# Referring Human Pose and Mask Estimation in the Wild

**Bo Miao**[1]  **Mingtao Feng**[2]  **Zijie Wu**[3]  **Mohammed Bennamoun**[1]
**Yongsheng Gao**[4]  **Ajmal Mian**[1]
[1]University of Western Australia  [2]Xidian University  [3]Hunan University  [4]Griffith University
https://github.com/bo-miao/RefHuman

## Abstract

We introduce Referring Human Pose and Mask Estimation (R-HPM) in the wild, where either a text or positional prompt specifies the person of interest in an image. This new task holds significant potential for human-centric applications such as assistive robotics and sports analysis. In contrast to previous works, R-HPM (i) ensures high-quality, identity-aware results corresponding to the referred person, and (ii) simultaneously predicts human pose and mask for a comprehensive representation. To achieve this, we introduce a large-scale dataset named RefHuman, which substantially extends the MS COCO dataset with additional text and positional prompt annotations. RefHuman includes over 50,000 annotated instances in the wild, each equipped with keypoint, mask, and prompt annotations. To enable prompt-conditioned estimation, we propose the first end-to-end promptable approach named UniPHD for R-HPM. UniPHD extracts multimodal representations and employs a proposed pose-centric hierarchical decoder to process (text or positional) instance queries and keypoint queries, producing results specific to the referred person. Extensive experiments demonstrate that UniPHD produces quality results based on user-friendly prompts and achieves top-tier performance on RefHuman `val` and MS COCO `val2017`.

(a) Multi-person Pose Estimation  (b) Referring Human Pose and Mask Estimation

Figure 1: Task illustration of (a) multi-person pose estimation predicts numerous outcomes and requires selection strategies during deployment, potentially leading to false negatives or suboptimal target results, and (b) our referring human pose and mask estimation requires a *unified promptable* model to simultaneously predict accurate pose and mask for the person of interest, providing *comprehensive and identity-aware* human representations to benefit human-AI interaction.

## 1 Introduction

Human pose estimation in the wild is a fundamental yet challenging problem in the vision community, fueling advancements in various applications like human-AI interaction, activity analysis, video surveillance, assistive robotics, and sports analysis. This task aims to locate keypoints (joint locations) of humans within images in unconstrained environments.

Previous multi-person pose estimation techniques typically follow a two-stage paradigm, separating the problem into person detection and local keypoint regression. These techniques can be summarized as top-down and bottom-up approaches. Top-down approaches [9, 15, 63, 79, 84] use a detection model to identify human bounding boxes and a separate pose estimation model to predict keypoints on the cropped single-human image. The independent models in these methods lead to a non-end-to-end pipeline with substantial computational costs. Bottom-up approaches [3, 10, 31, 62] usually predict keypoint heatmaps to obtain instance-agnostic keypoints and assign them to individuals using heuristic grouping algorithms. The intricate grouping algorithms in these approaches introduce hand-designed parameters and face challenges in handling complex scenarios such as occlusions.

With advancements in attention mechanisms [2, 56, 57, 59, 77] and transformer architectures [4, 95], many recent approaches regard human pose estimation as a direct set prediction problem and design end-to-end differentiable transformers, leading to notable improvements. They employ bipartite matching to establish one-to-one instance correspondence for the set of predictions, avoiding the need for post-processing in the training stage. Among them, PETR [70] proposes a transformer that progressively predicts keypoint positions through pose and keypoint decoders. QueryPose [85], ED-Pose [87], and GroupPose [44] further incorporate a human detection stage for instance feature extraction or query initialization to improve performance and expedite model convergence. They all use keypoint-level queries to capture local details for accurate pose estimation.

Despite demonstrating favorable performance, these methods still require designed strategies to select the best match for a target individual during deployment, which can result in suboptimal outcomes or false negatives, and lack exploration of human-AI interaction to directly predict expected results based on natural prompts. Additionally, they overlook joint human pose and mask estimation, which provides comprehensive human representations to facilitate applications like assistive robotics and sports analysis. For example, accurate joint human pose and mask estimation in unconstrained environments enables robots to locate, analyze, and interact with target individuals, enhancing user experience and assistive tasks.

In this paper, we propose the new task of Referring Human Pose and Mask Estimation (R-HPM) in the wild. As illustrated in Figure 1, unlike multi-person pose estimation [3, 15, 24, 66], R-HPM is a multimodal reasoning task that requires a *unified* model to predict both pose and mask of a referred individual using user-friendly text or positional prompts, enabling comprehensive and identity-aware human representations without post-processing. To achieve this, we introduce RefHuman, a large-scale dataset that substantially extends MS COCO [40] with additional text, scribble, and point prompt annotations. Our dataset accommodates diverse task settings to enhance human-AI interaction. Manually annotating text descriptions and scribbles is expensive. To reduce annotation costs, we design a human-in-the-loop text generation strategy using powerful large language models and employ bezier curves to automate scribble generation.

To benchmark R-HPM, we propose an end-to-end promptable approach called UniPHD. Our approach directly adopts prompts as instance queries and coordinates them with keypoint queries to jointly predict identity-aware keypoint positions and masks for the target. Unlike previous works [44, 70, 85, 87], UniPHD introduces a pose-centric hierarchical decoder (PHD) that employs deformable and graph attention to effectively model local details and global dependencies, ensuring target-awareness and instance coherence. Furthermore, our approach follows a general paradigm that enables seamless integration with the decoders from recent transformer-based methods such as GroupPose [44] and ED-Pose [87] for R-HPM. We conduct extensive experiments on the RefHuman dataset. Integrating our approach with GroupPose and ED-Pose yields promising results in an end-to-end fashion. Moreover, our UniPHD approach achieves top-tier performance on both RefHuman `val` and MS COCO `val2017`, demonstrating the effectiveness of our approach and the significance of our task.

Our main contributions are summarized below:

- We propose Referring Human Pose and Mask Estimation (R-HPM) in the wild, a new task that simultaneously predicts the pose and mask of a specified individual using natural, user-friendly text or positional prompts. This task enhances models with identity-awareness and produces comprehensive human representations to benefit human-AI interaction.

- We introduce RefHuman, a large-scale dataset that substantially extends COCO for R-HPM. RefHuman contains pose and mask annotations for individuals in diverse, unconstrained environments and is enriched with corresponding text and positional prompts.

- We propose an end-to-end promptable UniPHD approach for R-HPM. Our approach performs pose-centric hierarchical decoding, achieving top-ranked performance and establishing a solid benchmark for future advancements in this field.

## 2 Related Work

### 2.1 2D Human Pose Estimation in the Wild

Human pose estimation aims to localize keypoints of individuals in unconstrained environments. For a long period, two-stage approaches have dominated this field, generally divided into top-down and bottom-up methods. Top-down methods [9, 15, 34, 63, 79, 84] first detect and crop each person using an object detector, then perform pose estimation on these cropped instance-level images using a separate model. Although effective, the redundancy from the extra detection step, region of interest operations, and separate training makes them suboptimal. Bottom-Up methods [3, 10, 31, 62] first detect abundant instance-agnostic keypoints and then group them into individual poses. While generally efficient, their intricate grouping algorithms pose challenges in handling complex scenarios, resulting in inferior performance. Furthermore, these two-stage approaches suffer from non-differentiable, hand-crafted post-processing steps that challenge optimization. Inspired by the one-stage object detectors [21, 76], pixel-wise regression methods [49, 53, 55, 64, 73, 75, 78, 81, 93] densely predict pose candidates in an end-to-end fashion and apply Non-maximum Suppression (NMS) to obtain poses for different individuals. However, these methods produce redundant results, challenging the removal of duplicates.

