HPE-CogVLM: Advancing Vision Language Models with a Head Pose Grounding Task

Yu Tian, Tianqi Shao, Tsukasa Demizu, Xuyang Wu, Hsin-Tai Wu

***Abstract*—Head pose estimation (HPE) requires a sophisticated understanding of 3D spatial relationships to generate precise yaw, pitch, and roll angles. Previous HPE models, primarily CNN-based, rely on cropped close-up human head images as inputs and often lack robustness in real-world scenario. Vision Language Models (VLMs) can analyze entire images while focusing on specific objects through their attention mechanisms. In this paper, we propose a novel framework to improve the HPE accuracy by leveraging the object detection grounding capability of a VLM, referred to as CogVLM. We empirically find that directly LoRA fine-tuning of this VLM for the HPE task fails to achieve desirable HPE accuracy, while some model merging methods can improve accuracy but frequently produce blended invalid response formats, struggling to handle both object detection and HPE tasks simultaneously. To integrate HPE capability into CogVLM effectively, we develop a novel LoRA layer-based model merging method. This merging approach applies a high cosine similarity threshold and a "winner-takes-all" layer selection strategy, aligning attention to the HPE task while preserving original object detection knowledge. It successfully resolves issues with blended invalid response formats and improves accuracy. Results show that our HPE-CogVLM achieves a 31.5% reduction in Mean Absolute Error over the current state-of-the-art CNN model, 6DRepNet, in cross-dataset evaluation. Furthermore, HPE-CogVLM outperforms both directly LoRA fine-tuned and task arithmetic-based merged VLMs across all HPE metrics.**

***Index Terms*—Vision language model, Model merging, Head pose estimation, Visual grounding task, Catastrophic forgetting problem.**

I. Introduction

NOWADAYS, the head pose estimation (HPE) technique is applicable in various fields such as attention estimation [1]–[3], face recognition [4]–[7], customer behavior analysis [8], [9], driver assistance systems [10]–[13] and human-robot interaction [14]. This task involves predicting the Euler angles (yaw, pitch, and roll) of human heads from images or videos. Recent research efforts on some CNN-based models like 6DRepNet [15], HopeNet [16] and WHENet [17] have made significant advancements in HPE.

Despite the recent surge of interest in HPE, the application of this technique still faces several challenges in real-world scenario. The CNN-based models typically rely on narrowly focused datasets such as 300W-LP [18] for training, and validate on similarly constrained datasets like AFLW2000 [18] and BIWI [19]. These datasets primarily feature close-up images focused on a single head, mostly displaying frontal faces with yaw angles from $-99°$ to $99°$, instead of covering the full range of head poses ranged from $-180°$ to $180°$. Additionally, the frequent use of close-up images in these datasets reduces input variability, leading to uniform backgrounds. This uniformity could result in limited robustness in diverse real-world environments. The DirectMHP [20] model attempts HPE prediction in a one-shot manner trained on the full-range HPE datasets, like Agora [21] and CMU Panoptic [22], but this model lacks stability and struggles to balance the head bounding box (BBox) detection and HPE task performance. Consequently, the model's effectiveness remains uncertain in real-world settings.

The appearance of Large Language Models (LLMs) has made substantial advancements in a wide range of applications, significantly enhancing our daily lives by offering sophisticated assistance across various tasks. Recently, Vision Language Models (VLMs) have attracted significant attention for their capability to process multimodal information [23]–[26]. By combining image and video understanding with language processing, VLMs can effectively perform complex tasks, such as visual question answering [23], [25], [27] and visual grounding [26], [28]. In this paper, we leverage the attention mechanism of VLMs to enhance the robustness and accuracy of the HPE task. The VLMs empower the model to process the entire image while focusing specifically on the head region, eliminating the need to crop out backgrounds. This capability reduces the risk of overfitting to a limited set of visual features and allows the model to leverage context from the full scene. As a result, it enhances the robustness of tasks that traditional CNN-based methods often struggle to address. We validate this approach by integrating HPE functionality into the grounding model of CogVLM [26]. The grounding CogVLM's capabilities include caption grounding, referring expression generation, referring expression comprehension and grounded visual question answering [26]. All of these functionalities involve in the description of object localization in the BBox format of $[[x_0, y_0, x_1, y_1]]$ as shown in Figure 1(a). This BBox grounding capability provides a foundational skill for learning the new HPE task introduced in this paper. By leveraging this capability in designed prompts, our approach can accurately identify the head position within the image, even when multiple people are present in the image.

Despite its advantages, incorporating the HPE task into the grounding CogVLM introduces several challenges. First, VLM tasks such as image description, visual reasoning, and visual perception usually contain answering questions with natural

Yu Tian, Tianqi Shao, Tsukasa Demizu, and Hsin-Tai Wu are with Docomo Innovations, Inc., Sunnyvale, US (e-mail: yu.tian@docomoinnovations.com, jshao@docomoinnovations.com, tsukasa.demizu@docomoinnovations.com, hwu@docomoinnovations.com).

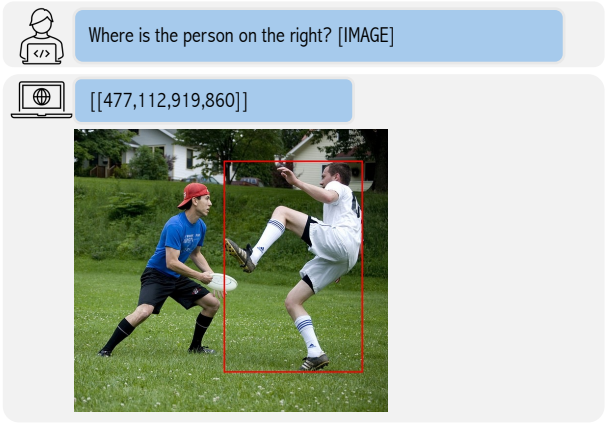
Xuyang Wu is with Department of Computer Science and Engineering, Santa Clara University, Santa Clara, US (e-mail: xwu5@scu.com).

Xuyang Wu and Hsin-Tai Wu are the corresponding author.

