X-Pose: Detecting Any Keypoints

Jie Yang[1,2], Ailing Zeng[1,★], Ruimao Zhang[2,★], and Lei Zhang[1]

[1]IDEA   [2]The Chinese University of Hong Kong, Shenzhen
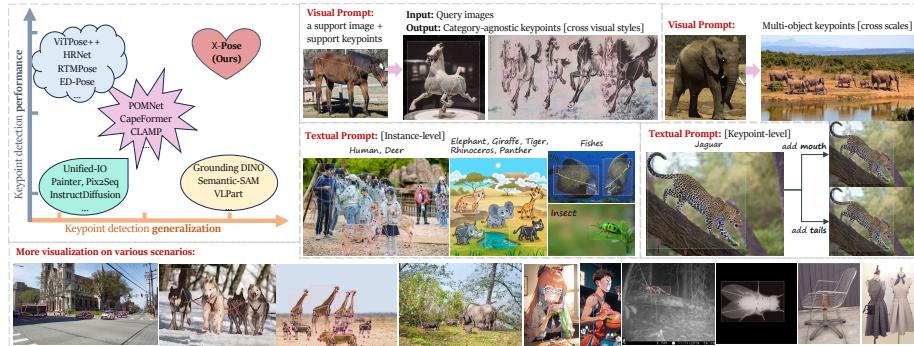https://github.com/IDEA-Research/X-Pose

Fig. 1: The proposed X-Pose achieves both strong keypoint detection generalization ability and high performance. X-Pose utilizes visual and textual prompts for training to learn fine-grained local-region visual representation via keypoint-text and keypoint-image alignments. Once trained, it can generalize cross-object and cross-keypoint categories, where it can detect multi-object keypoints on various challenging scenarios with diverse visual styles, scales, and poses.

Abstract. This work aims to address an advanced keypoint detection problem: how to accurately detect **any** keypoints in complex real-world scenarios, which involves massive, messy, and open-ended objects as well as their associated keypoints definitions. Current high-performance keypoint detectors often fail to tackle this problem due to their two-stage schemes, under-explored prompt designs, and limited training data. To bridge the gap, we propose X-Pose, a novel end-to-end framework with multi-modal (i.e., visual, textual, or their combinations) prompts to detect multi-object keypoints for any articulated (e.g., human and animal), rigid, and soft objects within a given image. Moreover, we introduce a large-scale dataset called UniKPT, which unifies 13 keypoint detection datasets with 338 keypoints across $1,237$ categories over 400K instances. Training with UniKPT, X-Pose effectively aligns text-to-keypoint and image-to-keypoint due to the mutual enhancement of multi-modal prompts based on cross-modality contrastive learning. Our experimental results demonstrate that X-Pose achieves notable improvements of 27.7 AP, 6.44 PCK, and 7.0 AP compared to state-of-the-art non-promptable, visual prompt-based, and textual prompt-based methods in each respective fair setting. More importantly, the in-the-wild test demonstrates X-Pose's strong fine-grained keypoint localization and

---

★Corresponding author.

generalization abilities across image styles, object categories, and poses, paving a new path to multi-object keypoint detection in real applications.

Keywords: Multi-object Keypoint Detection · Any Pose Estimation

1   Introduction

Fig. 2: In-the-wild test of X-Pose for any keypoint detection. We highlight the powerful detection performance from cross-category (the first row), multi-object (the second row), and cross-image-style (the third row) with various pose scenarios.

Multi-object keypoint detection, also known as multi-object pose estimation, stands as a fundamental computer vision task with board applications in VR/AR, biomedicine, and robots. It aims to estimate the 2D keypoint positions of various objects within an image, as shown in Fig. 1 and Fig. 2. With the rapid development of research areas, technologies are gradually transitioning from closed-set to open-world scenarios, leading to the formulation of three model paradigms: non-promptable model [4, 15, 54, 60], visual prompt-based model [6, 12, 29, 43, 53] and textual prompt-based model [68]. Although numerous models have been proposed, detecting any keypoints in complex real-world scenes remains a significant challenge due to the following difficulties:

The Bottleneck of Two-stage Schemes. Current high-performance non-promptable and promptable keypoint detectors all adopt two-stage strategies (e.g., top-down). Given an image containing multiple objects, they require an additional object detector to obtain object bounding boxes. Following this, they can crop the original image into single-object images and proceed with single-object keypoint detection. Consequently, the effectiveness of these models heavily relies on the performance of the object detector. Despite utilizing state-of-the-art object detectors [28,67], challenges persist, including inaccurate bounding boxes, missed detections, and excessive redundant bounding boxes due to inaccurate confidence scores, making them less reliable for real-world applications.

Under-explored Prompt Designs. Existing promptable keypoint detectors focus on single-object scenes with only keypoint-level prompts, overlooking object-

Table 1: Representative Model Comparisons of different paradigms. TD and E2E are the top-down and end-to-end methods. N is the object number in the image. **T.** and **V.** in the first row mean Textual Prompts and Visual Prompts, respectively.

Methods	T.	V.	Multi-object	Multi-class	Keypoints	Object	Training Images	Time [ms]
Non-promptable Methods								
ViTPose [54] (TD)	✗	✗	✗	✗	17	1	58K	$60 \times N$
ED-Pose [60] (E2E)	✗	✗	✓	✗	17	1	58K	55
Promptable Methods								
Capeformer [43] (TD)	✗	✓	✗	✗	293	100	17K	$57 \times N$
CLAMP [68] (TD)	✓	✗	✗	✗	14	54	10K	$63 \times N$
X-Pose (E2E)	✓	✓	✓	✓	**338**	**1237**	**226K**	59

level prompts. As a result, when faced with unseen objects with significant variations in appearance and different keypoint definitions, the model's ability to generalize at the keypoint level will be severely limited. Furthermore, these methods only accommodate a single type of prompt, either textual or visual, making user interactions unfriendly and inefficient.

Limited Training Data. Existing textual-prompt keypoint detectors [68] are trained using only 17 keypoint descriptions for animal pose estimation [65]. Similarly, present visual-prompt keypoint detectors [43, 53] rely on a small-scale dataset [53] (e.g., only 20K images with 100 instance classes). The constrained nature of their training data significantly hampers generalizability and effectiveness. Currently, there is no dataset encompassing a broader range of object categories, a more extensive set of keypoint categories, and a larger quantity of data. Organizing such a dataset would facilitate the learning of structured keypoint representations across various categories, thereby bolstering the models' generalization in real-world applications.