Recent human pose estimation methods [51, 52, 72] explore transformer-based architectures [2, 4, 77, 95] due to their sparse, end-to-end design and promising performance. These methods treat human pose estimation as a direct set prediction problem and use bipartite matching to establish one-to-one instance correspondence during training. Among these, [70] proposes a transformer with a pose decoder to predict keypoints and a joint decoder to refine them. [85] performs estimation on extracted object features to reduce noisy context. [85, 87] incorporate a human detection stage for query initialization to enhance performance. These methods achieve favorable performance but require complex strategies to identify the best match for a specified person during deployment due to the lack of exploration of prompt reasoning. In this work, we propose a multimodal reasoning task to directly and jointly predict the identity-aware pose and mask for a referred person, resulting in comprehensive human body representations and facilitating applications in human-AI interaction.

### 2.2 Referring Image Segmentation

Referring image segmentation (RIS) aims to segment target objects in images based on natural linguistic descriptions [18, 30, 91]. It related tasks include interactive image segmentation [7, 8, 25, 37, 41, 86], which segments targets based on user clicks. Early RIS methods [6, 17, 18, 35, 43, 54, 71, 90] employ convolution and recurrent neural networks for multimodal encoding and segmentation. Their intrinsic constraints in capturing long-range dependencies and handling free-form features often lead to suboptimal performance. To improve multimodal representation and alignment, attention-based bidirectional [19, 20] and progressive [16, 22, 23, 88, 89] cross-modal interaction modules are proposed. Additionally, [80, 92] leverage the strong cross-modal alignment capabilities of pre-trained large language models [67, 69].

With advancements in transformer architectures [4, 61, 77, 95], [11, 13, 29, 38, 42, 45, 58, 60, 83, 94] design end-to-end transformers for language-conditioned segmentation and achieve favorable performance. In this work, we investigate a unified promptable transformer that effectively processes both text and positional prompts, thereby broadening its generalizability and application scope. Moreover, our model coordinates prompts with keypoint queries, enabling the joint prediction of keypoint positions and segmentation masks for specified individuals.

## 3 RefHuman Dataset

We substantially extend COCO [40] to construct the RefHuman dataset. It contains pose and mask annotations for humans along with text and positional prompts to facilitate the new task of R-HPM.
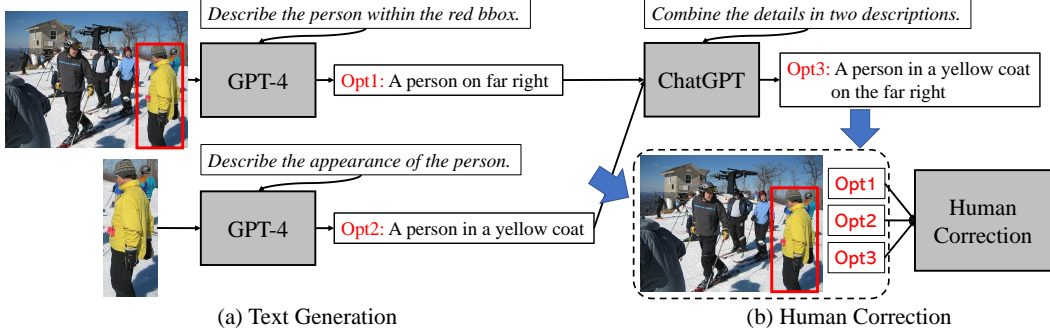
Figure 2: Human-in-the-loop text prompt generation. We use GPT to generate descriptions with complementary local details and global context, then manually review/correct the descriptions.

## 3.1 Data Annotation

Pioneer works on 2D human pose estimation datasets in RGB images include [5, 12, 14, 26, 28]. Recent efforts focus on human pose estimation 'in the wild', establishing datasets such as MPII [1], MS COCO [40], CrowdPose [33], COCO-WholeBody [27], and AI Challenger [82]. Despite their prevalence, these datasets lack crucial text and scribble prompt annotations needed for human-AI interaction. The large-scale RefHuman dataset extends the widely-used MS COCO with additional text and positional (scribbles and points) prompt annotations, facilitating promptable models to enhance human-AI interaction.

**Text Prompt Annotation.** Manually annotating text descriptions is costly. To mitigate this, we design a human-in-the-loop text generation strategy using the large language model, ChatGPT [65]. As shown in Figure 2 (a), we first provide the entire image with a bounding box indication to GPT-4 to generate a text description [Opt1] with global context. However, large language models often misidentify targets in complex scenes. To handle this, we crop the image to focus on the target person and let GPT-4 generate a description [Opt2] with correct local details. Given that individuals may have similar appearances, descriptions generated based on instance-level images are often not distinguishable. Therefore, we combine the global and local descriptions, integrating them into a comprehensive description [Opt3] of the target person. Finally, we manually select and revise these automatically generated descriptions to create accurate text prompts.

Our strategy successfully generates acceptable descriptions for images with simple scenarios. These correspond to over 35% of the cases. In the remaining cases, unsatisfactory descriptions are generated. Common issues include misidentification of individuals, incorrect orientation (*e.g.*, facing left or right), and indistinguishable context. Nevertheless, these descriptions often provide valuable context, expediting the annotation process. To further scale up the RefHuman training set, we integrate text annotations from RefCOCO/+/g [30, 50], doubling the number of referred instances.

**Positional Prompt Annotation.** Bézier curves are parametric curves based on Bernstein Polynomials and widely used in computer graphics. In this work, we employ cubic Bézier curves to simulate scribbles and generate clicks. Given four control points $\mathbf{p_0}$, $\mathbf{p_1}$, $\mathbf{p_2}$, and $\mathbf{p_3}$, the cubic Bézier curve starts at $\mathbf{p_0}$ moving toward $\mathbf{p_1}$ and reaches $\mathbf{p_3}$ from the direction of $\mathbf{p_2}$. The general equation for a Bézier curve $\mathbf{c}(t, n)$ of degree $n$ is:

$$\mathbf{c}(t, n) = \sum_{i=0}^{n} b_{i,n}(t)\mathbf{p}_i, \quad 0 \le t \le 1 \tag{1}$$

$$b_{i,n}(t) = \binom{n}{i}(1-t)^{n-i}t^i, \quad i = 0, ..., n \tag{2}$$

where $b_{i,n}(t)$ represents the Bernstein basis polynomials and $\binom{n}{i}$ is the binomial coefficient. To generate scribble prompts, we randomly sample four control points within the foreground mask to fit a cubic Bézier curve, represented as an ordered set of points $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. The curve is then discretized by uniformly sampling twelve points to form a scribble prompt $\mathbf{s} = \{(x_{\lfloor \frac{kn}{12} \rfloor}, y_{\lfloor \frac{kn}{12} \rfloor}) \mid k = 1, 2, ..., 12\}$. For point prompts, we randomly select a single point $\mathbf{p} = (x, y)$ from $\mathbf{s}$, where $x$ and $y$ are the horizontal and vertical coordinates.

