Exploring the Zero-Shot Capabilities of Vision-Language Models for Improving Gaze Following

Anshul Gupta* Pierre Vuillecard* Arya Farkhondeh Jean-Marc Odobez
Idiap Research Institute, Martigny, Switzerland
École Polytechnique Fédérale de Lausanne, Switzerland
{agupta, pvuillecard, afarkhondeh, odobez}@idiap.ch

Abstract

Contextual cues related to a person's pose and interactions with objects and other people in the scene can provide valuable information for gaze following. While existing methods have focused on dedicated cue extraction methods, in this work we investigate the zero-shot capabilities of Vision-Language Models (VLMs) for extracting a wide array of contextual cues to improve gaze following performance. We first evaluate various VLMs, prompting strategies, and in-context learning (ICL) techniques for zero-shot cue recognition performance. We then use these insights to extract contextual cues for gaze following, and investigate their impact when incorporated into a state of the art model for the task. Our analysis indicates that BLIP-2 is the overall top performing VLM and that ICL can improve performance. We also observe that VLMs are sensitive to the choice of the text prompt although ensembling over multiple text prompts can provide more robust performance. Additionally, we discover that using the entire image along with an ellipse drawn around the target person is the most effective strategy for visual prompting. For gaze following, incorporating the extracted cues results in better generalization performance, especially when considering a larger set of cues, highlighting the potential of this approach.

1. Introduction

Understanding where a person is looking in a scene, also known as gaze following, has diverse applications, including human-robot interaction [1, 18, 34], conversation analysis [11, 27], and the study of neurodevelopmental disorders [7, 21]. However, this is a challenging task, demanding a model to interpret a large spectrum of contextual cues such as the person's interactions with objects and other people in the scene. For instance, it has been shown that eye and hand movements are coordinated during manipulation activities [19]. Or that during conversations, people usually look at the person talking [37] and that leveraging this information can help in gaze target selection in meeting settings [28]. As seen in Figure 1, estimating the child's gaze target requires understanding their head and body pose, the interaction between the child and the adult through pointing and shared attention etc.

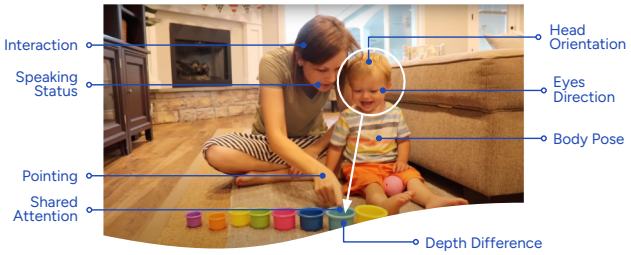

Figure 1. As humans, we rely on various sources of information to predict a person's gaze target. This image shows contextual information that could be valuable.

In order for a model to capture these cues and learn their impact on gaze target selection, it would need to be trained on a high-quality and large-scale labeled dataset. However, existing gaze following datasets [7, 39] are small-scale, hindering the effective utilization of these cues. To address these challenges, prior works have relied on dedicated cue extraction methods such as inferred body pose [13, 24], and supplied them to models for improved performance. However, these approaches focus on specific cues and do not supply the larger array of contextual cues which could be needed for accurate gaze target prediction. Traditional methods to address this limitation include: (1) manually annotating for relevant cues; but it is cost-intensive and not always available during inference, or (2) pseudo-labelling with expert models; however this requires access to a multitude of task specific models for each cue or a subset of cues. Hence, it is evident that novel solutions are required.

* indicates equal contribution

Given these challenges, we investigate the potential of Visual Language Models (VLMs) to extract valuable contextual cues for gaze following, aiming to overcome the constraints of traditional labeling approaches. VLMs have shown promising zero-shot performance for a variety of tasks [30, 45], owing to their ability to learn visual-text associations at scale. Hence, a single model may be capable of extracting all relevant contextual cues. At the same time, given the zero-shot setting, the set of cues to be considered can be adjusted based on the domain, further increasing this approach's applicability.

In this work, we consider cues related to pose, person-person interactions, and person-object interactions. We first evaluate the zero-shot performance of different VLMs for recognising these cues (Section 3), and leverage the best performing approach to extract them. We then investigate whether incorporating these extracted cues can improve gaze following performance (Section 4).