Considering the above challenges and motivations, we present X-Pose, a fully end-to-end framework that leverages multi-modal prompts (*i.e.,* textual, visual, or their combinations) to detect multi-object keypoints in complex real-world scenarios. **Firstly**, inspired by the DETR-like non-promptable human pose estimator [60], which integrates human-level and keypoint-level detection into an end-to-end framework, we extend this framework by incorporating the prompt mechanism to support any objects and keypoints. Such a scheme could dynamically adjust the object categories to be detected and accommodate diverse keypoint definitions, making the model effectively generalize across multi-class and multi-object scenarios in an open-world setting. **Secondly**, X-Pose is the first model to leverage multi-modal prompts for multi-object keypoint detection. During training, incorporating textual prompts into visual prompt-based keypoint detection tasks provides valuable high-level semantic guidance, enhancing the learning process. Similarly, visual prompts can complement textual prompts by providing detailed image-level information, thereby improving the effectiveness of both tasks. Moreover, during inference, a prompt design that supports textual, visual, and combination prompts can significantly enhance user experience and efficiency. **Lastly**, for effectively training X-Pose, we present UniKPT, a large-scale dataset that unifies 13 keypoint detection datasets with 338 keypoints across 1,237 categories over 400K instances. In UniKPT, we balance these

datasets by considering image appearance and style diversity, instances with varying poses, viewpoints, visibilities, and scales. Additionally, we reconstruct the semantic relationships between all keypoints and categories, while also standardizing the definitions of the same keypoints across categories (e.g., the left eye of different animals). Most importantly, each keypoint is meticulously annotated with a unique name to enhance keypoint-level semantic understanding.

Based on the above technical and data contributions, we show the functionalities supported and efficiency comparison with the previous representative models in Tab. 1. Furthermore, through comprehensive experiments, we demonstrate the remarkable generalization capabilities of X-Pose for both visual prompt-based and textual prompt-based keypoint detection. Compared to state-of-the-art methods, it achieves notable improvements of **6.44** PCK and **7.0** AP. Moreover, X-Pose significantly outperforms the state-of-the-art end-to-end model, particularly achieving a **27.7** AP improvement in AP-10K [65]. Its performance is also comparable to state-of-the-art results on all existing keypoint datasets. Additionally, X-Pose exhibits impressive text-to-image alignment at both object and keypoint levels, surpassing CLIP by **204**% when distinguishing between different animal categories and by **166**% across various image styles. From extensive qualitative results on in-the-wild images, we showcase the open-world keypoint detection performance and generalization ability of X-Pose, hoping it could benefit fine-grained visual perception and understanding.

2   Related Work

This section introduces three related areas, including non-promptable, visual prompt-based, and textual prompt-based keypoint detection. As summarized in Tab. 1, our X-Pose is the **first** end-to-end multi-modal prompt-based model, which could effectively and efficiently detect any keypoints in complex real-world scenarios involving multi-classes and multi-objects as well as their associated keypoints definitions.

2.1   Non-Promptable Keypoint Detection

Existing non-promptable methods have mainly focused on human or animal pose estimation [15, 31, 34, 46, 54, 55, 60, 64, 70], categorized into two-stage and one-stage paradigms. Among the two-stage methods, the top-down strategies dominate and demonstrate high performance [8, 23, 30, 45, 52, 54] by first detecting each object in an image using an independent object detector and then solely focusing on single-object keypoint detection with the proposed model. However, the use of separate object detectors and multiple inferences for each object incurs high computational costs. Moreover, any missed object detections directly lead to failures in keypoint detection. Recently, one-stage methods [27, 42, 59, 60] have proposed to detect multi-person keypoints in an end-to-end manner, which have shown superior performance and efficiency trade-offs. However, these methods only focus on single-class objects with specific pre-defined keypoints, limiting

their applicability and generality in real-world scenes. In light of these, our work aims to provide an end-to-end keypoint detector with strong keypoint generalization to detect any keypoints in complex real-world scenes.

2.2   Visual Prompt-based Keypoint Detection

Given a prompt image of a novel object and its corresponding keypoint definitions, visual prompt-based keypoint detection aims to detect the keypoints of the same object within an image. Existing methods [6, 12, 20, 29, 43, 46, 53, 64] typically focus on single-object scenes, simplifying this problem. Therefore, these methods are unable to address multi-class multi-object scenarios without known object detection, particularly situations where an image contains numerous objects of different categories with varying keypoint definitions. Moreover, most of them only consider one super-category, such as clothing or animal. To handle more objects, recent works [43, 53] train their models on the MP-100 dataset with 17K images spanning 100 categories. However, such a small-scale dataset makes them suffer from under-fitting and hard to learn the local keypoint representation effectively. To address the above problems, we introduce an end-to-end model that can leverage visual prompts to detect multi-object keypoints. For effective training, we unify the existing 13 datasets to generalize the model across the object and their keypoints.

2.3   Textual Prompt-based Keypoint Detection

Benefit from vision-language pretrained model CLIP [36], like open-vocabulary object detection and semantic segmentation tasks are actively explored [10, 21, 24, 25, 28, 47, 63, 66, 69]. In the field of keypoint detection, CLAMP [68] is the first work to leverage CLIP with language guidance to prompt the animal keypoints containing a fixed keypoint set (e.g., 20). CLAMP's primary emphasis is cross-species generalization within a predefined skeleton structure. However, its limited support for keypoint descriptions and the two-stage paradigm restrict its effectiveness. In this work, X-Pose introduces a large-scale dataset annotated with more keypoint names. Trained on such a dataset, we offer an end-to-end model that can leverage textual prompts to detect multi-object keypoints. Moreover, we are the first to explore multi-modal prompt-based keypoint detection.