4

Table 1: Data statistics of human-related images in RefCOCO [30], RefCOCO+ [30], RefCOCOg [50], their combined dataset RefCOCO/+/g, and our RefHuman.

| Dataset | Prompt | Target | Image | Instance | Expression |
|---|---|---|---|---|---|
| RefCOCO | Text | Mask | 9209 | 22819 | 65550 |
| RefCOCO+ | Text | Mask | 9209 | 22804 | 66463 |
| RefCOCOg | Text | Mask | 9141 | 16435 | 32299 |
| RefCOCO/+/g | Text | Mask | 13468 | 29470 | 149278 |
| RefHuman (Ours) | **Text, Scribble, Point** | **Pose, Mask** | **21634** | **50210** | **170017** |

## 3.2 Data Statistics

As shown in Table 1, RefHuman includes 50,210 human instances across 21,634 images, larger than RefCOCO [30], RefCOCO+ [30], and RefCOCOg [50], and with additional positional prompts. To construct RefHuman train set, we annotate prompts for all humans in MS COCO train2017 set with at least three surrounding people, a minimum of eight visible keypoints, and an area ratio of at least 2%. For the RefHuman val set, we annotate humans in MS COCO val2017 set, excluding those with non-visible keypoints or an area ratio below 1%, as instances below this threshold are often not visually clear and difficult to describe accurately. Note that each instance may have multiple text, scribble, and point annotations, with each image-prompt pair treated as a separate sample.
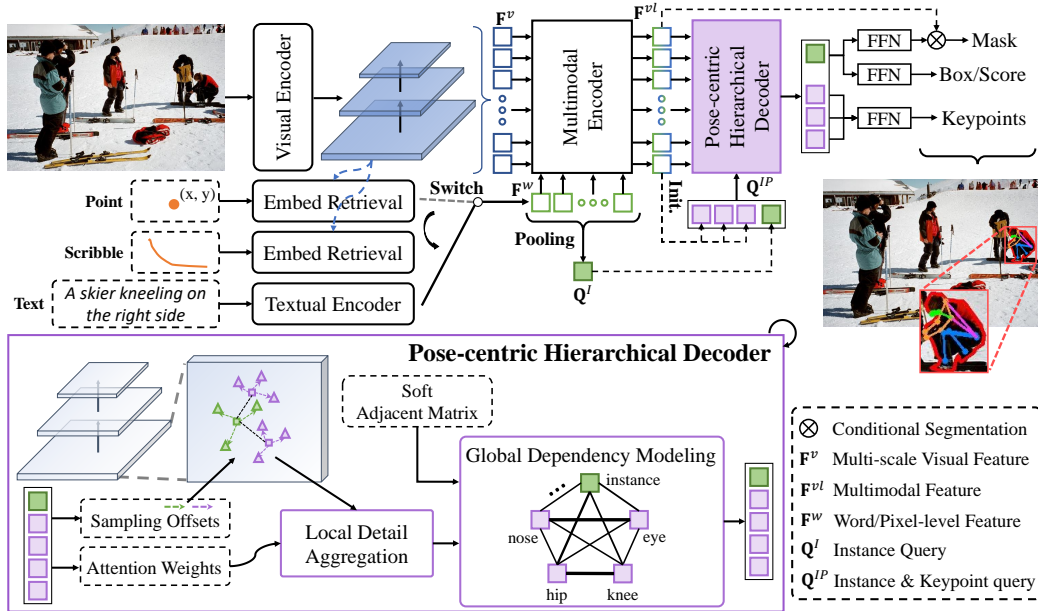
## 4 Method



Figure 3: Detailed architecture of our UniPHD, which contains a multimodal encoder that imbues visual features with prompt awareness and a pose-centric hierarchical decoder that enables prompt-conditioned queries to effectively capture local details and global dependencies within targets. Our unified model is end-to-end and accepts text descriptions, scribbles, or points as prompts to predict the keypoint positions and segmentation mask of the target person.

Given an image $\mathbf{I}$ and a text prompt $\mathbf{T}$ or positional prompt $\mathbf{P}$, our task aims to simultaneously predict the keypoint positions $\mathbf{K} \in \mathbb{R}^{N \times 2}$ and binary segmentation mask $\mathbf{M} \in \mathbb{R}^{H \times W}$ of the referred person, where $N$ is the number of keypoints, and $H$ and $W$ are the spatial dimensions. To this end, we propose a fully end-to-end promptable UniPHD approach, as illustrated in Figure 3. UniPHD consists of four key components: Backbone, Multimodal Encoder, Pose-centric Hierarchical Decoder, and Task-specific Prediction Heads. A small set of prompt-conditioned queries is employed to identify

the referred person and estimate results. During inference, we directly output expected predictions from the highest-scoring query group without any post-processing.

## 4.1 Backbone

**Visual Encoder.** We start by using a visual encoder to extract multi-scale visual features and apply a point-wise convolution to reduce the channel dimension of features to $D = 256$ for efficient multimodal interactions. Feature maps with downsampling rates of $\{8, 16, 32, 64\}$ are then flattened and concatenated into tokenized visual representations $\mathbf{F}^v$. We adopt Swin Transformer [47] as the visual encoder in this work.

**Prompt Encoder.** For a linguistic description with $L$ words as a prompt, we use RoBERTa [46] to extract word-level features $\mathbf{F}^w \in \mathbb{R}^{L \times D}$ and generate sentence-level feature as the instance query $\mathbf{Q}^I \in \mathbb{R}^{1 \times D}$ by pooling $\mathbf{F}^w$. For a discretized scribble or a single point prompt, we retrieve embeddings directly from visual features at stride 16 based on their positions to obtain pixel-level prompt features $\mathbf{F}^w$, which are then pooled to form the instance query $\mathbf{Q}^I$.

## 4.2 Multimodal Encoder

To generate target-related multimodal representations $\mathbf{F}^{vl}$, visual tokens $\mathbf{F}^v$ are first enhanced through a *modality-specific* cross-attention, which integrate information from prompt features:

$$\mathbf{F}^{vl} = \mathbf{F}^v + \text{Attn}_i(Q = \mathbf{F}^v, K = \mathbf{F}^w, V = \mathbf{F}^w) \tag{3}$$

where $i = 0$ for text prompt and $i = 1$ for positional prompts, *i.e.*, we adopt separate parameters for cross-modal fusion of different prompt types. After that, we follow previous works [44, 58, 87] to use a memory-efficient deformable transformer encoder to encode the multimodal features. The output of the transformer encoder $\mathbf{F}^{vl}$ is then forwarded to the decoder to update queries.

## 4.3 Pose-centric Hierarchical Decoder

We employ $n$ groups of queries for the referred human and generate $n$ groups of results. Each group consists of a prompt-conditioned instance query and $k$ learnable keypoint queries. The keypoint queries regard each keypoint as a target and aim to regress their respective positions, while each instance query, conditioned on the prompt, predicts a confidence score, dynamic segmentation filters, and a bounding box to locate the referred human. The scores are supervised by the losses of each query group to indicate result quality. During deployment, we directly generate results for the target using the highest-scoring query group without any post-processing.