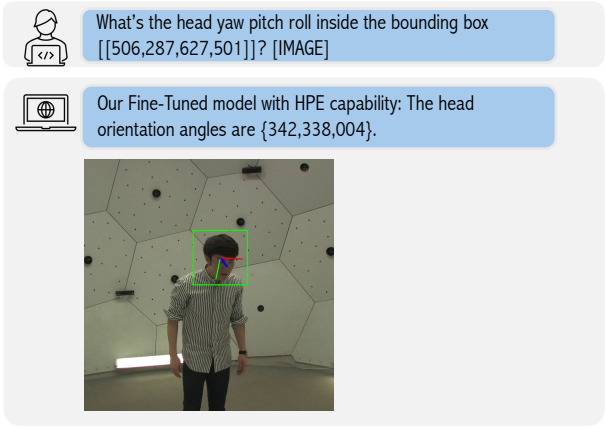
language responses. In contrast, our HPE task requires the VLM to produce precise numerical Euler angles. Although the grounding CogVLM can predict BBoxes, indicating its ability to produce numerical responses, the HPE task is significantly more complicated. HPE requires predicting the human head's orientation in terms of yaw, pitch, and roll angles, which involves interpreting 3D orientation from 2D images. This introduces additional dimensions of depth and angular perspective not required in the basic BBox detection task. Therefore, it raises the challenge of whether the grounding model can provide HPE answers with much higher accuracy. Secondly, catastrophic forgetting [29]–[31] poses a significant challenge in fine-tuning LLMs. The catastrophic forgetting problem is a phenomenon that LLMs tend to forget previously learned information when acquiring new data. Although extensive research has been conducted on mitigating catastrophic forgetting in general tasks, there is currently a lack of research specifically addressing this issue within the context of complex grounding tasks. Lastly, the original grounding CogVLM only involves in outputting responses with natural languages and BBoxes in $[[x_0, y_0, x_1, y_1]]$ format. In this paper, we introduce a new format $\{yaw\_angle, pitch\_angle, roll\_angle\}$ for answering HPE prompts as shown in Figure 1(b). This enriches the knowledge of the original grounding CogVLM, meanwhile increasing the complexity of output formats. Empirically, we have observed that the directly LoRA [32] fine-tuning and model merging methods frequently generate blended invalid outputs like $[[x_0, y_0, yaw\_angle\}$, which is referred as invalid answers in this paper. More invalid answers are detailed in Table II.

In this paper, for addressing the catastrophic forgetting problem in grounding tasks, we evaluate and improve the data rehearsal methods [29], [30] that were originally used in non-grounding VLMs. The results show that the visual grounding task, which demands accurate numerical outputs, requires a significantly larger rehearsal ratio than non-grounding VLMs. Here, the rehearsal ratio represents the percentage of images randomly selected from earlier training phases that are reintegrated during the training of new tasks [29], [30]. To improve HPE accuracy and address blended invalid outputs, we propose and validate a model merging method based on LoRA layers. Utilizing this approach, our model demonstrates exceptional robustness, achieving a 31.5% reduction in Mean Absolute Error (MAE) of Euler angles, compared to the CNN-based state-of-the-art (SOTA) in cross-dataset evaluations. Furthermore, we compare our LoRA layer-based merged model with both the directly LoRA fine-tuned model and the task arithmetic-based merged model within CogVLM. Our approach consistently shows superior performance in both of MAE and invalid answer ratio reduction. Our contributions can be concluded as following:

- Our work pioneers the improvement of HPE tasks through leveraging the visual grounding capability of CogVLM, demonstrating the VLM's ability to manage complex 3D spatial perception while retaining existing object localization knowledge.
- To the best of our knowledge, this is the first work to explore the issues of catastrophic forgetting and blended invalid response when multiple grounding tasks are involved.
- We propose a novel LoRA layer-based model merging method that adopts a "winner takes all" strategy, significantly outperforming the CNN-based SOTA, directly LoRA fine-tuned VLM, and task arithmetic-based merged VLM in terms of MAE and invalid answer ratio reduction. This demonstrates our method is able to achieve outstanding robustness and effectiveness in the HPE task, and holds potential for broader application in various grounding tasks.



(a) Example of CogVLM Grounding.



(b) Example of HPE-CogVLM.

Fig. 1: Examples of CogVLM and HPE-CogVLM. (a) shows an example of CogVLM grounding capability, which demonstrates the original grounding CogVLM's ability to identify objects based on prompts, a foundational skill useful for HPE task. (b) displays a visualization of head orientation predicted by our HPE-CogVLM from the CMU Panoptic dataset, using Euler angles. The head pose labels are depicted with pitch (red axis), roll (green axis), and yaw (blue axis) angles, each indicated in their respective directions.

II. Related Work

A. Head Pose Estimation (HPE)

Traditional approaches for HPE include both landmark-based [33] and landmark-free methods [34], [35]. Landmark-based methods rely on detecting specific facial landmarks,

such as the eyes, nose, and mouth, to estimate head orientation. While these methods perform well under controlled conditions, they struggle with full-range HPE because extreme head poses often obscure facial features, making it challenging to detect landmarks accurately [36]. Given our focus on full yaw range HPE, we prioritize landmark-free approaches, which are better suited for this task. Under this approach, several models divide continuous rotation variables into discrete bins for classification purposes [16], [17], [37]–[39]. Besides those, FSA-Net [40] employs a stage-wise regression and feature aggregation scheme to predict Euler angles. Meanwhile, models such as 6DRepNet [15] and TriNet [41] take a different approach by estimating the rotation matrix instead of directly predicting Euler angles. Despite their potential, CNN-based methods face significant robustness issues in real-life scenarios, often struggling with lighting variations, occlusions, and complex backgrounds. In this paper, we use several traditional CNN-based landmark-free HPE approaches as baselines. This allows us to assess conventional limitations and benchmark the effectiveness of our proposed method.

B. Grounding in Vision Language Models

Some VLMs with grounding capabilities can provide accurate BBox information in the format of $[[x_0,\ y_0,\ x_1,\ y_1]]$ based on specific prompts [26], [42]–[44]. This functionality is crucial for tasks requiring precise spatial awareness, like object detection and image captioning, as accurately identifying object positions enhances visual scene understanding. Unlike traditional BBox prediction task, HPE task requires a understanding of 3D spatial relationships to produce accurate Euler angles in the format of $\{yaw\_angle,\ pitch\_angle,\ roll\_angle\}$. Most existing VLMs are not inherently designed to handle such queries, as they typically output 2D BBoxes rather than 3D rotational data. Currently there is limited research investigating the effectiveness of VLMs in HPE task. While some related efforts exist, such as the work on CLIP-Gaze [45], which focuses on the gaze estimation task by using CLIP [46] model, CLIP-Gaze is limited to applications requiring only a narrow yaw angle range and does not address catastrophic forgetting, leading to a loss of CLIP's multi-task capabilities.

C. Model Merging in LLMs

There has been extensive exploration of model merging techniques designed to enhance the capabilities of LLMs. These methods aim to combine multiple LLMs, each with specialized functionalities, into a unified model capable of addressing a range of tasks across various domains. The typical merging methods [47]–[53] usually apply rules or algorithms to trim or merge the parameters of LLMs. For example, task arithmetic [47] defines arithmetic rules to incorporate new capabilities or delete undesired ones, allowing task-specific parameters to be adjusted in a controlled manner. Other approaches, like evolutionary model merging [53] leverage evolutionary algorithms to iteratively optimize the merging process. These algorithms evaluate multiple merging configurations across generations, selecting the most effective combinations based on defined fitness criteria. However, we have empirically found that existing merging methods usually produce blended invalid outputs, when dealing with multiple grounding tasks that require distinct output formats. To address this challenge, we propose a novel model merging method designed specifically to mitigate blended invalid outputs.