Challenges. While VLMs have shown impressive zero-shot performance for a variety of tasks, these tasks (ex. image classification) usually involve processing the entire image. However, for accurate gaze following we also need to capture contextual cues related to each person in the scene. Hence, we need to consider an appropriate visual prompt to allow the VLM to focus on the person of interest. At the same time, it is important to consider the choice of text prompt as VLMs have been shown to benefit from prompt engineering [29]. Finally, given the extracted cues from the VLMs, we need to consider how to incorporate such information into a gaze following model. Following these research questions, we make the following contributions:

- *VLMs for contextual cues extraction*: We explore 4 state of the art VLMs [22, 23, 30] for this task. We also investigate different visual prompts to focus on the person of interest, and different text prompts to describe the cue of interest. We show that VLMs can indeed capture contextual cues although the choice of VLM, visual prompt and text prompt impacts performance.
- *Text improved Gaze Following*: We incorporate the extracted contextual cues into a recent transformer based gaze following model [14]. Our results indicate that incorporating these cues can result in better generalization performance, especially when considering larger sets of cues.

2. Related-Work

Vision-Language Models (VLMs). VLMs began receiving significant attention following the introduction of CLIP (Contrastive Language-Image Pretraining) [30]. This contrastive learning framework learns effective multi-modal representations using image-text pairs. CLIP demonstrated impressive zero-shot classification performance on standard image classification benchmarks. Subsequent VLMs, such as BLIP [22] and BLIP-2 [23], have been introduced with notable differences from CLIP. These include varied objectives (incorporating additional losses beyond the contrastive approach), the use of more curated training datasets, and enhanced image caption generation through caption filtering. Specifically, BLIP-2 has introduced advanced pre-training strategies, integrating a frozen image encoder and a Querying Transformer (Q-Former), which enables a nuanced extraction of visual representations. They have even shown strong performance for video tasks such as action recognition [30] and text-to-video retrieval [22] despite not processing the temporal dimension. However, their performance for localizing the actions/cues of people in an image has not been explored.

While there has been some work on video language models, they face certain limitations. Firstly, available video-text pairs for pretaining is limited compared to the scale of available image-text pairs [44], so these methods do not generalize as well. To cope with this issue, some works leveraged pre-trained image-based VLMs and adapted them for video input, however, it harmed their zero-shot performance [20, 42]. Secondly, these models are computationally more expensive, and cannot be applied on static gaze following datasets.

More recently, VLMs such as BLIP2 have leveraged Large language models (LLMs) for generating textual output. LLMs have shown an impressive ability to act as models of the world, with a rudimentary understanding of agents, beliefs and actions [3], and an ability to perform commonsense and mathematical reasoning [8]. Hence, they may also be capable of capturing complex relationships between people and objects in the scene to better extract contextual cues for human behaviour understanding. Further, they have been shown to benefit from in-context learning [5] (ICL), wherein a few demonstrations of the task are provided at inference time without any weight updates. VLMs that exploit such LLMs have also displayed improvements in performance using ICL [2, 41], by being provided with a few sample visual and/or textual demonstrations. It is still a research question whether LLMs and ICL can improve the recognition of gaze contextual cues.

Visual and Textual Prompting. The recent work of Shtedritski et al. [35] explored different visual prompting approaches for CLIP. They compared cropping the visual area of interest versus drawing a red ellipse around it, and found the red ellipse approach to perform better for keypoint localization and referring expression comprehension tasks. They also observed that blurring or graying the region outside the ellipse can provide additional benefits. However, the performance of these approaches for action/cue recognition hasn't been explored. On the textual prompting side, the original CLIP paper [30] showed that prompt engineering, such as using the prompt 'a photo of a {label}' improved perfor-

mance over using just the label text on ImageNet [33]. They also observed that ensembling different prompts in the embedding space can reliably improve performance. Recent works [46, 47] introduced learnable prompts that were fine-tuned on a specific task. However this approach requires data to adapt, and hence cannot be applied in a zero-shot manner. Also, although the learnable prompts can then be included with prompts for unseen classes, the results lag behind manual prompt engineering efforts [46]. In this work, we evaluate different manual prompt engineering approaches that can be applied to any new set of classes.