**Prompt-conditioned Query Initialization.** We first construct a query group template $\mathbf{Q}^{IP}_{Init} \in \mathbb{R}^{(k+1) \times D}$ by concatenating the text/positional prompt embedding $\mathbf{Q}^I$ with $k$ (*e.g.*, 17) randomly initialized learnable embeddings $\mathbf{Q}^P$. To generate $n$ query groups, we apply linear layers to the prompt-aware multimodal features $\mathbf{F}^{vl}$ to identify the top-$n$ highest-scoring positions $\mathbf{p}$ and estimate their corresponding keypoint positions and centers as reference points, denoted as $\mathbf{c} \in \mathbb{R}^{n \times (k+1) \times 2}$. The template $\mathbf{Q}^{IP}_{Init}$ is then repeated $n$ times and enhanced with reference points $\mathbf{c}$, their associated positional embeddings, and the top-$n$ highest-scoring multimodal embeddings $\mathbf{F}^{vl}_{\mathbf{p}}$, to form $\mathbf{Q}^{IP}$. Finally, the instance query in each copy incorporates the prompt embedding and a positional embedding derived from the keypoints center of each candidate for pose-centric decoding. During decoding process, we employ graph-attention to model global dependencies for each query group, enabling all queries to utilize the prompt as guidance to identify the referred person only.

**Hierarchical Local Context and Global Dependency Modeling.** We introduce a pose-centric hierarchical decoder to effectively model complementary local details and global dependencies for target-aware decoding. As illustrated in Figure 3, local details are first aggregated using the efficient deformable attention [95], similar to [44, 58, 87]. Each instance and keypoint query in $\mathbf{Q}^{IP}$ independently predicts its sampling offsets and corresponding attention weights on the multimodal features $\mathbf{F}^{vl}$. We then aggregate the sampled features accordingly for each query to capture local details. However, after local detail aggregation, the keypoint queries struggle to perceive the prompt information and lacks interactions with each other, challenging the target-awareness and instance coherence. Additionally, the instance query lacks sufficient relevant global context to accurately locate the referred person at pixel-level.

To address these issues, we model each query group as a bipartite graph $\mathbf{G} = \{\mathbf{V}, \mathbf{E}, \mathbf{A}\}$, with nodes $\mathbf{V}$ representing different queries and edges $\mathbf{E}$ denoting relations between nodes, constrained by a learnable soft adjacent matrix $\mathbf{A} \in \mathbb{R}^{(k+1) \times (k+1)}$, which simulates *inherent* keypoint-to-keypoint, keypoint-to-instance, and instance-to-keypoint relations. The edge $\mathbf{E}_{ij}$ from the $i$-th node to the $j$-th node can be formulated as:

$$\mathbf{E}_{ij} = (W^q \mathbf{V}_i)(W^k \mathbf{V}_j)^\top + \mathbf{A}_{ij} \tag{4}$$

where $W^q$ and $W^k$ are learnable projection matrices. Through our graph attention, both instance and keypoint queries capture global dependencies of the target and thus ensure instance coherence. Ultimately, the instance query generates dynamic filters for mask prediction, while the keypoint queries regress target keypoint positions.

### 4.4 Task-specific Prediction Heads

As illustrate in Figure 3, four lightweight heads are built on top of the decoded queries to predict bounding boxes, class scores, masks, and keypoint positions for the target. The box head predicts the bounding box location of the target to aid the learning process. The class head outputs confidence scores, supervised by losses from each query group, to indicate the prediction quality of each query group. The mask head produces dynamic filters for conditional segmentation on the multimodal features $\mathbf{F}^{vl}$. The keypoint head predicts keypoint positions and their visibility for the referred person in images. Detailed training loss functions for these outputs are supplied in the Appendix.

**Conditional Segmentation.** After extracting semantically-rich multimodal features $\mathbf{F}^{vl}$, we address pixel-level target localization using dynamic filters generated by instance queries, which capture both local details and global dependencies. Similar to [58], we standardize the resolution of the multi-scale multimodal features $\mathbf{F}^{vl}$ to $H/8 \times W/8$ and combine them into a single feature map $\mathbf{F}^m$ through efficient element-wise addition. We then reshape the prompt-conditioned dynamic filters to form two point-wise convolutions that apply to $\mathbf{F}^m$ to obtain the segmentation mask for the referred person.

## 5 Experiments

**Metrics.** We use standard metrics to evaluate our task. For pose estimation, we adopt OKS-based average precision (AP) and PCKh at a 0.5 threshold (PCKh@0.5) [1]. For segmentation, we report mask AP and overall Intersection-over-Union (IoU). The AP metrics are evaluated across all query groups, consistent with prior works [44, 87], while PCKh@0.5 and oIoU are measured using only the highest-scoring query group to better reflect real-world deployment.

**Implementation Details.** We train our model on the RefHuman `train`, which contains approximately 46K person instances with 17 keypoints per instance, and evaluate on the RefHuman `val`. We also evaluate our model on MS COCO by generating positional prompts for *all* images in the dataset. Due to the page limit, we leave the further implementation details in the Appendix.

### 5.1 Main Results

In Table 2, we evaluate our models and the recent advances [44, 87] on the RefHuman `val` set. For fairness, all models are evaluated with inputs resized to a maximum of 640 pixels on the longer side.

**Effectiveness of Our Promptable Paradigm.** Recent human pose estimation methods like Group-Pose [87] and ED-Pose [44] predict poses for multiple humans but require hand engineered selection strategies to identify the best match for a specified person, which can result in suboptimal outcomes or false negatives. By integrating their decoders into our end-to-end promptable paradigm, we directly generate poses and masks in one go for the referred person, achieving up to **73.0** pose AP, **89.6** PCKh@0.5, and **85.6** oIoU with scribble prompts. This performance rivals that of previous models which use 3× more training data and hand engineered result selection strategies. Intersection-based result selection is chosen in Table 2 because distance- and IoU-based strategies lead to inferior performance, as discussed in the ablation study. Furthermore, our paradigm achieves advanced performance even with a single point prompt, while previous models struggle with such simple guidance, as the random positive points/clicks can be near poses of unintended humans in crowds. These results demonstrate the effectiveness of our end-to-end promptable paradigm and the significance of our proposed R-HPM task, *i.e.*, accurate joint pose and mask estimation based on user-friendly prompts.

Table 2: Results on RefHuman `val` split. Uni-ED-Pose and Uni-GroupPose integrate ED-Pose [87] and GroupPose [44] into our end-to-end paradigm. *Intersection-based result selection*: selects results covering at least 30% of the ground truth box. [†]: trains models using complete images in COCO `train2017`. FPS is measured on RTX 3090 with a batch size of 24. Uni-ED-Pose and Uni-GroupPose are trained with less data but rival the performance of vanilla models, which perform post-processing with ground truth boxes. Our UniPHD approach achieves top-tier performance.