D. Catastrophic Forgetting Problem

Catastrophic forgetting has been a significant issue that limits the effectiveness of LLMs, as they tend to forget previously knowledge when learning new knowledge. This problem is particularly pronounced in continual learning settings, where models are sequentially exposed to new tasks and risk losing their ability to perform earlier ones [54]. The rehearsal method [29]–[31], [55], [56] is the most widely used method to mitigate catastrophic forgetting. It involves reusing a small portion of old task datasets into the new task fine-tuning process. By periodically revisiting earlier knowledge, the model maintains a balanced representation of all tasks, reducing the likelihood of forgetting. Despite these advancements, catastrophic forgetting remains under-explored in grounding tasks, which require precise numerical outputs. In this paper, we evaluate and refine the rehearsal method for grounding tasks, aiming to balance prior knowledge retention with effective adaptation to new tasks.

III. HPE-CogVLM Framework

The proposed framework of HPE-CogVLM as shown in Figure 2 is structured through a multi-stage process. Each stage of this framework is designed to enhance different aspects of the model's capabilities, gradually refining the parameters to balance HPE and BBox tasks. The fine-tuning process at each stage follows the CogVLM's fine-tuning scripts[1], which implement LoRA [32] across transformer blocks, including the query, key, value of attention layers and dense layers. Subsequently, the LoRA matrices of each layer are accumulated into the corresponding layer in original model. To support this multi-stage process, the framework utilizes a variety of datasets, each serving a distinct role in model training and evaluation. Table I provides an overview of how these datasets contribute at different stages, such as enhancing human head BBox detection, improving HPE task accuracy, and preventing catastrophic forgetting. Below is a detailed description of each stage in the framework:

A. Stage 1: Pre-training of the Original Grounding CogVLM on Weak Label Data

As indicated in the CrowdHuman dataset [57] functionality column of Table I, the primary goal at this initial stage is to train the model to develop its capability for human head BBox detection. Additionally, this stage acts as a warm-up for the model's HPE capability by utilizing weak label images. This prepares the model to initiate the HPE task in real-world scenarios.

[1]https://github.com/THUDM/CogVLM

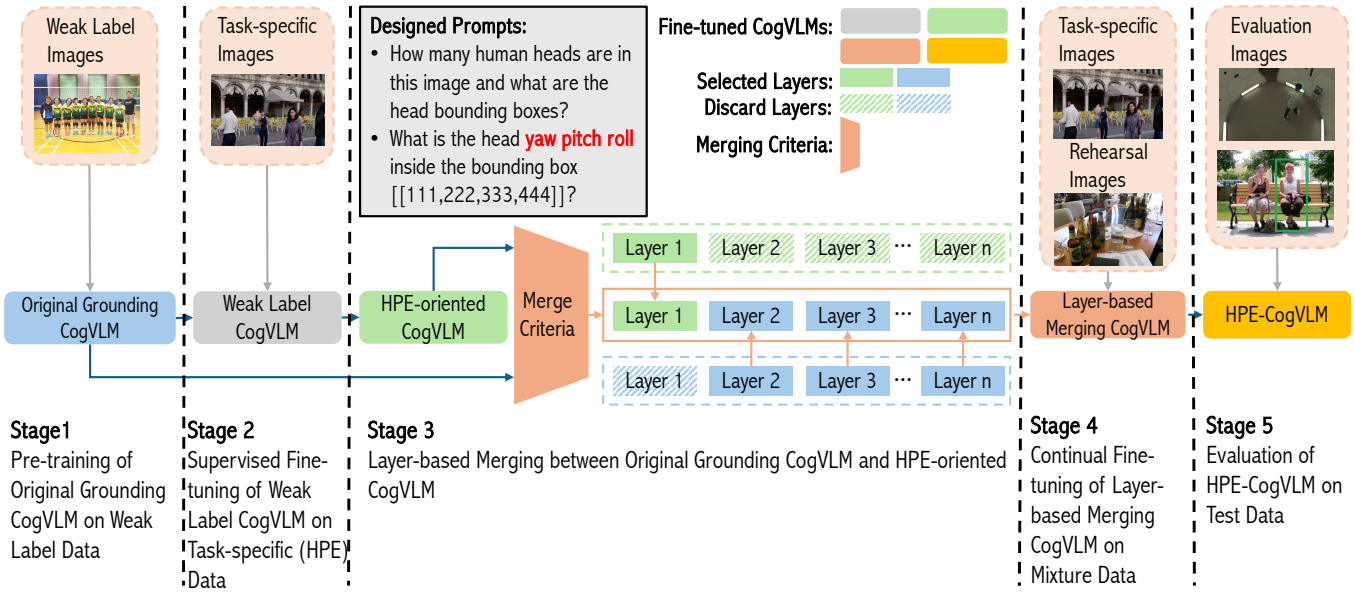


Fig. 2: The framework of integrating HPE task into the original grounding CogVLM. This diagram illustrates our multi-stage integration process of HPE task into the original grounding CogVLM model with the information of dataset usages, designed prompts and model merging strategy.

TABLE I: A detailed overview of various datasets used in our framework.

Task	Dataset	# of Images		# of Heads		Usage	Functionality
		Train	Test	Train	Test		
Weak Label Images	CrowdHuman [57]	11,731	-	94,795	-	Stage 1	Focus on human head BBox detection Warm-up for HPE task using weak label images
Task-specific Images	Agora [21]	9,654	-	64,187	-	Stage 2, 4	Training with precise HPE labels from synthetic images Focus on HPE task to improve accuracy
Rehearsal Images	Refcoco [28] Refcoco+ [28] Refcocog [58]	42,404 42,278 42,226	- - -	- - -	- - -	Stage 4	Preventing catastrophic forgetting with Refcoco/+/g training set Investigating rehearsal ratio optimization
Evaluation Images	CMU Panoptic [22] Refcoco [28] Refcoco+ [28] Refcocog [58]	- - - -	16,216 3785 3773 5023	- - - -	32,738 - - -	Stage 5	Evaluating the accuracy of the HPE task on CMU Panoptic test set Evaluating the BBox detection knowledge on Refcoco/+/g test set

To achieve this, the original grounding CogVLM undergoes pre-training on the CrowdHuman dataset, which consists of images of real people in diverse scenes. Since the original CrowdHuman dataset provides only accurate head BBox annotations and lacks ground truth (GT) annotations for HPE, we infer the weak HPE annotations using 6DRepNet. Therefore, the output model from this stage is termed as weak label CogVLM as shown in Figure 2. As the CrowdHuman dataset offers a rich collection of human head images in various poses and contexts, this approach enables the model to accurately locate the BBoxes of human heads and establishes a foundational understanding for the subsequent HPE task.