Contextual Cues for Gaze Following. Previous works in estimating the Visual Focus of Attention (VFOA) of a person leveraged cues such as the head pose and speaking information of people [28, 34, 38] for improved performance. However, these methods typically require access to frontal views of people and knowledge of the 3D scene structure (ex. using multiple calibrated cameras) for inference. This makes it challenging to deploy these algorithms in new environments where such information may not be available.

Hence, Recasens et al. [32] proposed the Gaze Following task to estimate the scene gaze target of a person in an image using only the image and with no prior assumptions about the scene or camera placement. However, due to the complexities of scenes and the lack of data, models have encountered challenges in capturing pertinent information, ultimately leading to sub-optimal predictions. Hence, recent methods have shown that inferred cues such as depth [4, 10, 13, 15, 18, 39, 40] and body pose [13, 24] can be leveraged for improved performance. However, incorporating person-specific auxiliary information in these models is not straightforward.

More recently, [14] proposed a new transfomer based model for gaze following and social gaze prediction. They showed that this architecture allows for easily incorporating person-specific auxiliary cues, with improved performance from the addition of people's speaking status. Whether it can benefit from additional cues remains to be explored.

3. Contextual Cues Extraction

In the first stage, we evaluate the zero-shot performance of different VLMs and prompting approaches for recognition of cues. Note that we interchangeably refer to a specific cue as a class.

3.1. Method

We investigate different visual and textual prompting strategies, as well as two different variants of VLMs for zero-shot contextual cues extraction, namely image-text matching (ITM) and visual question-answering (VQA).

Visual Prompting. In a complex scene involving multiple people, ITM becomes challenging as our task requires conditioning on a specific target person. To address this, we investigate various visual prompting techniques that enable ITM to focus on a chosen individual. We employ several approaches, including no prompting, drawing a red ellipse around the person following [36], blurring or graying the background. These techniques are applied either to the entire image (image-based) or to the cropped target person (person-based), resulting in a total of eight distinct visual prompting approaches. We provide an example of the different visual prompts in Figure 10 in the supplementary.

Text Prompting. In our approach to text prompting, we employed a structured method for generating prompts systematically based on templates. This method allows us to meticulously examine the impact of each textual component within the prompt. A template, in this context, is a fixed sentence where only specific parts can be altered. For examples, "a {photo} of a {person} {class}", and "a {person} is {class}", are two instances of templates. Beyond the varied sentence structures, the placeholders {photo}, {person}, and {class} can be substituted with semantically related components. For instance, {photo} could be replaced with "picture" or "snapshot," {person} might be substituted with "individual" or "human," and {class} can refer to class synonyms such as "talking" or "narrating" if the original class is "speaking". In this work, synonyms refer to changes in class synonyms otherwise mentioned explicitly. In the supplementary, Figure 8 presents all the different templates and synonyms used in this section.

Image-Text Matching (ITM). In ITM, the objective is to compute a cosine similarity between the visual and textual embedding. A high similarity suggests that the image contains the textual description. Formally, given an image I of size $H \times W \times 3$ and a set of K class names, we use the visual and text encoders of a pre-trained VLM (e.g., CLIP) to get a visual embedding $e_I \in \mathbb{R}^d$ and K text embeddings $e_T \in \mathbb{R}^{K \times d}$. We perform the following matching:

$$S = dot(e_I \cdot e_T^{\mathrm{T}}) \tag{1}$$

where, $S \in \mathbb{R}^K$ are the resulting similarity scores. When multiple textual prompts refer to the same class name k, i.e. $e_{T_k} \in \mathbb{R}^{P \times d}$, we can perform an *Ensemble* to get the score.

$$S_k = dot(e_I \cdot \frac{1}{P} \sum_{e \in e_{T_k}} e^{\mathrm{T}}) \tag{2}$$

The *Ensemble* approach utilizes the mean embedding, acting as a centroid for a given class and thus is expected to be more robust. The scores for each class are then normalized across samples to have a zero mean and a standard deviation of one. In this work, we investigate three different pre-trained VLMs such as CLIP [30], BLIP [22] and BLIP-2 [23]. For more details regarding these models, we refer the readers to the original papers and details in Sec. 2.

