boldsymbol{n}$. The sampling process is then modified to $\boldsymbol{\mathcal{S}}(\hat{\mathcal{P}}_c) = \hat{\boldsymbol{\mu}}_c + \gamma\boldsymbol{n}\hat{\boldsymbol{\sigma}}_c$.

To obtain distribution samples that can sufficiently represent the original distribution, we sample each category distribution for N_s times to obtain the final distribution sampling for each category. This final sampling is expressed by $[\hat{\mathcal{P}}]_c = \prod_{i=1}^{N_s}(\hat{\boldsymbol{\mu}}_c + \gamma\boldsymbol{n_i}\hat{\boldsymbol{\sigma}}_c)$, where $N_s < K$. $[\hat{\mathcal{P}}]_c$ has shape $\mathbb{R}^{N_s \times L \times e}$. $\prod$ represents concatenating N_q samples of the same distribution along a new dimension. The sampling of the universal HOI category space can be represented by $[\hat{\boldsymbol{\mathcal{P}}}]$, with shape $\mathbb{R}^{N \times N_s \times L \times e}$. Such sampled distribution space representations are suitable for the feed-forward process and can be updated by propagated gradients.

3.4 Model Implementation

Knowledge distillation. We implement our HOI decoder with two sequential transformer decoders, including instance decoder and interaction decoder, according to (Ning et al. 2023). The instance decoder and interaction decoder are equipped with learnable instance queries $\boldsymbol{Q}_{ins} \in \mathbb{R}^{(N_q \times N_s) \times C}$ and interaction queries $\boldsymbol{Q}_{int} \in \mathbb{R}^{(N_q \times N_s) \times C}$.

Given the learned category distribution spaces $\boldsymbol{\mathcal{P}}_{sub}$, $\boldsymbol{\mathcal{P}}_{obj}$ and $\boldsymbol{\mathcal{P}}_{int}$, we will sample from them during both training and inference. The sampling generates subject-aware, object-aware, and interaction-aware distribution guidance, termed $[\hat{\boldsymbol{\mathcal{P}}}]_{sub}$, $[\hat{\boldsymbol{\mathcal{P}}}]_{obj}$ and $[\hat{\boldsymbol{\mathcal{P}}}]_{int}$ respectfully. Then, we will take mean pooling on their token dimension and map them to the hidden space of HOI queries by learnable linear layers. We set the compression target N' of our space aggregation module to be the same as N_q. Finally, we integrate the distribution guidance and the HOI queries by:

$$\begin{aligned} \boldsymbol{Q}_d^{ins} &= \boldsymbol{Q}_{ins} \oplus \text{Linear}(\text{cat}([\hat{\boldsymbol{\mathcal{P}}}]_{sub}, [\hat{\boldsymbol{\mathcal{P}}}]_{obj})), \\ \boldsymbol{Q}_d^{int} &= \boldsymbol{Q}_{int} \oplus \text{Linear}([\hat{\boldsymbol{\mathcal{P}}}]_{int}), \end{aligned} \tag{4}$$

where $\oplus$ refers to weight addition. Then, the obtained distribution queries $\boldsymbol{Q}_d^{ins}$ and $\boldsymbol{Q}_d^{int}$ will be fed into the corresponding HOI decoders for further process.

Training. We train InterProDa with an end-to-end fully tuning approach. Both the interaction distribution learning process and the HOI detectors are tuned simultaneously, with a multi-task loss:

$$\mathcal{L} = \mathcal{L}_{HOI} + \lambda\mathcal{L}_{do}, \tag{5}$$

where $\mathcal{L}_{HOI}$ is the HOI set-prediction loss according to (Tamura et al. 2021) and λ is a hyperparameter.