B. Stage 2: Supervised Fine-tuning of the Weak Label CogVLM on Task-specific (HPE) Data

Following the pre-training stage, the model progresses to a supervised fine-tuning phase which is exclusively trained on task-specific HPE images. As shown in Table I, we utilize the Agora dataset as the task-specific images. Unlike the broader CrowdHuman dataset used previously, the agora dataset consists of synthetic images that offer more accurate annotations tailored to capture precise head pose information. The detailed annotations in this dataset allow the model to focus on subtle variations in head orientation, enhancing its ability to make fine-grained predictions. This stage focuses on improving the HPE accuracy, addressing the weaknesses in the weak label model's performance due to lower-quality annotations. The output model is referred as the HPE-oriented CogVLM as shown in Figure 2. This refined model combines the real-image exposure and contextual background knowledge gained from the previous stage with the task-specific precision acquired in this stage.

C. Stage 3: Layer-based Merging between Original Grounding CogVLM and HPE-oriented CogVLM

During this key stage, the original grounding CogVLM is merged with HPE-oriented CogVLM (from Stage 2) based on cosine similarity criteria. In our framework, cosine similarity

is calculated as the average cosine similarity between the layer parameter tensors of the original grounding CogVLM and the HPE-oriented CogVLM along the last dimension. Cosine similarity is used to gauge the amount of information shared between layers. The threshold of cosine similarity can help to determine whether layers from the HPE-oriented CogVLM should be integrated into the final model. Since LoRA fine-tuning is applied in previous stages, most of the original model parameters are only minimally altered, necessitating a high threshold for cosine similarity. In our experiments, We empirically set the threshold at 0.95. If the similarity falls below this threshold, we opt to completely retain the original knowledge. Otherwise, if the similarity exceeds the threshold, which indicates a substantial overlap in information due to the stringent criteria, we select the entire layer from the HPE-oriented CogVLM to guarantee the minimal risk of losing important existing knowledge. Consequently, the merging criteria is detailed as below:

- We calculate and rank the cosine similarities across all layers from both models, and always select the layer from the original grounding CogVLM model within the smallest 1% of cosine similarities.
- When the cosine similarity between two layers from each model is less than the threshold, we also select the layer from the original grounding CogVLM.
- Otherwise, we choose the layer from the HPE-oriented CogVLM.

Precedent methods usually involve in directly LoRA fine-tuning or merging models at the parameter-level by setting hyper parameters and developing algorithms to discard and merge specific parameters [47], [51]–[53]. However, we empirically find that directly LoRA fine-tuning does not achieve the desired HPE accuracy, and existing merging methods can achieve better accuracy but struggle when handling multiple grounding tasks with different output formats. They often blend output structures, which leads to invalid answers. For example, when we query with HPE prompts, the merged model may return a NLP response like "a person of head" or provide nonsensical responses like "[[999,231,123,389}". More examples are detailed in Table II. To address these issues, we set a high cosine similarity threshold and adopt a "winner-takes-all" method to select entire layers from either the original grounding CogVLM or the HPE-oriented CogVLM. This enables the merged model's attention to align with the new HPE task at minimal cost. By applying a high similarity threshold, we introduce new knowledge only when there is substantial informational overlap between the two models, ensuring that even when a layer is chosen from the HPE-oriented model, existing knowledge remains well-preserved. This approach minimizes attention loss while also aligning the model effectively with the new task. Although this new capability introduced at this stage may be basic, they lay a strong foundation for further refinement, allowing the model to build on this baseline effectively in stage 4. Additionally, by incorporating entire layers rather than individual parameters, the method preserves the integrity of the model's structure, reducing the risk of response blending when handling multiple grounding tasks with varying requirements.

D. Stage 4: Continual Fine-tuning of Layer-based Merging CogVLM on Mixture Data

After merging, the layer-based merging CogVLM undergoes an additional round of fine-tuning with both the task-specific HPE images and the rehearsal images. The optimal rehearsal ratio for training rehearsal images is pre-defined during Stage 1 by running several parallel experiments. In these experiments, we tune the original grounding CogVLM with weak label images, each combined with varying proportions of rehearsal images (0%,1%,10%,25%). The rehearsal ratio that yields the best performance is then used to fine-tune the merged model in this stage. Unlike the fine-tuning in stage 2, this phase involves only a brief period of fine-tuning, less than one epoch. The rationale for incorporating additional brief fine-tuning is that while layer merging maintains parameter integrity and optimize model's focus, it lacks the fine-tuned parameters necessary to enhance HPE prediction accuracy. In our approach, the merging model can be quickly fine-tuned to deliver accurate numerical predictions. The final output model of this stage is HPE-CogVLM as shown in Figure 2.

E. Stage 5: Evaluation of HPE-CogVLM on Test Data

To demonstrate the robustness of our model, we utilize real-world CMU Panoptic images to evaluate the model's performance on the HPE task. Meanwhile, we use rehearsal test datasets to assess the model's performance on the BBox prediction task.

Compared to directly LoRA fine-tuning and some model merging methods, this framework ensures that HPE-CogVLM maintains a high level of accuracy in HPE task, effectively mitigating the issues of blended invalid outputs and catastrophic forgetting problem. Experimental results will be shown in the section V.

IV. Experiments Setup

A. HPE Task Prompt Design

In some traditional CNN-based models, such as 6DRepNet and HopeNet, cropping the human head region is required as the initial step. In this paper, a new prompt method is proposed, allowing us to train HPE task utilizing the information of full images. In our prompts, BBox coordinates are leveraged to specify the human head of interest when multiple people are present. Therefore, the system is capable of effectively focusing on specific heads, which makes it easier to reduce the need for labor-intensive manual annotations and automate the inference process. Meanwhile, the global features from self-attention and head of interest features from cross-attention are both learnt to improve the robustness of HPE task. Figure 1 (b) illustrates a sample of our custom-designed prompts and responses tailored for the HPE task, demonstrating how the system interprets and responds to specific queries. Additionally, table II provides more examples of prompts and responses for the HPE and BBox prediction tasks. The BBox

TABLE II: Prompts and Responses Design for HPE Task.