3   Method

As shown in Fig. 3, X-Pose is an end-to-end multi-modal prompt-based keypoint detection framework. It takes an image accompanied by textual or visual prompts as input and outputs all the object bounding boxes and the corresponding keypoints. In the following subsection, we first introduce how to encode multi-modal inputs and enhance each other in Sec. 3.1. Then, we illustrate how to end-to-end decode the prompt-oriented information, including the desired object bounding boxes and keypoints in Sec. 3.2. Finally, we provide the training loss function and inference pipeline in Sec. 3.3.
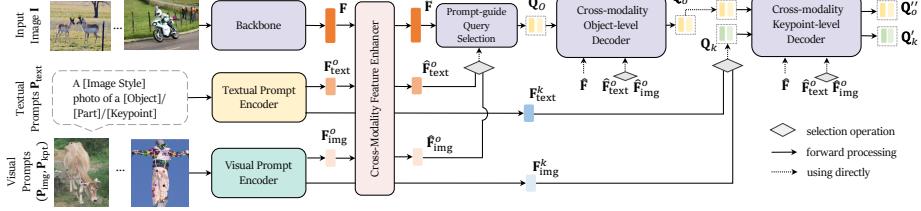
Fig. 3: The overview architecture of X-Pose. Given an input image, X-Pose follows the coarse-to-fine strategy to detect keypoints of any object via textual or visual prompts.
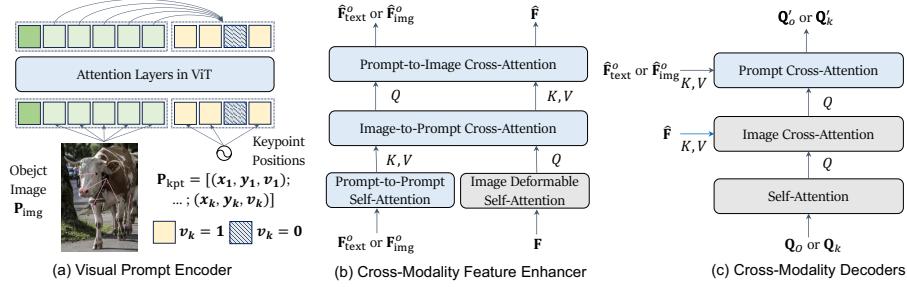

Fig. 4: The detailed illustration of a) Visual Prompt Encoder, b) Cross-Modality Interactive Encoder, and c) Cross-Modality Interactive Decoder. In (b) and (c), blue modules are newly introduced to incorporate prompt interactions.

3.1   Multi-Modality Inputs Encoding and Enhancing

X-Pose supports three kinds of inputs: input image $\mathbf{I}$, textual prompts $\mathbf{P}_{\text{text}}$, and visual prompts $(\mathbf{P}_{\text{img}}, \mathbf{P}_{\text{kpt}})$, where $\mathbf{P}_{\text{img}}$ is an image of object and $\mathbf{P}_{\text{kpt}}$ is 2D positions of defined keypoints. The input image is processed through a backbone network, to extract multi-scale and tokenized features $\mathbf{F}$. For multimodal prompts, we introduce two novel encoding mechanisms as follows:

Textual Prompts are encoded by the CLIP text encoder to produce object-level and keypoint-level textual embeddings $\mathbf{F}_{\text{text}}^{o}$ and $\mathbf{F}_{\text{text}}^{k}$, respectively. In particular, we formulate the textual prompt $\mathbf{P}_{\text{text}}$ as a hierarchical structure to describe objects, parts (e.g., face and hand), and keypoints. The template can be written as: "An [image style] photo of a [object]/[part]/[keypoint]".

Visual Prompts are processed through the CLIP image encoder to acquire visual embeddings for objects $\mathbf{F}_{\text{img}}^{o}$ and keypoints $\mathbf{F}_{\text{img}}^{k}$. As shown in Fig. 4-(a), while the original CLIP image encoder accepts only images as input, X-Pose extends its capability to incorporate keypoints for position encodings. **Firstly**, given a specific keypoint (x, y, v), where (x, y) denotes its 2D coordinate and v indicates the visibility of the keypoint, we introduce two token initialization methods for different keypoints: i) for visible keypoints ($v{=}1$), we use the Fourier embedding [32] to map the 2D coordinate to the corresponding feature dimensions; ii) for invisible keypints ($v{=}0$), we employ a shared learnable mask token [11] to represent the invisible position. **Secondly**, since the initialized keypoint tokens solely contain position information, we adopt two encoding mechanisms: i)

the "keypoint token to keypoint token" attention to capture skeletal structure relations; ii) the "image patch token to keypoint token" attention to propagate global image feature into each keypoint token.

Cross-Modality Enhancing. As in Fig. 4-(b), after obtaining the image features $\mathbf{F}$ and prompt features $\mathbf{F}^o_{\text{text}}$, $\mathbf{F}^o_{\text{img}}$, we employ the Deformable self-attention [71] to enhance image features and the vanilla self-attention [48] for enhancing prompt features. Moreover, we leverage an image-to-prompt cross-attention and a prompt-to-image cross-attention for cross-modality enhancement.

3.2   Cross-Modality Object and Keypoint Decoding

The decoders of X-Pose are decoupled into the object-level decoder and the keypoint-level decoder. Leveraging the enhanced image features and multi-modal prompt features in Sec. 3.1, X-Pose initializes the object queries $\mathbf{Q}_o$ and keypoint queries $\mathbf{Q}_k$ to decode all the objects $\mathbf{Q}''_o$ with their associated keypoints $\mathbf{Q}'_k$.

Object Decoding. Firstly, we utilize prompt-guided query selection [28,67] to initialize object queries $\mathbf{Q}_o$ from the enhance image features $\widehat{\mathbf{F}}$, which is highly associated with the enhanced object-level prompt features $\widehat{\mathbf{F}}^o_{\text{text}}$ or $\widehat{\mathbf{F}}^o_{\text{img}}$. Then, a cross-modality object decoder is employed to update these object queries to $\mathbf{Q}'_o$. Illustrated in Fig. 4-(c), object queries are inputted into a self-attention layer, an image cross-attention layer to integrate image features, and a prompt cross-attention layer to integrate prompt features.

Keypoint Decoding. The keypoint queries $\mathbf{Q}_k$ are directly initialized by keypoint level prompt features $\widehat{\mathbf{F}}^k_{\text{text}}$ or $\widehat{\mathbf{F}}^k_{\text{img}}$. Then, a cross-modality keypoint decoder is employed to update these keypoint queries to $\mathbf{Q}'_k$. Similar to object decoding, keypoint queries are processed through a self-attention layer, an image cross-attention layer to combine image features, and a prompt cross-attention layer to combine prompt features.

3.3   Training and Inference Pipeline

We adopt the same object and keypoint regression losses as previous end-to-end keypoint detectors [42,59,60]: the L1 loss and the GIOU loss [40] for object's bounding box regression $\mathcal{L}^{obj}_{reg}$; the L1 loss and the OKS loss [42] for keypoint regression $\mathcal{L}^{kpt}_{reg}$. Moving forward, X-Pose introduces prompt-to-object and prompt-to-keypoint contrastive losses for alignments.