Method	Feature Extractor	HICO-DET (Default)			vcoco
		Full	*Rare*	*Nonrare*	$AP^{\#1}_{role}$
HOTR (Kim et al. 2021)	R50	25.10	17.34	27.42	55.2
QPIC (Tamura et al. 2021)	R50	29.07	21.85	31.23	58.8
CPC (Park et al. 2022)	R50	29.63	23.14	31.57	63.1
GENVLKT (Liao et al. 2022)	R50	33.75	29.25	35.10	62.4
PViC (Zhang et al. 2023)	R50	34.69	32.14	35.45	62.8
HOICLIP (Ning et al. 2023)	R50 + C	34.69	31.12	35.74	63.5
CQL (Xie et al. 2023)	R50	35.36	32.97	36.07	66.4
RLIPv2 (Yuan et al. 2023)	R50	35.38	29.61	37.10	65.9
LOGICHOI(Li et al. 2023)	R50	35.47	32.03	36.22	64.4
MPHOI(Cao et al. 2023a)	R50 + C&D	36.50	35.48	36.80	66.2
RmLR(Cao et al. 2023a)	R50	36.93	29.03	39.29	63.7
ViPLO (Part et al. 2023)	R50 + C	37.22	35.45	37.75	62.2
SCTC (Jiang et al. 2024)	R50	37.92	34.78	38.86	67.1
CMMP (Lei et al. 2024)	R50 + C	38.14	37.75	38.25	64.0
ADA-CM (Lei et al. 2023)	R50 + C	38.40	37.52	38.66	58.5
BCOM (Wang et al. 2024)	R50 + C	39.34	39.90	39.17	65.8
UniHOI (Cao et al. 2023b)	R50 + B	40.06	39.91	40.11	65.6
InterProDa (Ours)	R50 + C	**42.67**	**45.21**	**41.92**	**67.6**

Table 1: **Performance comparison with state-of-the-art on benchmark HICO-DET and vcoco.** under fully-supervision setting. R50 refers to ResNet50, C refers to CLIP (Radford et al. 2021), B refers to BLIP2 (Li et al. 2022), D refers to Stable Diffusion (Rombach et al. 2022). Our model achieves competitive performance, especially 45.21 mAP on the $rare$ split.

4 Experiments

4.1 Datasets & Metrics

Datasets and benchmarks. We train and evaluate InterProDa on two widely-used HOI Detection datasets, HICO-DET (Chao et al. 2018) and vcoco (Gupta and Malik 2015). HICO-DET includes 80 object categories, 117 verb categories and 600 HOI triplet combinations. We split HICO-DET into 38,118 training&evaluation images and 9,658 test images according to (Tamura et al. 2021). vcoco has 5,400 training images and 4,946 validation images, consisting of 81 object categories, 29 verb categories, and 263 HOI triplet combinations.

Metrics. We evaluate mean Average Precision (mAP) as our main performance on HICO-DET, following (Chao et al. 2018). Specifically, for an HOI triplet prediction, we first calculate if the IoUs of the human bounding box and object bounding box between groud-truths are larger than 0.5. Then, we check if the predicted categories of interactions and objects are positive. We report both $Default$ and $Known\ Object$ (KO) split of HICO-DET. For vcoco, we report scenario 1 AP.

4.2 Implementations

Here, we detail the implementation of our best practices. In prompt distribution learning, we set the number of prompts per category, K, to 8, with each prompt having a token length L of 77. The reparameterization sampling number N_s is set to 2 per distribution, and the noise scale γ for sampling is initialized at 1e-2. The loss weight λ for dynamic orthogonal loss is set to 5e-2, and the dynamic margin tuning parameter α is set to 0.5. We use openAI CLIP-ViT-L/32 text encoder for generating prompt embeddings.

Method	Type	*Unseen*	*Seen*	*Full*
GEN-VLKT (Liao et al. 2022)	RF-UC	21.36	32.91	30.56
HOICLIP (Ning et al. 2023)	RF-UC	25.53	34.85	32.99
ADA-CM (Lei et al. 2023)	RF-UC	27.64	34.35	33.01
BCOM (Wang et al. 2024)	RF-UC	28.52	35.04	33.74
CMMP (Lei et al. 2024)	RF-UC	35.98	37.42	37.13
Ours	RF-UC	**36.38**	**40.88**	**39.58**
GEN-VLKT (Liao et al. 2022)	NF-UC	25.05	23.38	23.71
HOICLIP (Ning et al. 2023)	NF-UC	26.39	28.10	27.75
ADA-CM (Lei et al. 2023)	NF-UC	32.41	31.13	31.39
BCOM (Wang et al. 2024)	NF-UC	33.12	31.76	32.03
CMMP (Lei et al. 2024)	NF-UC	33.52	35.53	34.50
Ours	NF-UC	**33.64**	**36.47**	**35.50**

Table 2: **Zero-shot HOI Detection**. RF-UC and NF-UC refer to the rare- and non-rare-first unseen combination settings following (Liao et al. 2022).