BBox Prediction Task	
Designed Prompt	How many human heads are in this image and what are the head bounding boxes?
Correct Answer	Their head bounding boxes are [[106,168,148,242;245,168,270,230]].
Invalid Answer (Reason)	[[000,111,222,333... (*Recycled output error*) {112,432,211} (*Angle format output error*) A man in Red (*NLP output error*) [[212,123,212} (*Mixed output error*) [[234,134,100,111]] (*Logical error*)
Head Pose Estimation Task	
Designed Prompt	What is the head yaw pitch roll inside the bounding box [[106,168,148,242]]?
Correct Answer	The head orientation angles are {072,354,002}.
Invalid Answer (Reason)	{112,432,211,201} (*Wrong number error*) [[234,134,100,111]] (*BBox format output error*) A person head (*NLP output error*) [[212,123,212} (*Mixed output error*) {999,389,001} (*Logical error*)

format adheres to the specifications set by CogVLM [26]. Euler angles in the output are first converted to positive floats. These values are then rounded to the nearest integer, formatted as strings with a fixed length of three characters, padded with zeros where necessary. This table also includes several typical examples of invalid answers to highlight how invalid answers can lead to completely ineffective outputs when multiple grounding tasks requires accurate numerical output varied in range and quantity. In response to these issues, we define a new metric described in Section IV-D to assess the model's availability.

B. Datasets

Table I outlines datasets used in various stages of our framework. The CrowdHuman dataset [57] serves as the pre-training dataset due to its extensive collection of human images. Its head pose annotations are derived from pseudo-labels inferred by the pre-trained 6DRepNet [15] model, and thus are referred as weak label images. It enables the model to obtain the capability of detecting real human heads and warm up the HPE task in stage 1. The synthetic Agora dataset [21] serves as the task-specific HPE images, which encompasses full-range of human head yaw angle images and provides the GT of SMPL-X parameters [59]. Its head pose annotations are generated using the method of DirectMHP [20][2]. This dataset is used in stages 2 and 4 of our framework to enhance HPE accuracy. The Refcoco [28], Refcoco+ [28], and Refcocog [58] train datasets, which are originally utilized by CogVLM to learn BBox prediction, are chosen as rehearsal images to help mitigate the catastrophic forgetting of existing BBox capability. In our experiments, various portions of the rehearsal images are applied to determine the optimal rehearsal ratio [29], [30].

For evaluation, a subset of the CMU Panoptic dataset serves as the test dataset for evaluating HPE task, as its panoptic images of real people closely mirror real-life scenarios. The selection of images and labels follows the DirectMHP approach [20]. To evaluate object BBox localization, the test datasets including testA and testB data from Refcoco and Refcoco+, as well as the test dataset from Refcocog, are selected as the BBox evaluation datasets.

[2]https://github.com/hnuzhy/DirectMHP

C. Implementation Details

Throughout the LoRA fine-tuning process, a LoRA rank of 10 is used. The learning rate of 1×10^{-4} is used in pre-training stage. All other training parameters follow the default settings of the CogVLM. The experiments are performed on using two NVIDIA A100 80GB GPUs with a training batch size of 8. The training processes in stages 1, 2, and 4 of our framework cost 20, 50, and 10 hours, respectively.

D. Evaluation Metrics

We define four evaluation metrics for assessing HPE and BBox prediction tasks as follows:

Angle Error Ratio ($\mathrm{E_{angle}}$): $\mathrm{E_{angle}} = \frac{e_{\mathrm{angle}}}{t_{\mathrm{angle}}}$, where e_{angle} denotes the number of invalid HPE answers and t_{angle} denotes the number of total HPE answers. This new metric is defined to assess the capability of models to provide relevant numerical outputs for HPE task. When we prompt with a HPE query, the CogVLM could produce irrelevant responses such as an natural language processing (NLP) task response like "a person head", a BBox task response like "[[111,222,333,444]]", or a blended response like "[[111,999,999,99}" as shown in Table II.

BBox Error Ratio ($\mathrm{E_{bbox}}$): $\mathrm{E_{bbox}} = \frac{e_{\mathrm{bbox}}}{t_{\mathrm{bbox}}}$, where e_{bbox} denotes the number of invalid BBox answers and t_{bbox} denotes the number of total BBox answers. This new metric is defined to assess the capability of models to provide relevant numerical outputs for BBox prediction task.

BBox accuracy (ACC.): Acc. = $\frac{m}{\hat{m}}$, where m denotes the number of valid BBox answers with IoU > 0.5 and $\hat{m}$ denotes the number of total valid BBox answers. A BBox prediction is considered to be accurate if the intersection over union (IoU) between the GT and the prediction exceeds 0.5 [28]. And the invalid answers are excluded from accuracy and MAE calculation.

MAE of Euler angles (MAE): For the HPE task, the MAE between the GT Euler angles and the predicted Euler angles is defined as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \min(360^\circ - |\hat{A}_i - A_i|, |\hat{A}_i - A_i|) \quad (1)$$

where $\hat{A}_i$ represents the GT's Euler angles, A_i represents the predicted Euler angles, and variable n denotes the number of valid HPE answers. The MAE is measured in a circular manner rather than linearly, leading to the inclusion of a term that minimizes the difference between the predicted and actual angle by considering a full 360-degree rotation [15], [20]. In this paper, the MAE value is considered as the average of MAE for yaw, pitch and roll Euler angles.

E. Baseline Methods

In this paper, three types of baseline methods are considered to be compared with our HPE-CogVLM.

Traditional CNN-based models, including 6DRepNet, HopeNet and WHENet, serve as the baselines for CNN-based approaches. The 6DRepNet model, recognized as the current SOTA, is specifically retrained and tested on the same Agora and CMU datasets used in the VLM experiment to ensure a fair comparison. This model is trained 100 epochs, and the best MAE is selected for baseline analysis with our HPE-CogVLM. The pre-trained models of HopeNet and WHENet are utilized because HopeNet scripts are hard-coded and WHENet training scripts are not publicly available.

Non-merging CogVLM, directly LoRA fine-tuned model without applying model merging technique, serves as a comparison point to evaluate the effectiveness of our merging approach versus the LoRA fine-tuning only method [29]–[31]. The difference of Non-merging CogVLM and our HPE-CogVLM methods is that the Non-merging CogVLM bypasses stages 2 and 3, instead it undergoes significantly more training iterations in stage 4 which is equal to the total iterations of stages 2 and 4 in the HPE-CogVLM framework. For examples, our HPE-CogVLM is fine-tuned 25k and 5k iterations in stages 2 and 4 respectively, while the Non-merging CogVLM is solely fine-tuned in stage 4 for 30k iterations. This ensures fair comparison with respect to HPE task training iterations.

Task Arithmetic (TA) merging CogVLM, which adheres to our framework but replacing the layer-based merging with TA based merging, is to provide a baseline for comparing our merging approach with another merging method. The TA merging process is chosen as it forms the foundation for many other merging algorithms [51], [52]. In this process, we set the lambda parameter of task arithmetic to 0.5 [47], assigning equal importance to both the BBox prediction task and the HPE task.

V. Experimental Results

A. Baseline Comparison

The results in Table III show the performance comparison between our HPE-CogVLM and various baselines described in Section IV-E. In comparison with traditional CNN-based models, our HPE-CogVLM presents the significantly lower MAE. The HPE-CogVLM MAE is 75.1%, 66.8%, and 31.5% lower than WHENet, HopeNet, and 6DRepNet, respectively. The CNN-based models also perform worse than other CogVLM-based models, which highlights the superior robustness of VLM-based models over CNN-based models.