Object-level Alignment. Previous keypoint detectors [45,54,60] mainly focus on close-set objects and typically use a simple linear layer as the object classifier. In contrast, X-Pose encodes multi-modal prompts (text or image) into the object-level prompt features $\widehat{\mathbf{F}}^o_{\text{text}}, \widehat{\mathbf{F}}^o_{\text{img}} \in \mathbb{R}^{L \times C}$, where L is the number of object classes in prompts and C indicates the feature dimension. Following [24,28], we employ contrastive loss between predicted objects $\mathbf{Q}''_o$ and prompt features for classification. More specifically, we compute the dot product between each object query and the prompt features to predict logits and then calculate the Focal loss of each logit $\mathcal{L}^{obj}_{align}$ for optimization.

Table 2: Statistics of UniKPT with 13 keypoint datasets.

Datasets	Keypoints	Class	Images	Instances	Unified Images	Unified Instances
COCO [26]	17	1	58,945	156,165	58,945	156,165
300W-Face [41]	68	1	3,837	4,437	3,837	4,437
OneHand10K [50]	21	1	11,703	11,289	2,000	2000
Human-Art [16]	17	1	50,000	123,131	50,000	123,131
AP-10K [65]	17	54	10,015	13,028	10,015	13,028
APT-36K [62]	17	30	36,000	53,006	36,000	53,006
MacaquePose [19]	17	1	13,083	16,393	2,000	2,320
Animal Kingdom [34]	23	850	33,099	33,099	33,099	33,099
AnimalWeb [17]	9	332	22,451	21,921	22,451	21,921
Vinegar Fly [35]	31	1	1,500	1,500	1,500	1,500
Desert Locust [9]	34	1	700	700	700	700
Keypoint-5 [51]	55/31[1]	5	8,649	8,649	2,000	2,000
MP-100 [53]	561/293[1]	100	16,943	18,000	16,943	18,000
UniKPT	338	1237	-	-	226,547	418,487

[1] Keypoint-5 and MP-100 have different categories with varying numbers of keypoints. While the cumulative count of keypoints reaches 55 and 561 by aggregating across categories, we consolidate them into unified counts of 31 and 293 keypoints by leveraging textual descriptions.
[2] MP-100 includes training subsets from two other datasets, Deepfashion2 [7] and Carfusion [38].

Keypoint-level Alignment. In previous keypoint detectors, the classification problem for keypoints is often overlooked, and the learning process is to establish a one-to-one mapping between predicted and labeled keypoints. In contrast, X-Pose takes the first step toward prompts-to-keypoint alignment using a unified set of keypoint definitions. Similar to object-level alignment, given keypoint prompt features $\widehat{\mathbf{F}}_{\text{text}}^k, \widehat{\mathbf{F}}_{\text{img}}^k \in \mathbb{R}^{K \times C}$, where K is the number of keypoint categories in prompts. We utilize contrastive loss between predicted keypoints $\mathbf{Q}_k'$ and prompt features for classification.

The Overall Loss. The overall training pipeline of X-Pose can be written as follows,

$$\mathcal{L} = \mathcal{L}_{reg}^{obj} + \mathcal{L}_{reg}^{kpt} + \mathcal{L}_{align}^{obj} + \mathcal{L}_{align}^{kpt} \tag{1}$$

Training & Inference Details. During training, we employ a 50% probability to randomly select either a visual prompt or textual prompt for each iteration. We sample two images containing the same object category from our UniKPT dataset and then choose one as the visual prompt. During inference, 1) Textual Prompts: We can utilize pre-defined object classes with keypoints definitions as text prompts to obtain quantitative results. In practical scenarios, users can provide the text to predict the desired objects with keypoints or any keypoint. 2) Visual Prompt: We can randomly sample a set of image prompts from the training data to obtain quantitative results. In practical scenarios, users can provide a single object image (1-shot) with the corresponding keypoint definition to predict all the similar objects in the test images.

4   UniKPT: A Unified Keypoint Dataset

Unifying 13 Keypoint Datasets into UniKPT. As summarized in Tab. 2, we observe that each dataset only focuses on a single super-category (e.g., "human only" and "animal only"), making it challenging to achieve keypoint generalization when using them individually. Additionally, these datasets have significant differences in quality, quantity, and appearance styles. Motivated by these,

we propose to unify 13 existing keypoint detection datasets into UniKPT. which addresses several crucial aspects: 1) Balanced Diversity: We ensure a balance across the 13 datasets by considering diverse factors such as image appearance, style, poses, viewpoints, visibilities, and scales. 2) Semantic Relationships: We reconstruct the semantic relationships between all keypoints and categories in the 13 datasets. 3) Standardized Definitions: We standardize the definitions of the same keypoints across different categories. For example, the left eye of various animals would be consistently defined. 4) **Enhanced Annotations**: Each keypoint within UniKPT has been meticulously annotated with a unique name to enhance keypoint-level semantic understanding.

Statistical Analysis. In total, the unified dataset comprises $226,547$ images and $418,487$ instances, featuring 338 keypoints and $1,237$ instance categories. In particular, for articulated objects like humans and animals, we further categorize them based on biological taxonomy, resulting in $1,216$ species, 66 families, 23 orders, and 7 classes.

5 Experiment

5.1 Experimental Setup

Dataset. We follow the fair and standard benchmarks: 1) MP-100 [53] for visual prompt-based keypoint detection in Sec. 5.3; 2) AP-10K [65] for textual prompt-based keypoint detection in Sec. 5.4; 3) UniKPT for general keypoint detection in Sec. 5.5. Please refer to the Appendix for more details about each dataset.

Implementation details. 1) Network Details. We use 6-layer cross-modality feature enhancer, 2-layer cross-modality object decoder, and 4-layer cross-modality keypoint decoder. We adopt the CLIP model with the ViT-Base network for visual and textual prompt encoding. The feature dimension is set to 256. 2) Training & Inference Details. We use the exact same training details as all the end-to-end models [42,60]. Specifically, we augment the training images through random cropping, flipping, and resizing. The shorter sides are kept within $[480, 800]$, while the longer sides are less than or equal to 1333. The size of the prompt image is set to 224, aligning with the requirement of CLIP. We utilize the AdamW optimizer with a weight decay of 1e-4. Our models are trained on 8 Nvidia A100 GPUs with a batch size of 16. During inference, the images are resized with shorter sides of 800 and longer sides less than or equal to 1333.