Model setting	*Full*	*Rare*	*Non-rare*
base	36.70	32.58	37.93
+ prompt query	37.47	34.38	38.40
+ distribution	38.08	36.14	38.65
+ VLM $\mathcal{F}$	39.56	40.91	39.16
+ $\mathcal{L}_{do}$	41.85	43.80	41.27
+ Reparam. γ	42.19	43.36	41.81
+ prompt design	**42.67**	**45.21**	**41.92**

Table 3: **Ablations** on proposed components.

methods	Prior	*Full*	*Rare*	*Non-rare*	params
HOICLIP (Ning et al. 2023)	CLIP-B	34.69	31.12	35.74	66M
HOICLIP + ours	CLIP-B	**38.04**	**35.92**	**38.68**	**68M**
ADA-CM (Lei et al. 2023)	CLIP-L	38.40	37.52	38.66	4M
ADA-CM + ours	CLIP-L	**41.71**	**42.92**	**41.34**	**5M**

Table 4: **Performances of traditional models equipped with our InterProDa**.

We train the distribution space and the HOI decoder concurrently for 60 epochs with an AdamW optimizer on both HICO-DET and vcoco. We set the initial learning rate to 1e-4 and decays after 40 epochs. We set the random seed at 42. For zero-shot training, we follow the setup in (Ning et al. 2023), and we mask the category priors of prompts for unseen categories and initialize them with 8 placeholder tokens,

Our HOI decoder follows an end-to-end transformer encoder-decoder architecture. The encoder and instance decoder have their hidden dimension of 256, while our interaction decoder has the same hidden dimension with the CLIP hidden dimension following (Ning et al. 2023). We construct the HOI detector with 6-layer encoders, 3-layer instance decoders and 3-layer interaction decoders. We set query dimension N_q as 80 for both $\boldsymbol{Q}_d^{ins}$ and $\boldsymbol{Q}_d^{int}$, and their pattern dimension N_s are set to 2, which is also the number of reparameterization samplings. We implement the distribution space aggregation module with a learnable linear layer, which first permutates the distribution space and then maps the category dimension of the space feature to $N' = N_q$.

4.3 Performance Comparison

Fully supervised performance. We present a performance comparison between InterProDa and SOTA methods trained or fine-tuned with full supervision on HICO-DET and vcoco, detailed in Table 1. InterProda with a ResNet50 backbone and a CLIP feature extractor achieves new SOTA performances of 42.67 $full$ mAP and 45.21 $rare$ mAP on HICO-DET (default). Noticing that our performance enhancement is more significant in the $rare$ split, it highlights the role of interaction distribution learning in exploring hard and uncommon HOI visual content. For vcoco, InterProDa also achieves SOTA 67.6 scenario 1 mAP.

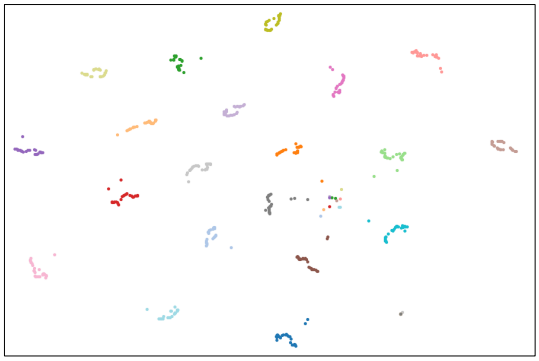

Figure 5: **T-SNE visualization of learned prompt distributions of 20 random interaction categories on HICO-DET.** Each cluster with the same color refers to the category prompt embeddings belonging to a distinct HOI category distribution. The learned distribution space has clear margins between categories, showing suitable cross-category dependencies. Best viewed in color.