In comparison with Non-merging CogVLM, HPE-CogVLM MAE is 10% lower than the Non-merging CogVLM. Meanwhile the $E_{\textbf{angle}}$ of our model is 2.5 times smaller than the Non-merging CogVLM. This indicates that our LoRA layer-based merging method is more proficient in HPE than the method that does not utilize any model merging technique. Regarding the BBox results, the HPE-CogVLM's BBox prediction accuracy in test datasets is 0.6%, 0.5% and 1.1% lower than the Non-merging CogVLM, however, the Non-merging CogVLM costs five times more iterations for training rehearsal images as discussed in Section IV-E. This demonstrates that even with only 1/5 of the rehearsal image training iterations, our HPE-CogVLM achieves a comparable level of BBox accuracy.

Comparing with TA merging CogVLM, our HPE-CogVLM wins in all the metrics. For instance, when evaluated on test datasets, the BBox prediction accuracy of the HPE-CogVLM exceeds that of the TA merging CogVLM by 1%, 2.4%, and 1.7%, respectively. For the HPE task performance, $E_{\textbf{angle}}$ of TA merging CogVLM is 68.9%, which is 1325 times larger than that of HPE-CogVLM, indicating that only 31.1% of the responses for the HPE task are valid. Due to the high number of invalid HPE responses, the MAE metric becomes ineffective for assessing the performance. This highlights that even with an additional round of fine-tuning in stage 4, the TA merging fails to produce relevant numerical responses within our research domain, ultimately proving ineffective for the HPE task. In contrast, by applying our LoRA layer-based merging method, HPE-CogVLM successfully achieves the lowest MAE and invalid output ratio, demonstrating the superiority of this approach.

B. Catastrophic Forgetting Pattern in HPE Task

Table IV illustrates the profound impact of catastrophic forgetting in a model trained for HPE task only using Agora dataset. The Refcoco test accuracy starts at a high of 91.4% at iteration 0, indicating initial proficiency in object detection. As the number of training iteration increases and the model is increasingly exposed to the HPE task, the Refcoco test accuracy drastically decreases to 10.8% at iteration 1000. This sharp decline illustrates significant forgetting of the original BBox knowledge. $\mathbf{E}_{\textbf{bbox}}$ rises significantly from 0% to 36.2% at iteration 500 and then decreases to 10.2% at iteration 1000. This trend suggests that the model initially adapts to the new task at the expense of previously learned behaviors, causing a temporary increase in errors before stabilizing. The MAE improves from "Not capable" at iteration 100 to 42.16 at iteration 1000, indicating that the model begins to gain proficiency in the new task. The decline in $E_{\textbf{angle}}$ from 2.5% to 0.1% implies that the model's HPE output format becomes more consistent over time.

TABLE III: Comparison of HPE-CogVLM performance with various baselines. The best results are highlighted in bold.

Model	Refcoco		Refcoco+		Refcocog		CMU Panoptic	
	$\mathbf{Acc_{test}}$	$\mathbf{E_{bbox}}$	$\mathbf{Acc_{test}}$	$\mathbf{E_{bbox}}$	$\mathbf{Acc_{test}}$	$\mathbf{E_{bbox}}$	$\mathbf{MAE_{test}}$	$\mathbf{E_{angle}}$
WHENet	-	-	-	-	-	-	29.55	-
HopeNet	-	-	-	-	-	-	22.16	-
6DRepNet	-	-	-	-	-	-	10.74	-
Original Grounding CogVLM	91.4%	0%	86.7%	0%	90.2%	0%	-	-
Non-merging CogVLM	91.1%	0%	85.2%	0%	88.9%	0%	8.18	0.13%
TA merging CogVLM	89.5%	0%	82.3%	0%	86.1%	0%	7.72	68.9%
HPE-CogVLM	90.5%	0%	84.7%	0%	87.8%	0%	**7.36**	**0.052%**

Note: A dash ("-") indicates that the model is either not capable of performing the specified task or not applicable to the specified metric. For instance, WHENet, HopeNet, and 6DRepNet only accept detected human head bounding boxes for HPE prediction, meaning they cannot perform bounding box detection. As a result, their performance on the Refcoco, Refcoco+, and Refcocog tasks is marked with a "-". This applies to all the tables in the following results subsections.

TABLE IV: The impact of catastrophic forgetting when no data rehearsal are applied.

Iterations	$\mathbf{Acc_{test}}$	$\mathbf{E_{bbox}}$	$\mathbf{MAE_{test}}$	$\mathbf{E_{angle}}$
0	91.4%	0%	-	-
100	91.3%	0%	-	-
500	28.1%	36.2%	41.16	2.5%
1000	10.8%	10.2%	42.16	0.1%

TABLE V: Performance of weak label CogVLM under various rehearsal ratios.

Iterations	Rehearsal Ratio	$\mathbf{Acc_{test}}$	$\mathbf{E_{bbox}}$	$\mathbf{MAE_{test}}$	$\mathbf{E_{angle}}$
0k	0%	91.4%	0%	-	-
10k	0%	21.8%	0.026%	17.20	0.48%
10k	1%	77.5%	0.19%	21.51	0.85%
10k	10%	91.0%	0%	19.32	0.32%
10k	25%	91.5%	0%	19.92	0.23%

What is particularly noteworthy in this scenario is the nature of forgetting and learning displayed by the model—old knowledge is significantly diminished before new knowledge is solidified. This contrasts with human learning process, where new and old knowledge often coexist and can even enhance each other's acquisition. In human cognition, learning new tasks frequently involves integrating new information with existing knowledge, without the catastrophic forgetting seen in this model.

C. Selecting Optimal Rehearsal Ratios for Mitigating the Catastrophic Forgetting Problem

Table V presents the performance of weak label CogVLM across different proportions (0%, 1%, 10%, 25%) [29]–[31] of the rehearsal dataset in stage 1. The primary aim is to determine the appropriate data rehearsal ratio to retain old knowledge for the fine-tuning in stage 4. The Refcoco test accuracy at iteration 0 is 91.4%, indicating proficiency with the BBox prediction tasks. After the training is finished, the results demonstrate a clear trend that as the rehearsal ratio increases, the Refcoco test accuracy substantially improves. Starting at a low of 21.8% when no Refcoco data is used, the accuracy spikes to 77.5% with just 1% of rehearsal ratio, eventually reaching over 91% with 10% and 25% of rehearsal ratio. This clearly shows that the more original task data used in learning a new task, the less catastrophic forgetting occurs. In the $\mathbf{E_{bbox}}$ column, the consistently low $\mathbf{E_{bbox}}$ values suggest that the availability of BBox predictions tend to stabilize after 10K iterations. MAE and $E_{\mathbf{angle}}$ for HPE task show a fluctuating trend. Since the head pose weak label is provided for this pre-training stage, they may not fully reflect the model's true HPE performance. Rehearsal ratios of 10% and 25% are selected for the stage 4 experiment due to high refcoco BBox prediction accuracy. These ratios are significantly higher than the commonly used 1% rehearsal ratio in non-grounding tasks [29], [30].