| | Prompt | Backbone | Pose Estimation | | Segmentation | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | | | AP | PCKh@0.5 | oIoU | AP | Params | FPS |
| *with intersection-based result selection using ground truth boxes* | | | | | | | | |
| GroupPose[†] [44] | BBox | Swin-T | 72.0 | - | - | - | - | - |
| ED-Pose[†] [87] | BBox | Swin-L | 72.8 | - | - | - | - | - |
| GroupPose[†] [44] | BBox | Swin-L | 73.4 | - | - | - | - | - |
| *our R-HPM methods supporting **various** prompts, without result selection strategy* | | | | | | | | |
| Uni-ED-Pose | Text | Swin-T | 65.3 | 78.3 | 74.5 | 61.5 | - | - |
| Uni-GroupPose | Text | Swin-T | 65.0 | 78.0 | 74.7 | 61.4 | - | - |
| **UniPHD** | Text | Swin-T | **66.7** | **79.0** | **75.0** | **62.2** | - | - |
| Uni-ED-Pose | Point | Swin-T | 71.9 | **88.9** | 82.8 | 67.7 | - | - |
| Uni-GroupPose | Point | Swin-T | 71.5 | 88.5 | 82.2 | 67.6 | - | - |
| **UniPHD** | Point | Swin-T | **73.5** | 88.7 | **83.1** | **68.9** | - | - |
| Uni-ED-Pose | Scribble | Swin-T | 72.9 | 89.6 | 84.9 | 69.2 | 175.7M | 35.4 |
| Uni-GroupPose | Scribble | Swin-T | 73.0 | 89.3 | 85.6 | 69.0 | 177.7M | 35.6 |
| **UniPHD** | Scribble | Swin-T | **74.7** | **90.4** | **85.7** | **70.0** | 184.0M | 35.3 |

Table 3: Comparison with state-of-the-art methods on MS COCO `val2017`. Our method achieves leading performance in pose estimation while also offering segmentation capabilities.

| | Max Res. | Backbone | Pose AP | $AP_{50}$ | $AP_{75}$ | $AP_M$ | $AP_L$ |
| --- | --- | --- | --- | --- | --- | --- | --- |
| PETR [70] | 640 | Swin-L | 66.4 | 88.1 | 73.0 | 56.5 | 80.0 |
| ED-Pose [87] | 640 | Swin-L | 68.9 | 89.8 | 75.3 | 60.1 | 81.1 |
| GroupPose [44] | 640 | Swin-T | 68.5 | 88.9 | 75.4 | 60.5 | 79.9 |
| **UniPHD** w/ Point | 640 | Swin-T | 73.5 | 93.7 | 81.1 | 68.0 | 81.8 |
| **UniPHD** w/ Scribble | 640 | Swin-T | **73.9** | **93.9** | **81.9** | **68.5** | **82.2** |

**Effectiveness of Our UniPHD Approach.** In Table 2, our UniPHD approach sets a new benchmark in overall performance for R-HPM through pose-centric hierarchical decoding, outperforming the nearest competitor, Uni-GroupPose w/ Scribble, by **1.7** Pose AP and **1.0** Mask AP, with comparable FPS. Intuitively, the interdependent influence of human keypoints makes graph networks suitable for modeling their dependencies. After applying deformable attention to capture local details, we employ graph attention with a learnable soft adjacent matrix to simulate keypoint-to-keypoint, keypoint-to-instance, and instance-to-keypoint relations. This matrix guides edge construction to effectively model global dependencies and ensure instance-awareness and coherence.

**Results for Different Prompts.** Table 2 shows that scribble prompts outperform point and text prompts in R-HPM. Scribble prompts offer more explicit positional guidance, enhancing instance query robustness. Point prompts, while slightly less accurate, offer greater user convenience in human-AI interaction due to their single-click simplicity. Text prompts face challenges in multimodal alignment, especially for crowded scenarios, but provide linguistic flexibility and enable non-physical interactions. Overall, our unified model effectively handles various prompts, showing great potential to facilitate human-AI interaction.

**Evaluation on MS COCO `val2017`.** We further evaluate UniPHD on MS COCO by aggregating results from all instances in each image. As shown in Table 3, all models are evaluated using inputs resized to a maximum of 640 pixels on the longest side for comparison. Trained solely with positional prompts, UniPHD achieves state-of-the-art performance on COCO `val2017`, even when compared

Figure 4: Qualitative results of our UniPHD with text prompts in various challenging scenarios.

Table 4: Comparison with state-of-the-art text-based segmentation methods on RefHuman.

|  | Backbone | oIoU |
|---|---|---|
| LAVT [88] | Swin-T | 74.5 |
| CGFormer [74] | Swin-T | 75.3 |
| ReLA [42] | Swin-T | 75.9 |
| SgMg [58] | Swin-T | 75.9 |
| **Ours** | Swin-T | **76.3** |

Table 5: Ablation of result selection strategies for GroupPose [44] with Swin-T.

| Selection Strategy | AP | $AP_M$ | $AP_L$ |
|---|---|---|---|
| None | 33.5 | 30.3 | 36.8 |
| w/ L1 | 47.4 | 45.9 | 48.8 |
| w/ IoU ($\tau$=0.3) | 69.1 | 65.8 | 72.3 |
| w/ Intersection ($\tau$=0.5) | 70.9 | 68.7 | **73.1** |
| w/ Intersection ($\tau$=0.3) | **72.0** | **73.6** | 72.7 |

Table 6: Ablation of multi-task learning.

|  | Text Prompt | | Scribble Prompt | |
|---|---|---|---|---|
|  | Pose | Mask | Pose | Mask |
| w/o Pose Head | - | 61.8 | - | 68.7 |
| w/o Mask Head | 63.3 | - | 72.7 | - |
| **Ours** | **66.7** | **62.2** | **74.7** | **70.0** |

Table 7: Ablation of global dependency modeling.

|  | Text Prompt | | Scribble Prompt | |
|---|---|---|---|---|
|  | Pose | Mask | Pose | Mask |
| w/o Global Dep. | 53.6 | 54.4 | 63.5 | 65.8 |
| Self-Attention | 64.9 | 61.1 | 73.8 | 69.4 |
| **Ours** | **66.7** | **62.2** | **74.7** | **70.0** |

to competitors using higher-resolution inputs, further demonstrating the efficacy of our approach and the significance of the proposed task.

**Comparison with Text-based Segmentation Methods.** To further validate the effectiveness of UniPHD in segmentation, we compare it with popular open-source text-based segmentation methods trained on RefHuman, as shown in Table 4. Using only text prompts for training, our model achieves superior performance compared to competitive methods like SgMg [58] and ReLA [42].

**Qualitative Results.** In Figure 4, we present qualitative results of UniPHD in various challenging scenarios such as crowded scenes, occlusions, similar appearances, dim lighting, and viewpoint changes. UniPHD effectively captures appearance, location, action, and context information to generate high-quality outputs. More visualizations are supplied in the Appendix.

## 5.2 Ablation Study

**Result Selection Strategies.** Table 5 shows the performance of a recent human pose estimation model [44] using various result selection strategies on RefHuman `val`. Without carefully designed strategies, the model struggles to accurately predict target individuals. The distance-based strategy fails to effectively filter out high-scoring irrelevant results, while the intersection-based strategy delivers the best performance.

**Multi-task Learning.** In Table 6, we evaluate the impact of multi-task learning in our approach on RefHuman `val` using the AP metric. Removing either head adversely affects the performance of the other. This validates the effectiveness of our decoder, which enables bidirectional information flow between keypoint and instance queries to enhance both predictions.

Table 8: Ablation of query initialization.

|  | Text Prompt | | Scribble Prompt | |
|---|---|---|---|---|
|  | Pose | Mask | Pose | Mask |
| w/o Initialization | 65.5 | 61.0 | 74.0 | 69.6 |
| **Ours** | **66.7** | **62.2** | **74.7** | **70.0** |

Table 9: Training UniPHD with extra data.

| Prompt | Pose Estimation | | Segmentation | |
|---|---|---|---|---|
|  | AP | PCKh@0.5 | oIoU | AP |
| Point | 81.0 | 94.3 | 88.3 | 72.3 |
| Scribble | 81.2 | 94.9 | 89.4 | 73.2 |

**Pose-centric Hierarchical Decoder.** In Table 7, we analyze the impact of global dependency modeling in our decoder. Without global dependencies, keypoint queries struggle to perceive the prompts for predicting target-aware results. Our graph attention clearly outperforms self-attention because it not only models dynamic node relations but also simulates inherent keypoint-to-keypoint, keypoint-to-instance, and instance-to-keypoint relations via soft adjacent matrix.