Visual Question Answering (VQA). In order to explore

the potential of LLMs for our task, we investigate a recent VQA model, BLIP-2 VQA [23], that leverages a LLM called FlanT5 [31]. In VQA models, a textual question is jointly input with an image to the model, and the model outputs a textual answer. We convert the text prompts described previously into a set of questions that result in simple "yes" or "no" answers, which we then convert into a binary score. Examples of prompts are displayed in supplementary Fig 9. To further explore the benefits of ICL, we provide additional textual context in the form of a generated caption from the same model. Thus, the text input to the model is of the form {*generated caption*} {*text prompt*}. It is worth noting that the BLIP-2 VQA model is much slower to run than the ITM models as (1) the model is much larger due to the LLM, and (2) the answer is conditioned on the image *and* question, so we need to run a forward pass for each image-prompt pair. This is unlike the ITM models where the images and prompts can be processed separately, with a similarity score computed afterward.

3.2. Experiments

Datasets. We employ two datasets to shed light on the VLMs' ability to extract meaningful cues.

ChildPlay: We manually annotated 6 cues from the ChildPlay [39] dataset, which is a recently proposed dataset for gaze following. For each class, we selected around 50 clear positives and 50 clear negatives. The classes and statistics are presented in the supplementary Table 5.

AVA-Actions: Then, to scale our evaluation we used the validation split of the AVA dataset [12], which is a human action localization dataset. This dataset is much more challenging since it is heavily unbalanced and large scale containing around 41000 images. A subset of the classes of interest was selected. In Table 1, a summary of the dataset classes and distribution is shown.

Metrics. We leverage two metrics:

- *AP*: To evaluate the performance of different VLMs and prompting approaches, we use Average Precision (AP). It is computed per class between the ground truth and the scores obtained from the VLMs. We also consider the mean of the AP scores across all classes or mean Average Precision (mAP).
- *Accuracy*: Since the output of the VQA variants is a binary decision, we cannot compute AP; instead, we compute accuracy. To compare with ITMs, we binarize their output by applying a threshold of zero since the scores are normalized with a zero mean (however they may benefit from optimizing the threshold).

3.3. ITM Results

Visual prompting. We compare the performance of the different visual prompts described in Section 3.1 in Fig. 2. The results are aggregated across VLMs and different text prompts, and categorized by the type of visual input, i.e. image-based versus person-based. We see that image-based approaches outperform person-based variants. This suggests that a broader visual input provides additional context, enhancing the zero-shot recognition for the target person in the image. Furthermore, among the visual prompts, the red ellipse approach outperforms others, aligned with findings in [36]. Therefore, in subsequent experiments, we employ the image-based red ellipse as the visual prompt.

VLMs. We compare the performance of three VLMs, namely CLIP, BLIP, and BLIP-2. In Fig. 3, we present a class-wise comparison of the three VLMs on AVA. Note that, for each VLM, we aggregate the results from different

Selected Classes - AVA	Support
Pose (P)	
stand	23424
sit	16660
bend/bow (at the waist)	1512
Person-Person Interaction (P-P)	
talk to (e.g., self, a person, a group)	25985
hug (a person)	340
hand clap	330
give/serve (an object) to (a person)	313
Person-Object Interaction (P-O)	
carry/hold (an object)	17199
touch (an object)	5099
read	658
write	273
lift/pick up	118
text on/look at a cellphone	112
work on a computer	111

Table 1. Selected classes from the AVA Dataset (validation set) categorized as Pose (P), Person-Person Interaction (P-P), and Person-Object Interaction (P-O), including the number of samples (support) for each.

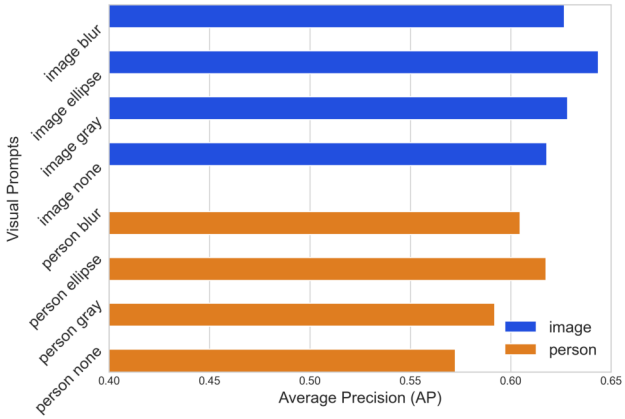

Figure 2. Results of different visual prompting approach on Childplay. *Image* corresponds to input the full image whereas *person* refers to the use of person crop as input.