Zero-shot performance. As CLIP is utilized in InterProDa in the prompt distribution learning, we also compare zero-shot performance with related works. As shown in Table 2, InterProDa outperforms recent SOTA zero-shot HOI detectors on both $unseen$ and $seen$ evaluations than recent method considering prompt learning (Lei et al. 2023; Wang et al. 2024). It is worth noticing that InterProDa with single-modal prompts performs better than CMMP (Lei et al. 2024) with multi-modal prompts, highlighting the effectiveness of prompt distribution learning.

Incorporation with previous works. We provide portability experimentation on representative works of both one-stage and two-stage HOI detectors, HOICLIP (Ning et al. 2023) and ADA-CM (Lei et al. 2023). For HOICLIP, we incorporate InterProDa into its decoder queries. For ADA-CM, we combine its prompt learner with our distribution learning. As illustrated in Table 4, InterProDa enhances HOICLIP (Ning et al. 2023) to achieve both $full$ and $rare$ mAP with only 2M additional parameters. For ADA-CM, we introduce a significant increase in $rare$ mAP with less than 1M additional trainable parameters.

4.4 Component Ablation

We evaluate the effectiveness of our proposed components in Table 3. The base item refers to the fundamental HOI detector detailed in the implementation details. In the next row, we initialize three sets of learnable prompts without HOI

(a)

$\mathcal{L}_{do}$ design	*Full*	*Rare*	*Non-rare*
w/o Δ	41.62	43.75	40.98
w / Contra.	41.60	43.49	41.04
w/ Δ	**41.85**	**43.80**	**41.27**

(b)

Sampling strategy	*Full*	*Rare*	*Non-rare*
VAE Reparam.	42.09	42.40	41.99
Reparam. w/ γ	**42.67**	**45.21**	41.92
Repeat $\boldsymbol{\mu}$	42.39	43.23	**42.14**

(c)

Space basis	*Full*	*Rare*	*Non-rare*
Naive	40.94	40.25	41.31
Gaussian	**42.01**	**44.44**	**41.29**
Fourier	42.40	44.42	41.79

(d)

K	*Full*	*Rare*	*Non-rare*	N_s	*Full*	*Rare*	*Non-rare*	λ	*Full*	*Rare*	*Non-rare*
2	41.62	43.75	40.96	8	39.79	40.42	39.80	1e-1	41.62	43.75	40.98
4	41.83	43.96	**41.19**	4	41.49	43.71	40.82	1e-2	41.72	43.52	41.19
8	**41.93**	**44.77**	41.08	2	**41.62**	**43.75**	**40.96**	5e-2	**42.09**	**44.19**	**41.46**

Table 5: **Ablations** of (a) designs of dynamic orthogonal loss, (b) sampling strategy, (c) feature space bases and (d) model hyperparameters. All experiments are conducted with ResNet50 backbone and CLIP-L on HICO-DET.

category priors. For the $distribution$ item, we expand these prompts to prompt collections and model them as Gaussian distributions. By mapping the prompts to VLM text embeddings, the model gains higher $rare$ mAP. Equipping the distribution space with $\mathcal{L}_{do}$, the capability to represent fertile visual patterns is simulated. By updating our distribution prompts to HOI-aware prompts with different tailored structures, our model receives more improvement.

4.5 Analysis of Learned Distributions

Visualization of intra-category patterns. We provide visualizations to analyze the learned intra-category patterns in each HOI category distribution, as shown in Figure 4. The plots display the standard variances of prompts within each category. We observe that distributions with higher standard variances tend to correspond to HOI categories with more complex and diverse patterns. For example, in the figure, the category hit a $ball$ covers various activities, such as playing basketball and hitting a ping-pong ball. In contrast, the $herd$ a $sheep$ category show similar poses and spatial arrangements. These distribution characteristics align well with our expectations.

Visualization of inter-category dependencies. We also visualize the learned inter-category dependencies in our HOI distribution space, as presented in Figure 5. The distributions maintain a relatively clear margin between others, making the space representation suitable for HOI determination.