**Query Initialization.** Similar to recent transformer-based pose estimation methods [44, 87], we use prompt-conditioned query initialization to enrich queries with dynamic spatial priors. Table 8 reveals that omitting these dynamic priors results in a performance decrease of 0.4-1.2% AP. Nonetheless, our model maintains robust performance without query initialization, demonstrating its efficacy.

**Training with Extra Data.** To unleash the capability of UniPHD for positional prompt-based prediction, we expand our training data beyond RefHuman to encompass the entire MS COCO `train2017` images by generating additional point and scribble prompt annotations based on Bézier curves. As shown in Table 9, UniPHD, trained exclusively with positional prompts on the expanded dataset, achieves impressive performance on RefHuman `val`, with up to **94.9** PCKh@0.5 and **89.4** oIoU, demonstrating its scalability.

# 6 Conclusion

We introduced Referring Human Pose and Mask Estimation in the wild, a new task aimed at simultaneously predicting the poses and masks of specified individuals using natural, user-friendly text or positional prompts. This task holds significant potential for enhancing human-AI interactions in fields such as assistive robotics and sports analysis. To achieve this, we introduced the RefHuman dataset that substantially extends MS COCO with additional text and positional prompt annotations. We also proposed the first end-to-end promptable approach named UniPHD, which employs pose-centric hierarchical decoder to model local details and global dependencies for R-HPM. UniPHD achieves top-tier performance, establishing a solid benchmark for future advancements in this field. We hope our new task, dataset, and approach could foster advancements in human-AI interaction and related areas.

**Broader Impacts.** Malicious use of R-HPM models can bring potential negative societal impacts, such as unauthorized mass surveillance. We believe that the model itself is neutral with positive human-centric applications, including assistive robotics and sports analysis.

**Limitations.** The proposed UniPHD approach produces high-quality results with various prompts. However, text prompts perform worse than positional prompts due to misidentification. Future research could enhance vision-language alignment to reduce this performance disparity.

# References

[1] Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR. pp. 3686–3693 (2014)

[2] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR (2015)

[3] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. pp. 7291–7299 (2017)

[4] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020)

[5] Charles, J., Pfister, T., Magee, D., Hogg, D., Zisserman, A.: Personalizing human video pose estimation. In: CVPR. pp. 3063–3072 (2016)

[6] Chen, D.J., Jia, S., Lo, Y.C., Chen, H.T., Liu, T.L.: See-through-text grouping for referring image segmentation. In: ICCV. pp. 7454–7463 (2019)

[7] Chen, X., Cheung, Y.S.J., Lim, S.N., Zhao, H.: Scribbleseg: Scribble-based interactive image segmentation. arXiv preprint arXiv:2303.11320 (2023)

[8] Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., Zhao, H.: Focalclick: Towards practical interactive image segmentation. In: CVPR. pp. 1300–1309 (2022)

[9] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR. pp. 7103–7112 (2018)

[10] 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: CVPR. pp. 5386–5395 (2020)

[11] Chng, Y.X., Zheng, H., Han, Y., Qiu, X., Huang, G.: Mask grounding for referring image segmentation. arXiv preprint arXiv:2312.12198 (2023)

[12] Dantone, M., Gall, J., Leistner, C., Van Gool, L.: Human pose estimation using body parts dependent joint regressors. In: CVPR. pp. 3041–3048 (2013)

[13] Ding, H., Liu, C., Wang, S., Jiang, X.: Vlt: Vision-language transformer and query generation for referring segmentation. TPAMI (2022)

[14] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)

[15] Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: ICCV. pp. 2334–2343 (2017)

[16] Feng, G., Hu, Z., Zhang, L., Lu, H.: Encoder fusion network with co-attention embedding for referring image segmentation. In: CVPR. pp. 15506–15515 (2021)

[17] Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: CVPR. pp. 1115–1124 (2017)

[18] Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: ECCV. pp. 108–124. Springer (2016)

[19] Hu, Y., Wang, Q., Shao, W., Xie, E., Li, Z., Han, J., Luo, P.: Beyond one-to-one: Rethinking the referring image segmentation. In: ICCV. pp. 4067–4077 (2023)

[20] Hu, Z., Feng, G., Sun, J., Zhang, L., Lu, H.: Bi-directional relationship inferring network for referring image segmentation. In: CVPR. pp. 4424–4433 (2020)

[21] Huang, L., Yang, Y., Deng, Y., Yu, Y.: Densebox: Unifying landmark localization with end to end object detection. arXiv preprint arXiv:1509.04874 (2015)

[22] Huang, S., Hui, T., Liu, S., Li, G., Wei, Y., Han, J., Liu, L., Li, B.: Referring image segmentation via cross-modal progressive comprehension. In: CVPR. pp. 10488–10497 (2020)

[23] Hui, T., Liu, S., Huang, S., Li, G., Yu, S., Zhang, F., Han, J.: Linguistic structure guided context modeling for referring image segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. pp. 59–75. Springer (2020)

[24] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. pp. 34–50. Springer (2016)

[25] Jang, W.D., Kim, C.S.: Interactive image segmentation via backpropagating refinement scheme. In: CVPR. pp. 5297–5306 (2019)

[26] Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV. pp. 3192–3199 (2013)

[27] Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 196–214. Springer (2020)

[28] Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: bmvc. vol. 2, p. 5. Aberystwyth, UK (2010)

[29] Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: ICCV. pp. 1780–1790 (2021)

[30] Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014)

[31] Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: CVPR. pp. 11977–11986 (2019)

[32] Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)

[33] Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: CVPR. pp. 10863–10872 (2019)

[34] Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., Tu, Z.: Pose recognition with cascade transformers. In: CVPR. pp. 1944–1953 (2021)

[35] Li, R., Li, K., Kuo, Y.C., Shu, M., Qi, X., Shen, X., Jia, J.: Referring image segmentation via recurrent refinement networks. In: CVPR. pp. 5745–5753 (2018)

[36] Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced nlp tasks. In: ACL (2020)

[37] Li, Z., Chen, Q., Koltun, V.: Interactive image segmentation with latent diversity. In: CVPR. pp. 577–585 (2018)

[38] Li, Z., Wang, M., Mei, J., Liu, Y.: Mail: A unified mask-image-language trimodal network for referring image segmentation. arXiv preprint arXiv:2111.10747 (2021)

[39] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)

[40] 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)

[41] Lin, Z., Zhang, Z., Chen, L.Z., Cheng, M.M., Lu, S.P.: Interactive image segmentation with first click attention. In: CVPR. pp. 13339–13348 (2020)

[42] Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: CVPR. pp. 23592–23601 (2023)

[43] Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., Yuille, A.: Recurrent multimodal interaction for referring image segmentation. In: ICCV. pp. 1271–1280 (2017)

[44] 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. In: ICCV. pp. 15029–15038 (2023)