5.2 Qualitative In-the-wild Test

As shown in Fig. 2, we show the powerful detection performance of X-Pose in real-world scenarios, which could end-to-end address the challenges of cross-category (the first row), multi-object (the second row), and cross-image-style (the third row). Furthermore, Fig. 5 presents a surprising observation: despite being trained on a limited dataset containing human faces with the 68 keypoints defined by [41], X-Pose demonstrates remarkable cross-object capabilities when tasked with any face detection.

Fig. 5: In-the-wild test of X-Pose for any face keypoint detection. We showcase the model's strong generalization to detect face keypoints of any object with 68 keypoint definitions, despite being trained only on the person's face with these definitions.

5.3   Visual Prompt-based Keypoint Detection

We evaluate X-Pose against the previous methods, such as ProtoNet [44], MAML [5], Fine-tune [33], POMNet [53], and Capeformer [43] on the MP-100 dataset, as shown in Tab. 3. As an end-to-end framework, X-Pose offers efficiency by requiring only a single forward pass for scenes with multiple objects. Surprisingly, it outperforms all top-down methods, establishing a new state-of-the-art performance. This achievement can be attributed to X-Pose's utilization of instance-level visual prompt information, which enhances keypoint-level generalization, especially in such scenarios involving unseen objects with significant variations in appearance and different keypoint definitions.

5.4   Textual Prompt-based Keypoint Detection

To compare with the previous textual prompt-based model-CLAMP [68], we follow its zero-shot setting to evaluate the model's generalization ability on unseen animal species, using AP-10K. In particular, we select the Bovidae or Canidae animal orders as the training set and the unseen orders Canidae and Felidae as the testing set. The results are shown in Tab. 4. Compared to CLAMP, X-Pose achieves much better performance in both settings, e.g., there is a 6.9 AP and 7.0 AP increase in these two settings, respectively. Moreover, X-Pose as an end-to-end model has greater efficiency when dealing with multi-object scenes.

Table 3: Comparisons of visual prompt-based keypoint detection on the MP-100 benchmark. † means the model needs keypoint identifiers during the test, however, which is not available in real-world scenarios. The inference times for all methods are tested on an A100 with a batch size of 1. Top-down methods need multiple inferences when N objects are detected in an image. The compared <u>results</u> are highlighted.

	Method	Backbone	Split1	Split2	Split3	Split4	Split5	Mean (PCK)	Time [ms]
TD	ProtoNet	ResNet-50	46.05	40.84	49.13	43.34	44.54	44.78	-
	MAML	ResNet-50	68.14	54.72	64.19	63.24	57.20	61.50	-
	Fine-tune	ResNet-50	70.60	57.04	66.06	65.00	59.20	63.58	-
	POMNet	ResNet-50	84.23	78.25	78.17	78.68	79.17	<u>79.70</u>	$151\times N$
	CapeFormer†	ResNet-50	**89.45**	84.88	83.59	83.53	85.09	85.31	$57\times N$
	CapeFormer	ResNet-50	<u>85.81</u>	-	-	-	-	-	$57\times N$
E2E	X-Pose	ResNet-50	89.07$_{\uparrow 3.26}$	**85.05**	**85.26**	**85.52**	**85.79**	**86.14**$_{\uparrow 6.44}$	**59**

Note: We train our models only on the MP-100 dataset to ensure a fair comparison. During evaluation, all methods use the same visual prompts paired with test images.

Table 4: Comparisons of textual prompt-based keypoint detection on AP-10K. Following CLAMP [68]'s zero-shot setting, we train the model only on Bovidae or Canidae animal order and test it on unseen Canidae or Felidae animal order, respectively.

Methods	Backbone	Train	Test	AP	AP_M	AP_L	Time [ms]
CLAMP (TD)	ResNet50	Bovidae	Canidae	46.9	30.3	46.9	$63 \times N$
X-Pose (E2E)	ResNet50	Bovidae	Canidae	**53.8**$_{\uparrow 6.9}$	**34.7**	**54.1**	**59**
CLAMP (TD)	ResNet50	Canidae	Felidae	48.4	13.6	48.9	$63 \times N$
X-Pose (E2E)	ResNet50	Canidae	Felidae	**55.4**$_{\uparrow 7.0}$	**20.8**	**54.2**	**59**

5.5 General Keypoint Detection

Comparison with the SOTA on Various Keypoint Datasets. Existing state-of-the-art results are achieved by non-promptable models with two-stage schemes. Compared to them, X-Pose possesses a unique advantage in superior generalization to handle unseen objects with different keypoint definitions. Furthermore, as shown in Tab. 5, X-Pose surprisingly achieves new state-of-the-art performance on most datasets. Despite slightly lower performance on some datasets like COCO, X-Pose demonstrates for the first time the performance upper ceiling of an end-to-end model.

Comparison with the SOTA End-to-End Model. For a fair comparison, we train both our X-Pose and ED-Pose using the same datasets, *i.e.*, COCO, Human-Art, AP-10K, and APT-36K. The complete results are shown in the Appendix. Tab. 6 highlights the performance comparison on AP-10K, which involves the classification of 54 different species. X-Pose surpasses ED-Pose with a 27.7 AP improvement, thanks to instance-level and keypoint-level alignments.

Qualitative Results on Existing Datasets. Given an input image and textual prompts, X-Pose shows powerful qualitative results across existing closed datasets, encompassing articulated, rigid, and soft objects, as shown in Fig. 6.

5.6 Compared with Open-Vocabulary Models

Comparison with the Vision-Language Model. We assess X-Pose's text-to-image alignment capabilities at different granularities, i.e., object and keypoint. In Tab. 7, we report the CLIP score of X-Pose and CLIP [37] on AP-10K,

Table 5: Comparison with absolute SOTA results on all existing keypoint datasets. † indicates results using the flipping test. Results marked with * rely on ground-truth bounding boxes for top-down methods. The **best** results are highlighted in **bold**, and the second best results are highlighted with a underline. T and V denote textual and visual prompts used.