4.6 Other Ablations

Dynamic orthogonal constraint. We compare different designs of $\mathcal{L}_{do}$ in Table 5 (a). Contra. refers to adding hard contrastive margins without the tuning parameter α. It shows the effectiveness of the dynamic margin Δ in capturing suitable distribution space dependencies.

Different sampling Strategies. We also test different sampling strategies in Table 5 (b). VAE Reparam. setting follows (Kingma and Welling 2022), and Reparam. w/γ indicates our proposed approach. Repeat $\boldsymbol{\mu}$ refers to filling each distribution sample with the mean of the distribution. From the table, our learnable noise factor minimizes the impact of the noise on deterministic inference.

Model hyper-parameters. We detailed our hyperparameter experiments in Table 5 (d). For each experiment, we set irrelevant K, N_s and λ to 2, 2 and 1e-2 by default. It is noticeable that K can be theoretically set to ∞ with ideal memory storage, we set K to 8 for model efficiency.

Different space bases. We evaluate different space bases in modeling the continuous HOI category representation in Table 5 (c). $Naive$ means sampling prompts from collections of the category prompt embeddings by max-pooling in each prompt group. $Fourier$ is stated in section 4.8.

4.7 Exploration of Query Pattern Dimension

For the performances illustrated in Figure 2, we learn randomly initialized prompt embeddings with different category dimensions and pattern dimensions, like $\mathcal{P} \in \mathbb{R}^{N_q \times N_p \times C}$, where N_q is the category dimension, N_p is the pattern dimension. Orthogonal loss is added to the N_q dimension to distinguish different categories. In the HOI detector, we reshape their queries to $\mathcal{P} \in \mathbb{R}^{(N_q \times N_p) \times C}$ for calculating $\mathcal{L}_{hoi}$. In this setting, for instance, models with labels (32×2) and (64×1) in Figure 2 share the same parameter amounts but with different learning processes before being fed into the detector.

4.8 Prompts in Frequency Space

In the view of signal processing, modeling category embeddings as distributions is actually representing them with Gaussian bases. The orthogonal Fourier basis is another signal basis suitable for reconstructing varying signals due to its continuous nature. We attempt to utilize Fourier transformation to project pooled category prompt embeddings to the frequency domain. This approach also obtains similar performance with using Gaussian distributions.

5 Conlusion

We first introduce a novel interaction distribution learning approach, InterProDa, to HOI detection, which learns prompt distributions for HOI categories. InterProDa learns category distribution queries to represent nearly infinite intra-category HOI patterns. It also learns universal cross-category dependencies. InterProDa, to our best practice, achieves new State-of-the-Art performance on HOI detection benchmarks and can significantly boost existing query-based HOI detectors with lightweight additional parameters.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62172021 and the Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology under Grant 2024B1212010006.

References

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20.

Cao, S.; Yin, Y.; Huang, L.; Liu, Y.; Zhao, X.; Zhao, D.; and Huang, K. 2023a. Efficient-VQGAN: Towards High-Resolution Image Generation with Efficient Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7368–7377.

Cao, Y.; Tang, Q.; Su, X.; Chen, S.; You, S.; Lu, X.; and Xu, C. 2023b. Detecting Any Human-Object Interaction Relationship: Universal HOI Detector with Spatial Prompt Learning on Foundation Models. In *Advances in Neural Information Processing Systems*, volume 36, 739–751.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to Detect Human-Object Interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 381–389.

Chen, Y.; Sun, S.; Li, G.; Gao, W.; and Li, T. H. 2023. Closing the Gap Between Theory and Practice During Alternating Optimization for GANs. *IEEE Transactions on Neural Networks and Learning Systems*.

Gupta, S.; and Malik, J. 2015. Visual Semantic Role Labeling. arXiv:1505.04474.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2021. Masked Autoencoders Are Scalable Vision Learners. arXiv:2111.06377.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning*.

Jiang, W.; Ren, W.; Tian, J.; Qu, L.; Wang, Z.; and Liu, H. 2024. Exploring Self- and Cross-Triplet Correlations for Human-Object Interaction Detection. arXiv:2401.05676.

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*.

Kim, B.; Lee, J.; Kang, J.; Kim, E.; and Kim, H. J. 2021. HOTR: End-to-End Human-Object Interaction Detection With Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 74–83.

Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Lei, T.; Caba, F.; Chen, Q.; Jin, H.; Peng, Y.; and Liu, Y. 2023. Efficient Adaptive Human-Object Interaction Detection with Concept-guided Memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6480–6490.

Lei, T.; Yin, S.; Peng, Y.; and Liu, Y. 2024. Exploring Conditional Multi-Modal Prompts for Zero-shot HOI Detection.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Li, L.; Wei, J.; Wang, W.; and Yang, Y. 2023. Neural-Logic Human-Object Interaction Detection. In *Advances in Neural Information Processing Systems*, volume 36, 21158–21171.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. arXiv:2101.00190.

Liao, Y.; Zhang, A.; Lu, M.; Wang, Y.; Li, X.; and Liu, S. 2022. GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 20091–20100.

Liu, Y.; Xiong, Z.; Li, Y.; Lu, Y.; Tian, X.; and Zha, Z.-J. 2023a. Category-Stitch Learning for Union Domain Generalization. *ACM Trans. Multimedia Comput. Commun. Appl.*

Liu, Y.; Xiong, Z.; Li, Y.; Tian, X.; and Zha, Z.-J. 2023b. Domain Generalization Via Encoding and Resampling in a Unified Latent Space. *IEEE Transactions on Multimedia*, 25: 126–139.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt Distribution Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5206–5215.

Ning, S.; Qiu, L.; Liu, Y.; and He, X. 2023. HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,bCVPR 2023*, 23507–23517.

Park, J.; Lee, S.; Heo, H.; Choi, H. K.; and Kim, H. J. 2022. 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, CVPR 2022*.

Part et al. (2023) Park, J.; Park, J.-W.; and Lee, J.-S. 2023. ViPLO: Vision Transformer Based Pose-Conditioned Self-Loop Graph for Human-Object Interaction Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17152–17162.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139, 8748–8763.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tamura et al.(2021) Masato, T.; Hiroki, O.; and Tomoaki, Y. 2021. QPIC: Query-Based Pairwise Human-Object Interaction Detection With Image-Wide Contextual Information. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 10410–10419.

Wang, G.; Guo, Y.; Xu, Z.; and Kankanhalli, M. 2024. Bilateral Adaptation for Human-Object Interaction Detection with Occlusion-Robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27970–27980.

Williams, C.; and Barber, D. 1998. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12): 1342–1351.

Xie, C.; Zeng, F.; Hu, Y.; Liang, S.; and Wei, Y. 2023. Category Query Learning for Human-Object Interaction Classification. In *CVPR*.

Yang, J.; Li, B.; Zeng, A.; Zhang, L.; and Zhang, R. 2024. Open-World Human-Object Interaction Detection via Multimodal Prompts. *arXiv preprint arXiv:2406.07221*.

Yuan, H.; Zhang, S.; Wang, X.; Albanie, S.; Pan, Y.; Feng, T.; Jiang, J.; Ni, D.; Zhang, Y.; and Zhao, D. 2023. RLIPv2: Fast Scaling of Relational Language-Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 21649–21661.

Zhang, F. Z.; Yuan, Y.; Campbell, D.; Zhong, Z.; and Gould, S. 2023. Exploring Predicate Visual Context in Detecting of Human-Object Interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10411–10421.

Zhang, Y.; Gong, M.; Liu, T.; Niu, G.; Tian, X.; Han, B.; Scholkopf, B.; and Zhang, K. 2022. Adversarial Robustness through the Lens of Causality. *ArXiv*, abs/2106.06196.

Zhang et al. (2022) Zhang, F. Z.; Campbell, D.; and Gould, S. 2022. Efficient Two-Stage Detection of Human-Object Interactions with a Novel Unary-Pairwise Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20104–20112.

Zhao, B. N.; Xiao, Y.; Xu, J.; Jiang, X.; Yang, Y.; Li, D.; Itti, L.; Vineet, V.; and Ge, Y. 2023. DreamDistribution: Prompt Distribution Learning for Text-to-Image Diffusion Models. arXiv:2312.14216.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision (IJCV)*.