[45] Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R.K., Mahadevan, V., Manmatha, R.: Polyformer: Referring image segmentation as sequential polygon generation. In: CVPR. pp. 18653–18663 (2023)

[46] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[47] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)

[48] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[49] Lu, P., Jiang, T., Li, Y., Li, X., Chen, K., Yang, W.: Rtmo: Towards high-performance one-stage real-time multi-person pose estimation. arXiv preprint arXiv:2312.07526 (2023)

[50] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR. pp. 11–20 (2016)

[51] Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., Wang, Z.: Tfpose: Direct human pose estimation with transformers. arXiv preprint arXiv:2103.15320 (2021)

[52] Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., Wang, Z., den Hengel, A.v.: Poseur: Direct human pose regression with transformers. In: ECCV. pp. 72–88. Springer (2022)

[53] Mao, W., Tian, Z., Wang, X., Shen, C.: Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In: CVPR. pp. 9034–9043 (2021)

[54] Margffoy-Tuay, E., Pérez, J.C., Botero, E., Arbeláez, P.: Dynamic multimodal instance segmentation guided by natural language queries. In: ECCV. pp. 630–645 (2018)

[55] McNally, W., Vats, K., Wong, A., McPhee, J.: Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation. In: ECCV. pp. 37–54. Springer (2022)

[56] Miao, B., Bennamoun, M., Gao, Y., Mian, A.: Regional video object segmentation by efficient motion-aware mask propagation. In: DICTA. pp. 1–6 (2022)

[57] Miao, B., Bennamoun, M., Gao, Y., Mian, A.: Self-supervised video object segmentation by motion-aware mask propagation. In: ICME (2022)

[58] Miao, B., Bennamoun, M., Gao, Y., Mian, A.: Spectrum-guided multi-granularity referring video object segmentation. In: ICCV. pp. 920–930 (2023)

[59] Miao, B., Bennamoun, M., Gao, Y., Mian, A.: Region aware video object segmentation with deep motion modeling. IEEE Transactions on Image Processing (2024)

[60] Miao, B., Bennamoun, M., Gao, Y., Shah, M., Mian, A.: Temporally consistent referring video object segmentation with hybrid memory. IEEE Transactions on Circuits and Systems for Video Technology (2024)

[61] Miao, B., Zhou, L., Mian, A.S., Lam, T.L., Xu, Y.: Object-to-scene: Learning to transfer object knowledge to indoor scene recognition. In: IROS. pp. 2069–2075 (2021)

[62] Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. Advances in neural information processing systems **30** (2017)

[63] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. pp. 483–499. Springer (2016)

[64] Nie, X., Feng, J., Zhang, J., Yan, S.: Single-stage multi-person pose machines. In: ICCV. pp. 6951–6960 (2019)

[65] OpenAI: Gpt-4 technical report (2023)

[66] Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: CVPR. pp. 4903–4911 (2017)

[67] 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: ICML. pp. 8748–8763. PMLR (2021)

[68] 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: CVPR. pp. 658–666 (2019)

[69] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)

[70] Shi, D., Wei, X., Li, L., Ren, Y., Tan, W.: End-to-end multi-person pose estimation with transformers. In: CVPR. pp. 11069–11078 (2022)

[71] Shi, H., Li, H., Meng, F., Wu, Q.: Key-word-aware network for referring expression image segmentation. In: ECCV. pp. 38–54 (2018)

[72] Stoffl, L., Vidal, M., Mathis, A.: End-to-end trainable multi-instance pose estimation with transformers. arXiv preprint arXiv:2103.12115 (2021)

[73] Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR. pp. 5693–5703 (2019)

[74] Tang, J., Zheng, G., Shi, C., Yang, S.: Contrastive grouping with transformer for referring image segmentation. In: CVPR. pp. 23570–23580 (2023)

[75] Tian, Z., Chen, H., Shen, C.: Directpose: Direct end-to-end multi-person pose estimation. arXiv preprint arXiv:1911.07451 (2019)

[76] Tian, Z., Chu, X., Wang, X., Wei, X., Shen, C.: Fully convolutional one-stage 3d object detection on lidar range images. Advances in Neural Information Processing Systems **35**, 34899–34911 (2022)

[77] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)

[78] Wang, D., Zhang, S.: Contextual instance decoupling for robust multi-person pose estimation. In: CVPR. pp. 11060–11068 (2022)

[79] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020)

[80] Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: CVPR. pp. 11686–11695 (2022)

[81] Wei, F., Sun, X., Li, H., Wang, J., Lin, S.: Point-set anchors for object detection, instance segmentation and pose estimation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. pp. 527–544. Springer (2020)

[82] Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., et al.: Ai challenger: A large-scale dataset for going deeper in image understanding. arXiv preprint arXiv:1711.06475 (2017)

[83] Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: CVPR. pp. 4974–4984 (2022)

[84] Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: ECCV. pp. 466–481 (2018)

[85] Xiao, Y., Su, K., Wang, X., Yu, D., Jin, L., He, M., Yuan, Z.: Querypose: Sparse multi-person pose regression via spatial-aware part-level query. Advances in Neural Information Processing Systems **35**, 12464–12477 (2022)

[86] Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.S.: Deep interactive object selection. In: CVPR. pp. 373–381 (2016)

[87] 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)

[88] Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: CVPR. pp. 18155–18165 (2022)

[89] Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: CVPR. pp. 10502–10511 (2019)

[90] Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: CVPR. pp. 1307–1315 (2018)

[91] Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 69–85. Springer (2016)

[92] Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: ICCV. pp. 5729–5739 (2023)

[93] Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)

[94] Zhu, C., Zhou, Y., Shen, Y., Luo, G., Pan, X., Lin, M., Chen, C., Cao, L., Sun, X., Ji, R.: Seqtr: A simple yet universal network for visual grounding. In: ECCV. pp. 598–615. Springer (2022)

[95] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2021)

# A Appendix

## A.1 Training Loss Functions

The overall loss function for UniPHD consists of four parts:

$$\mathcal{L}_{train} = \mathcal{L}_{box} + \mathcal{L}_{class} + \mathcal{L}_{pose} + \mathcal{L}_{mask}, \tag{5}$$

where $\mathcal{L}_{box}$ is for human box regression that contains L1 loss and GIOU [68] loss, $\mathcal{L}_{class}$ is for classification includes focal loss [39] with $\alpha = 0.25$ and $\gamma = 2$, $\mathcal{L}_{pose}$ is for keypoint regression and visibility prediction that includes L1 loss, the constrained L1 loss-OKS loss [70], and cross entropy loss, $\mathcal{L}_{mask}$ is for segmentation that includes Dice [36] loss and focal loss. The loss coefficients $\lambda_{box}^{L1}$, $\lambda_{box}^{GIOU}$, $\lambda_{class}^{focal}$, $\lambda_{pose}^{L1}$, $\lambda_{pose}^{OKS}$, $\lambda_{pose}^{CE}$, $\lambda_{mask}^{Dice}$, $\lambda_{mask}^{focal}$ are 5, 2, 2, 10, 4, 4, 5, 2, following [44, 58, 87]. The Hungarian algorithm [32] is used to identify the optimal assignment (query group) with the highest similarity to the ground truth for training. The class label for the optimal instance query with the minimum loss from the ground truth is set to one, while all others are set to zero.