Methods	Backbone	COCO	AP-10K	Human-Art AP↑	Macaque	300W	Hand	AK	Fly	Locust PCK↑	KPT-5	DF2	Carfusion
SOTA (TD)	-	**78.6**†	**80.4***†	35.6	51.9*	**99.8***	99.5*	-	-	-	-	-	-
X-Pose-T (E2E)	Swin-T	74.4	74.0	72.5	78.0	98.1	95.7	95.3	99.6	99.7	94.3	95.7	78.1
X-Pose-V (E2E)	Swin-T	74.3	73.6	72.1	77.3	99.4	95.9	94.3	99.8	99.6	87.4	91.0	72.1
X-Pose-T (E2E)	Swin-L	76.8	79.2	**75.9**	79.4	98.5	**99.8**	**96.1**	**99.9**	99.8	**95.5**	**97.5**	**88.7**
X-Pose-V (E2E)	Swin-L	76.6	79.0	75.5	77.8	99.3	99.5	95.5	**99.9**	**99.9**	91.6	95.5	85.0

Fig. 6: Visualization of the detected keypoints via X-Pose on UniKPT.

which involves 54 animal categories, and Human-Art, which features 15 image styles. Results show that X-Pose consistently provides higher-quality text-to-image similarity scores at the object and keypoint levels.

Comparison with Open-Vocabulary Detection Model. We compare X-Pose with the state-of-the-art open-vocabulary object detector, GroundingDINO [28], in terms of instance-level and keypoint-level detection. We present the COCO results in Tab. 8, while results for other datasets are provided in the Appendix. Grounding-DINO fails to localize fine-grained keypoints; however, X-Pose successfully addresses these challenges, achieving significant improvements across all datasets. X-Pose maintains comparable performance with GroundingDINO. Additionally, we find that although fine-tuning GroundingDINO for instance detection can be beneficial, it negatively impacts keypoint detection.

5.7 Ablation Study

In the first two ablation studies, we train X-Pose with the Swin-T backbone on four datasets: COCO, Human-Art, AP-10K, and APT36K. For fair comparisons, we report the results on AP-10K, which enables comprehensive evaluation in classification and localization. In the third ablation study, we present the results on both the seen dataset AP-10K in UniKPT and the unseen dataset AnimalPose [1] to demonstrate generalization ability.

Contrastive Loss. We introduce $\mathcal{L}_{Align}^{obj}$ and $\mathcal{L}_{Align}^{kpt}$ to facilitate prompt-to-object and prompt-to-keypoint alignment, respectively, as in Sec. 3.3. We present the results using textual prompts in Tab. 9, highlighting the significant improvement in detection performance, particularly in AP_L, due to $\mathcal{L}_{Align}^{obj}$. This under-

Table 6: Comparisons with the end-to-end non-prompted model on AP-10K. We train ED-Pose with the same dataset as X-Pose.

Methods	Backbone	AP	AP_M	AP_L
ED-Pose	Swin-T	45.5	31.0	46.5
X-Pose-V	Swin-T	72.8	**47.2**	74.0
X-Pose-T	Swin-T	**73.2**$_{\uparrow 27.7}$	45.6	**74.3**

Table 7: Comparisons of CLIP score.

Methods	AP-10K `val`		Human-Art `val`	
	Instance	Keypoint	Instance	Keypoint
CLIP	28.36	21.75	23.60	23.81
X-Pose	**58.59**$_{\uparrow 106\%}$	**66.01**$_{\uparrow 204\%}$	**68.41**$_{\uparrow 190\%}$	**63.46**$_{\uparrow 166\%}$

Table 8: Comparisons with the state-of-the-art open-vocabulary object detector, focusing on instance-level and keypoint-level detection. ‡ denotes the fine-tuning of GroundingDINO using the keypoint detection datasets. Note that we limit the instance-level comparison to AP_M (medium objects) and AP_L (large objects), as small objects do not have keypoints annotated.

Methods	Backbone	Instance-level		Keypoint-level			Training Datasets	Dataset Volume
		AP_M	AP_L	AP	AP_M	AP_L		
COCO `val` set								
GroundingDINO-T	Swin-T	70.8	82.0	3.1	2.8	3.2	O365,GoldG,Cap4M	1858K
GroundingDINO-T	Swin-B	69.7	79.5	6.8	6.6	7.2	COCO,O365,GoldG,Cap4M,OpenImage,ODinW-35,RefCOCO	-
GroundingDINO‡-T	Swin-T	**71.2**	**83.4**	1.8	1.7	1.9	COCO,Human-Art,AP-10K,APT-36K	1858K + 155K
X-Pose-T	Swin-T	71.1	80.2	**74.2**	**68.8**	**82.1**	COCO,Human-Art,AP-10K,APT-36K	155K
X-Pose-V	Swin-T	71.1	80.3	74.1	**68.8**	81.8	COCO,Human-Art,AP-10K,APT-36K	155K

scores its importance to benefit the model to distinguish between categories and enhance classification performance. The improved detection performance positively affects keypoint performance. Moreover, the inclusion of $\mathcal{L}_{Align}^{kpt}$ further helps the network learn keypoint distinctions, resulting in enhanced keypoint detection performance.

Multi-Modality Prompts. We explore whether the two modalities can benefit each other in Tab. 10. Single-modality training settings always perform worse than multi-modality settings, highlighting the mutual advantages of both textual and visual prompts.

Dataset Quantity in UniKPT. We first train our X-Pose using 4 datasets covering humans and 60 different animals. Then, we add extra 5 animal datasets to train X-Pose, as shown in Tab. 11. This results in significant improvements in both instance and keypoint detection on seen AP-10K datasets (using textual prompts). Moreover, we achieve a significant improvement on the unseen AnimalPose dataset (using visual prompts), thanks to more categories and the increased data size, making the model more generalizable. Furthermore, we incorporate additional part-level datasets (Face and Hand) as well as rigid and soft object datasets for training. Although these diverse datasets lead to a slight decrease in AP-10K performance, it further boosts the model's performance on unseen datasets.

5.8   The Analysis of Multi-modal Prompts

As in Fig. 7, we present a visualization comparison of using different prompts across AP-10K and Human-Art datasets. We make two key observations: i) Since AP-10K requires accurate animal category classification (e.g., from 54 categories), we find that the classification accuracy achieved through visual prompts
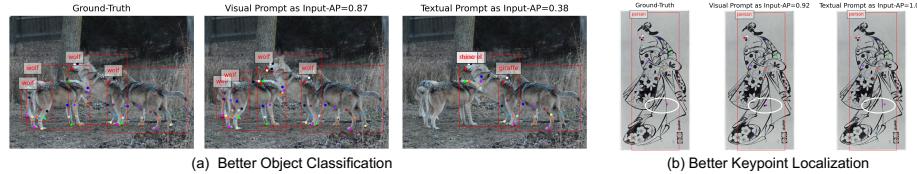
Table 9: Impact of constrastive loss on AP-10K.