D. The Influence of Rehearsal Ratios on Multiple grounding task Learning

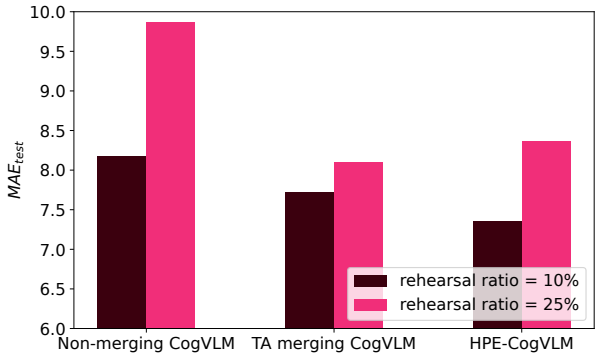
Figure 3 presents comparative results for both the BBox prediction task and the HPE task under two different rehearsal ratios in stage 4. Between the two HPE-CogVLM models, the one with a lower rehearsal ratio (10%) achieves a better MAE of 7.36, which is 12% lower than the 8.36 observed with the higher (25%) ratio. Conversely, the Refcoco test accuracy improves slightly with higher rehearsal ratios, showing increases of 0.3% compared to the lower ratio. The similar phenomenon also presents in Non-merging CogVLM and TA merging CogVLM results. The results clearly show that a higher rehearsal ratio helps retain existing knowledge by incorporating more data from previous tasks into fine-tuning, but this comes at the cost of new task performance. So we seek for balance between the retention of old knowledge and the performance on new tasks. In our case, the 10% rehearsal ratio achieves significantly better HPE performance, while the BBox prediction is only slightly better with the 25% rehearsal ratio. After balancing both factors, the HPE-CogVLM trained with the 10% rehearsal ratio is chosen as the optimal model.

E. Performance of HPE-oriented CogVLM on HPE Task Only

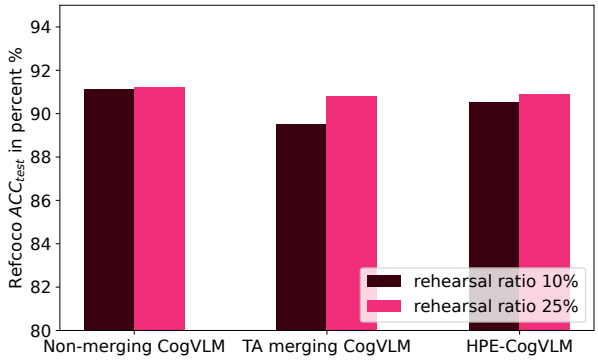
In our framework, The HPE-oriented CogVLM from stage 2 stands out as the most effective model dedicated solely to the HPE task. Table VI presents comparative performance results of 6DRepNet and the HPE-oriented CogVLM, both not accommodate BBox prediction capabilities, over similar training epochs. The low Refcoco test accuracy of HPE-oriented CogVLM is expected, given that no data rehearsal

TABLE VI: The HPE-oriented CogVLM model exhibits the highest HPE performance within our framework. The best results are highlighted in bold.

Model	Epochs	Acc_{test}	MAE_{test}	MAE_{train}	E_{angle}
6DRepNet	3	-	12.70	9.40	-
6DRepNet	6	-	12.76	8.80	-
6DRepNet	9	-	11.44	7.90	-
6DRepNet	50	-	11.37	2.91	-
6DRepNet	100	-	11.4	2.23	-
HPE-oriented CogVLM	3	8.8%	6.40	-	0.0092%
HPE-oriented CogVLM	6	12.6%	6.31	-	0%
HPE-oriented CogVLM	9	11.0%	**6.24**	-	0%



(a) The influence of rehearsal ratio on MAE.



(b) The influence of rehearsal ratio on BBox Prediction.

Fig. 3: The model performance under various rehearsal ratios (10% and 25%). (a) shows the MAE results under rehearsal ratio 10% and 25% on VLMs. (b) shows the Refcoco Test BBox accuracy results under rehearsal ratio 10% and 25% on VLMs.

is implemented in this stage so that the VLM can focus on the HPE task learning. In terms of MAE metric, the HPE-oriented CogVLM displays a gradual decrease in MAE from 6.4 at 3 epochs to 6.24 at 9 epochs. When compared our model with 6DRepNet, in the same epoch, our MAE shows much lower numbers than 6DRepNet. For example, in epoch 9, MAE of HPE-oriented CogVLM is 6.24 which is 45.5% lower than 6DRepNet. After extending the 6DRepNet training to 100 epochs, while its training MAE decreases from 9.40 to 2.23, the MAE on the CMU dataset does not improve, remaining stable around 11.4. This indicates that the model is overfitting to the training dataset, with no enhancement in cross-dataset inference performance. This difference emphasizes the superior performance of VLM than traditional CNN-based models.

F. Visualization of Cross Attention Maps Supervised by Designed Prompts

Fig. 4: The visualization displays cross attention maps generated in response to our custom prompts. The left image shows the attention map associated with the prompt "What is the head yaw pitch roll inside the bounding box [[335,179,445,332]]?" (BBox for the person on the left), and the right image corresponds to the prompt "What is the head yaw pitch roll inside the bounding box [[775,105,893,261]]?" (BBox for the person on the right).

Figure 4 provides a visual representation of cross attention maps created in response to specific prompts when multiple people are present in the image. The left image highlights the model's response to the prompt of HPE task within the BBox [[335,179,445,332]]. It effectively focuses on the head of the person on the left. The right image similarly demonstrates the model's precision in targeting the head of the person on the right within the BBox [[775,105,893,261]]. These visualizations confirm the model's accuracy in localizing attention within specified areas, demonstrating strong capabilities in spatial awareness. Moreover, they illustrate that CogVLM is capable not only of generating BBox outputs in response to queries but also of interpreting and responding to BBoxes specified within the prompts.

VI. Conclusions

In this paper, we present a framework that successfully enhances the HPE task by leveraging the visual grounding capabilities of CogVLM. Through a novel merging approach

that utilizes a high cosine similarity threshold and a "winner-takes-all" layer selection strategy, we effectively integrate HPE capabilities into the model while preserving the original BBox knowledge. This method not only improves prediction accuracy but also addresses the challenges of blended invalid response formats. Additionally, we mitigate catastrophic forgetting by optimizing the rehearsal ratio. Our experimental results demonstrate that HPE-CogVLM achieves a substantial 31.5% reduction in MAE compared to the current CNN-based state-of-the-art in cross-dataset evaluations. Furthermore, HPE-CogVLM consistently outperforms directly LoRA fine-tuned model and task arithmetic-based merging model in all HPE metrics, establishing it as a robust and effective solution for complex multimodal grounding tasks.