## A.2 Implementation Details

We follow the common optimization strategies in [44, 58, 83, 87] to train the models. During training, we augment input images through random flip, random crop, and random resize with the shorter sides within the range of 360 to 640 pixels and the longer sides up to 640 pixels. For each iteration, we randomly select either a text or positional prompt with equal probability. We use the AdamW [48] optimizer with a weight decay of $1 \times 10^{-4}$ and train our models on 24GB RTX 3090 GPUs with batch size 16 for 20 epochs. The initial learning rates are set to $1 \times 10^{-5}$ for the visual encoder and $1 \times 10^{-4}$ for other components, with a rate decay at the 18th epoch by a factor of 10. Both the multimodal encoder and pose-centric hierarchical decoder consist of 6 layers, and we use 20 query groups for our models. During testing, we resize the input images with their longer sides up to 640 pixels.

## A.3 Different Query Group Numbers

In Table 10, our approach perform well across different numbers of query groups, measured by the AP metric. This hyperparameter primarily affects pose estimation, with 20 query groups identifying more candidates to achieve the best performance.

Table 10: Ablation of different number of query groups.

| Num. | Text Prompt | | Scribble Prompt | |
|---|---|---|---|---|
| | Pose | Mask | Pose | Mask |
| 5 | 62.6 | 60.1 | 72.1 | 69.1 |
| 10 | 64.4 | 61.4 | 72.4 | 69.5 |
| 20 | 66.7 | 62.2 | 74.7 | 70.0 |

## A.4 Increasing Model Capacity

In Table 11, we enhance model capacity by adding multimodal encoder layers and increasing the feature dimensions from 256 to 384. The results indicate our current settings are already highly effective.

Table 11: Ablation of increasing model capacity.

| | Text Prompt | | Scribble Prompt | |
|---|---|---|---|---|
| | Pose | Mask | Pose | Mask |
| Baseline | **66.7** | 62.2 | 74.7 | 70.0 |
| w/ more layers | 66.3 | 62.0 | **74.9** | **70.4** |
| w/ higher dimensions | 66.6 | **62.8** | 74.5 | 70.2 |

## A.5 Additional Visualizations

In Figure 5, we present additional qualitative results of UniPHD with text and positional prompts. UniPHD effectively processes these prompts to produce high-quality results.
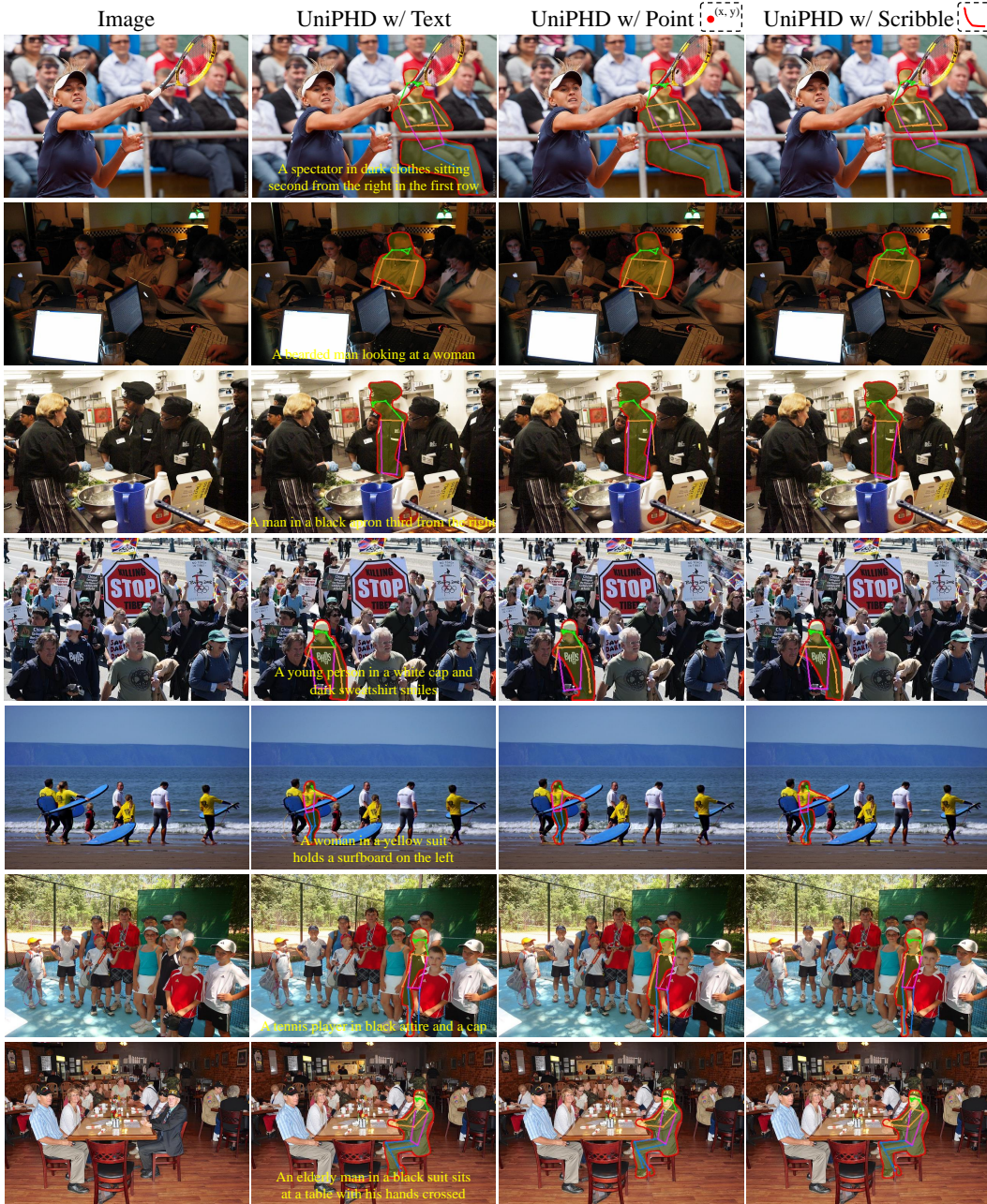


Figure 5: Qualitative results of our UniPHD with different prompts in various challenging scenarios.

**Licenses of MS COCO and RefCOCO/+/g.** The annotations in MS COCO belong to the COCO Consortium and are licensed under CC BY 4.0. The use of the images complies with Flickr Terms of Use. See https://cocodataset.org/#termsofuse for more details. The RefCOCO/+ datasets are licensed under Apache License 2.0 and the RefCOCOg dataset is licensed under CC BY 4.0.

**Terms of Use and License of RefHuman.** RefHuman is licensed under CC BY 4.0.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims, which are supported by our experimental results. See Abstract and Sec. 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Sec. 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not contain any theoretical assumptions or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We report the dataset, model, and training details in Sec. 3, Sec. 4, Sec. 5, and the Appendix. The introduced dataset and code are publicly available.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The introduced dataset and code are publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Sec. 5 and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We follow the usual format used in previous related works to report and compare the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See the Appendix.

   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:
   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Sec. 6.

    Guidelines:
    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the creators or original papers of all the related code, data, and models used in this paper. All the assets are free for research study and widely used in previous related works. We also provide the licenses of the used datasets in the Appendix.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We report the dataset, model, and training details in Sec. 3, Sec. 4, Sec. 5, and the Appendix. The introduced dataset and code are publicly available.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We did not use crowdsourcing or conduct research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We did not use crowdsourcing or conduct research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.