$\mathcal{L}_{Align}^{obj}$	$\mathcal{L}_{Align}^{kpt}$	Object-level		Keypoint-level		
		AP$_M$	AP$_L$	AP	AP$_M$	AP$_L$
		53.7	62.5	45.5	31.0	46.5
✓		53.8	78.5	72.6	43.6	73.4
✓	✓	**54.5**	**78.8**	**73.2**	**45.6**	**74.3**

Table 10: Impact of multi-modal prompts on AP-10K. The prompt used in the test is highlighted in gray.

Visual Prompt	Textual Prompt	Object-level		Keypoint-level		
		AP$_M$	AP$_L$	AP	AP$_M$	AP$_L$
✓		53.3	78.1	71.5	43.4	72.4
✓	✓	**55.8**	**79.0**	72.8	**47.2**	74.0
	✓	53.8	78.5	72.9	45.1	74.2
✓	✓	54.5	78.8	**73.2**	45.6	**74.3**

Table 11: Impact of dataset quantity on AP-10K and AnimalPose.

Training Data	AP-10K's Object		AP-10K's Keypoint			AnimalPose
	AP$_M$	AP$_L$	AP	AP$_M$	AP$_L$	PCK
COCO,Human-Art,AP-10K,APT-36K	54.5	78.8	73.2	45.6	74.3	52.7
+MacquePose,AnimalKingdom,AnimalWeb,Vinegar Fly,Desert Locust	**55.6**	**80.2**	**74.2**	**48.3**	**75.0**	70.1
+300w-Face,OneHand10K,Keypoint-5,MP-100	55.3	78.8	74.0	47.8	74.7	**73.4**

surpasses that of textual prompts (see Fig. 7-(a)). This is attributed to the fact that visual prompts can provide a greater volume of similar instance features, enhancing the accuracy of classifications. ii) We notice that text prompts offer a slightly better keypoint localization accuracy compared to visual prompts (see Fig. 7-(b)). However, this difference is exceedingly small, primarily due to the effective cross-modality contrastive learning strategies, which significantly enhance keypoint localization accuracy.

Fig. 7: Prompts Analysis in X-Pose on (a) AP-10K [65] and (b) Human-Art [16].

6   Conclusion

This work studies the problem of detecting any keypoints in real-world scenes. To solve this problem, we proposed an end-to-end multi-modal prompt-based framework trained on a unified keypoint dataset to learn general semantic fine-grained keypoint concepts and global-to-local keypoint structure. The extensive experiments and in-the-wild tests demonstrate that X-Pose achieves high keypoint detection performance and generalizability in real-world scenes.

Broader Impact: Based on the proposed X-Pose, we can provide 1) an end-to-end keypoint detector for any keypoints to benefit various downstream areas [2, 56–58, 61]; 2) a user-friendly connector with either textual prompts or visual prompts to first detect keypoints and then take them as user clicks for fine-grained detection, segmentation, and tracking [18, 22, 39, 72]; 3) the proposed UniKPT dataset could benefit the training of large vision models and its keypoint-level semantic annotations could promote better fine-grained vision-language understanding [3, 13, 14, 49].

Acknowledgement

References

1. Cao, J., Tang, H., Fang, H.S., Shen, X., Lu, C., Tai, Y.W.: Cross-domain adaptation for animal pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9498–9507 (2019) 12
2. Chen, L.H., Lu, S., Zeng, A., Zhang, H., Wang, B., Zhang, R., Zhang, L.: Motionllm: Understanding human behaviors from human motions and videos. arXiv preprint arXiv:2405.20340 (2024) 14
3. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024) 14
4. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5386–5395 (2020) 2
5. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning. pp. 1126–1135. PMLR (2017) 10
6. Ge, Y., Zhang, R., Luo, P.: Metacloth: Learning unseen tasks of dense fashion landmark detection from a few samples. IEEE Transactions on Image Processing **31**, 1120–1133 (2021) 2, 5
7. Ge, Y., Zhang, R., Wang, X., Tang, X., Luo, P.: Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5337–5345 (2019) 8
8. Geng, Z., Wang, C., Wei, Y., Liu, Z., Li, H., Hu, H.: Human pose as compositional tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 660–671 (2023) 4
9. Graving, J.M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B.R., Couzin, I.D.: Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. Elife **8**, e47994 (2019) 8
10. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021) 5
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) 6

12. He, X., Bharaj, G., Ferman, D., Rhodin, H., Garrido, P.: Few-shot geometry-aware keypoint localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21337–21348 (2023) 2, 5

13. Jiang, Q., Li, F., Ren, T., Liu, S., Zeng, Z., Yu, K., Zhang, L.: T-rex: Counting by visual prompting. arXiv preprint arXiv:2311.13596 (2023) 14

14. Jiang, Q., Li, F., Zeng, Z., Ren, T., Liu, S., Zhang, L.: T-rex2: Towards generic object detection via text-visual prompt synergy. arXiv preprint arXiv:2403.14610 (2024) 14

15. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: Rtm-pose: Real-time multi-person pose estimation based on mmpose. arXiv preprint arXiv:2303.07399 (2023) 2, 4

16. Ju, X., Zeng, A., Wang, J., Xu, Q., Zhang, L.: Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 618–629 (2023) 8, 14

17. Khan, M.H., McDonagh, J., Khan, S., Shahabuddin, M., Arora, A., Khan, F.S., Shao, L., Tzimiropoulos, G.: Animalweb: A large-scale hierarchical dataset of annotated animal faces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6939–6948 (2020) 8

18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023) 14

19. Labuguen, R., Matsumoto, J., Negrete, S.B., Nishimaru, H., Nishijo, H., Takada, M., Go, Y., Inoue, K.i., Shibata, T.: Macaquepose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture. Frontiers in behavioral neuroscience 14, 581154 (2021) 8

20. Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., Rahman, M.M., Di Santo, V., Soberanes, D., Feng, G., et al.: Multi-animal pose estimation, identification and tracking with deeplabcut. Nature Methods 19(4), 496–504 (2022) 5

21. Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, C., Zhang, L., Gao, J.: Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767 (2023) 5

22. Li, H., Zhang, H., Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, L.: Taptr: Tracking any point with transformers as detection. arXiv preprint arXiv:2403.13042 (2024) 14

23. Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., Tu, Z.: Pose recognition with cascade transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1944–1953 (2021) 4

24. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022) 5, 7

25. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7070 (2023) 5

26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 8

27. Liu, H., Chen, Q., Tan, Z., Liu, J.J., Wang, J., Su, X., Li, X., Yao, K., Han, J., Ding, E., et al.: Group pose: A simple baseline for end-to-end multi-person pose estimation. arXiv preprint arXiv:2308.07313 (2023) 4

28. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023) 2, 5, 7, 12

29. Lu, C., Koniusz, P.: Few-shot keypoint detection with uncertainty learning for unseen species. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19416–19426 (2022) 2, 5

30. Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., Wang, Z., den Hengel, A.v.: Poseur: Direct human pose regression with transformers. In: European Conference on Computer Vision. pp. 72–88. Springer (2022) 4

31. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M.: Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience (2018), https://www.nature.com/articles/s41593-018-0209-y 4

32. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 6

33. Nakamura, A., Harada, T.: Revisiting fine-tuning for few-shot learning. arXiv preprint arXiv:1910.00216 (2019) 10

34. Ng, X.L., Ong, K.E., Zheng, Q., Ni, Y., Yeo, S.Y., Liu, J.: Animal kingdom: A large and diverse dataset for animal behavior understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19023–19034 (2022) 4, 8

35. Pereira, T.D., Aldarondo, D.E., Willmore, L., Kislin, M., Wang, S.S.H., Murthy, M., Shaevitz, J.W.: Fast animal pose estimation using deep neural networks. Nature methods **16**(1), 117–125 (2019) 8

36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 5

37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 11

38. Reddy, N.D., Vo, M., Narasimhan, S.G.: Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1906–1915 (2018) 8

39. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024) 14

40. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: 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. pp. 658–666 (2019) 7

41. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: Database and results. Image and vision computing **47**, 3–18 (2016) 8, 9

42. Shi, D., Wei, X., Li, L., Ren, Y., Tan, W.: End-to-end multi-person pose estimation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11069–11078 (2022) 4, 7, 9

43. Shi, M., Huang, Z., Ma, X., Hu, X., Cao, Z.: Matching is not enough: A two-stage framework for category-agnostic pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7308–7317 (2023) 2, 3, 5, 10

44. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017) 10

45. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019) 4, 7

46. Sun, M., Zhao, Z., Chai, W., Luo, H., Cao, S., Zhang, Y., Hwang, J.N., Wang, G.: Uniap: Towards universal animal perception in vision via few-shot learning. arXiv preprint arXiv:2308.09953 (2023) 4, 5

47. Sun, P., Chen, S., Zhu, C., Xiao, F., Luo, P., Xie, S., Yan, Z.: Going denser with open-vocabulary part segmentation. arXiv preprint arXiv:2305.11173 (2023) 5

48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 7

49. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. Advances in Neural Information Processing Systems **36** (2024) 14

50. Wang, Y., Peng, C., Liu, Y.: Mask-pose cascaded cnn for 2d hand pose estimation from single color image. IEEE Transactions on Circuits and Systems for Video Technology **29**(11), 3258–3268 (2018) 8

51. Wu, J., Xue, T., Lim, J.J., Tian, Y., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: Single image 3d interpreter network. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. pp. 365–382. Springer (2016) 8

52. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European conference on computer vision (ECCV). pp. 466–481 (2018) 4

53. Xu, L., Jin, S., Zeng, W., Liu, W., Qian, C., Ouyang, W., Luo, P., Wang, X.: Pose for everything: Towards category-agnostic pose estimation. In: European Conference on Computer Vision. pp. 398–416. Springer (2022) 2, 3, 5, 8, 9, 10

54. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in Neural Information Processing Systems **35**, 38571–38584 (2022) 2, 3, 4, 7

55. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose+: Vision transformer foundation model for generic body pose estimation. arXiv preprint arXiv:2212.04246 (2022) 4

56. Yang, J., Li, B., Yang, F., Zeng, A., Zhang, L., Zhang, R.: Boosting human-object interaction detection with text-to-image diffusion model. arXiv preprint arXiv:2305.12252 (2023) 14

57. Yang, J., Li, B., Zeng, A., Zhang, L., Zhang, R.: Open-world human-object interaction detection via multi-modal prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16954–16964 (2024) 14

58. Yang, J., Wang, C., Li, Z., Wang, J., Zhang, R.: Semantic human parsing via scalable semantic transfer over multiple label domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19424–19433 (2023) 14

59. Yang, J., Zeng, A., Li, F., Liu, S., Zhang, R., Zhang, L.: Neural interactive keypoint detection. arXiv preprint arXiv:2308.10174 (2023) 4, 7

60. Yang, J., Zeng, A., Liu, S., Li, F., Zhang, R., Zhang, L.: Explicit box detection unifies end-to-end multi-person pose estimation. In: The Eleventh International Conference on Learning Representations (2022) 2, 3, 4, 7, 9

61. Yang, J., Zhu, Y., Wang, C., Li, Z., Zhang, R.: Toward unpaired multi-modal medical image segmentation via learning structured semantic consistency. arXiv preprint arXiv:2206.10571 (2022) 14

62. Yang, Y., Yang, J., Xu, Y., Zhang, J., Lan, L., Tao, D.: Apt-36k: A large-scale benchmark for animal pose estimation and tracking. Advances in Neural Information Processing Systems **35**, 17301–17313 (2022) 8

63. Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., Xu, H.: Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. Advances in Neural Information Processing Systems **35**, 9125–9138 (2022) 5

64. Ye, S., Filippova, A., Lauer, J., Vidal, M., Schneider, S., Qiu, T., Mathis, A., Mathis, M.W.: Superanimal models pretrained for plug-and-play analysis of animal behavior. arXiv preprint arXiv:2203.07436 (2022) 4, 5

65. Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D.: Ap-10k: A benchmark for animal pose estimation in the wild. arXiv preprint arXiv:2108.12617 (2021) 3, 4, 8, 9, 14

66. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary detr with conditional matching. In: European Conference on Computer Vision. pp. 106–122. Springer (2022) 5

67. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022) 2, 7

68. Zhang, X., Wang, W., Chen, Z., Xu, Y., Zhang, J., Tao, D.: Clamp: Prompt-based contrastive learning for connecting language and animal pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23272–23281 (2023) 2, 3, 5, 10, 11

69. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022) 5

70. Zhou, M., Stoffl, L., Mathis, M.W., Mathis, A.: Rethinking pose estimation in crowds: Overcoming the detection information bottleneck and ambiguity. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14689–14699 (October 2023) 4

71. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020) 7

72. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36** (2024) 14