REFERENCES

[1] T. Fischer, H. J. Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 334–352.

[2] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6912–6921.

[3] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 100–115.

[4] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4838–4846.

[5] I. Masi, F.-J. Chang, J. Choi, S. Harel, J. Kim, K. Kim, J. Leksut, S. Rawls, Y. Wu, T. Hassner, W. AbdAlmageed, G. Medioni, L.-P. Morency, P. Natarajan, and R. Nevatia, "Learning pose-aware models for pose-invariant face recognition in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 379–393, 2019.

[6] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 802–815, 2012.

[7] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4511–4520.

[8] M. C. d. F. Macedo, A. L. Apolinário, and A. C. d. S. Souza, "A robust real-time face tracking using head pose estimation for a markerless ar system," in *2013 XV Symposium on Virtual and Augmented Reality*, 2013, pp. 224–227.

[9] S. Wu, J. Liang, and J. Ho, "Head pose estimation and its application in tv viewers' behavior analysis," in *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 2016, pp. 1–6.

[10] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 300–311, 2010.

[11] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 3, pp. 254–265, 2018.

[12] T. Hu, S. Jha, and C. Busso, "Temporal head pose estimation from point cloud in naturalistic driving conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 8063–8076, 2021.

[13] ——, "Robust driver head pose estimation in naturalistic conditions from point-cloud data," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. Ieee, 2020, pp. 1176–1182.

[14] D. Strazdas, J. Hintz, and A. Al-Hamadi, "Robo-hud: Interaction concept for contactless operation of industrial cobotic systems," *Applied Sciences*, vol. 11, no. 12, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/12/5366

[15] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "6d rotation representation for unconstrained head pose estimation," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, Oct. 2022. [Online]. Available: http://dx.doi.org/10.1109/ICIP46576.2022.9897219

[16] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," *CoRR*, vol. abs/1710.00925, 2017. [Online]. Available: http://arxiv.org/abs/1710.00925

[17] Y. Zhou and J. Gregson, "Whenet: Real-time fine-grained estimation for wide range head pose," 2020.

[18] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3d total solution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, p. 78–92, Jan. 2019. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2017.2778152

[19] G. Fanelli, M. Dantone, and L. Van Gool, "Real time 3d face alignment with random forests-based active appearance models," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

[20] H. Zhou, F. Jiang, and H. Lu, "Directmhp: Direct 2d multi-person head pose estimation with full-range angles," 2023.

[21] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black, "Agora: Avatars in geography optimized for regression analysis," 2021.

[22] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social interaction capture," 2016.

[23] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," 2024.

[24] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: A family of highly capable multimodal models," 2024.

[25] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[26] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, and J. Tang, "Cogvlm: Visual expert for pretrained language models," 2024.

[27] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh, "Vqa: Visual question answering," 2016.

[28] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," 2016.

[29] T. Scialom, T. Chakrabarty, and S. Muresan, "Fine-tuned language models are continual learners," 2022.

[30] J. Huang, L. Cui, A. Wang, C. Yang, X. Liao, L. Song, J. Yao, and J. Su, "Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal," 2024.

[31] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang, "An empirical study of catastrophic forgetting in large language models during continual fine-tuning," 2024.

[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.

[33] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009. [Online]. Available: https://dl.acm.org/doi/10.5555/1577069.1755843

[34] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 2074–2083. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018_workshops/w41/html/Ruiz_Fine-Grained_Head_Pose_CVPR_2018_paper.html

[35] Y. Zhou and J. Gregson, "Whenet: Real-time fine-grained estimation for wide range head pose," in *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020. [Online]. Available: https://www.bmvc2020-conference.com/assets/papers/0907.pdf

[36] H. Hu, X. Wu, Y. Wang, Y. Fang, and H. Wu, "Mathematical foundation and corrections for full range head pose estimation," *CoRR*, vol. abs/2403.18104, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.18104

[37] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, "Quatnet: Quaternion-based head pose estimation with multiregression loss," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1035–1046, 2019.

[38] B. Huang, R. Chen, W. Xu, and Q. Zhou, "Improving head pose estimation using two-stage ensembles with top-k regression," *Image and Vision Computing*, vol. 93, p. 103827, 2020.

[39] H. Zhang, M. Wang, Y. Liu, and Y. Yuan, "Fdn: Feature decoupling network for head pose estimation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 789–12 796.

[40] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1087–1096.

[41] Z. Cao, Z. Chu, D. Liu, and Y. Chen, "A vector-based representation to enhance head pose estimation," 2020.

[42] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," 2023.

[43] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, "Shikra: Unleashing multimodal llm's referential dialogue magic," 2023.

[44] H. You, H. Zhang, Z. Gan, X. Du, B. Zhang, Z. Wang, L. Cao, S.-F. Chang, and Y. Yang, "Ferret: Refer and ground anything anywhere at any granularity," 2023.

[45] P. Yin, G. Zeng, J. Wang, and D. Xie, "Clip-gaze: Towards general gaze estimation via visual-linguistic model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 6729–6737.

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[47] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing models with task arithmetic," 2023.

[48] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, and L. Schmidt, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," 2022.

[49] D.-H. Jang, S. Yun, and D. Han, "Model stock: All we need is just a few fine-tuned models," 2024.

[50] M. Davari and E. Belilovsky, "Model breadcrumbs: Scaling multi-task model merging with sparse masks," 2023.

[51] P. Yadav, D. Tam, L. Choshen, C. A. Raffel, and M. Bansal, "Ties-merging: Resolving interference when merging models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[52] L. Yu, B. Yu, H. Yu, F. Huang, and Y. Li, "Language models are super mario: Absorbing abilities from homologous models as a free lunch," *arXiv preprint arXiv:2311.03099*, 2023.

[53] T. Akiba, M. Shing, Y. Tang, Q. Sun, and D. Ha, "Evolutionary optimization of model merging recipes," *arXiv preprint arXiv:2403.13187*, 2024.

[54] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[55] Z. Zhang, M. Fang, L. Chen, and M.-R. Namazi-Rad, "CITB: A benchmark for continual instruction tuning," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9443–9455. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.633

[56] J. Mok, J. Do, S. Lee, T. Taghavi, S. Yu, and S. Yoon, "Large-scale lifelong learning of in-context instructions and how to tackle it," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 12 573–12 589. [Online]. Available: https://aclanthology.org/2023.acl-long.703

[57] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd," *arXiv preprint arXiv:1805.00123*, 2018.

[58] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," 2016.

[59] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," 2019.