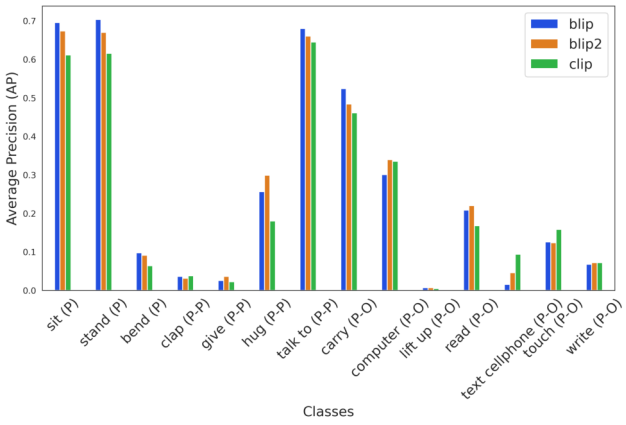

Figure 3. Results of different VLMs following the ITM approach on AVA. Three VLMs are compared across different classes categorized as Pose (P), Person-Person Interaction (P-P), and Person-Object Interaction (P-O).

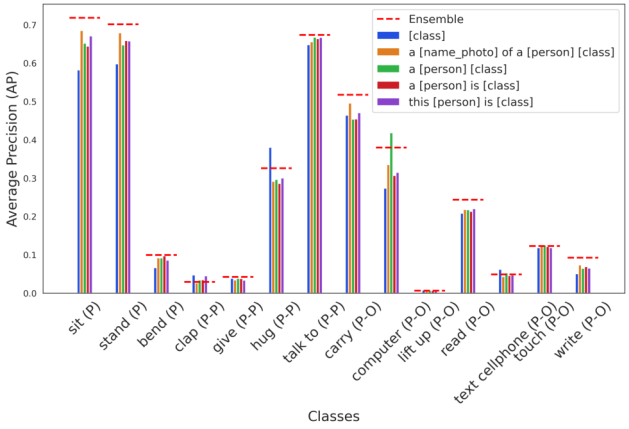

Figure 4. Results of different templates using BLIP-2 on AVA. Six templates are compared across different classes.

textual prompts. Firstly, we observe that no single model always outperforms the others. However, BLIP and BLIP-2 surpass CLIP in pose and person-to-person classes, while CLIP performs well when the class refers to a clear object, such as *work on a computer* or *text on a cellphone*. This may be related to differences in training data, and is a direction for investigation. On average, BLIP-2 is the top performing model. In the subsequent analysis, we continue focusing on BLIP-2 while varying the text prompting aspects.

Text prompting. We investigate the impact of the text prompts described in Section 3.1 at two different levels, at the template level and synonym level. When evaluating the template, we aggregate results over the other text prompt component variations. Similarly for when we evaluate the class synonym. In Figure 4, performance for different templates are shown on AVA per class. Firstly, there is no best template overall, which correlates to the finding of [30] that

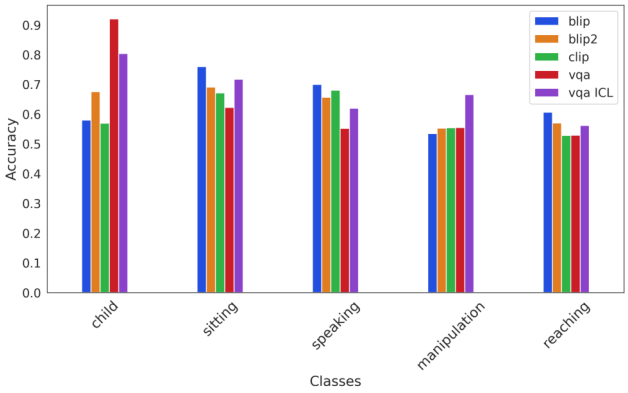

Figure 5. Results of BLIP-2 vqa with and without in-context learning, vqa ICL and vqa respectively, on ChildPlay. It is compared with the VLMs CLIP, BLIP and BLIP-2.

VLMs are prompt sensitive. However, using the *Ensemble* approach described in Section 3.1 provides more robust performance, often outperforming the best template, and always outperforming the worst template. In addition, the wording in textual prompts matters, as can be seen in the supplementary Figure 11, where different class synonyms can change the performance by a large margin. However, we notice that for most of the classes, including $\{person\}$ in the prompt improves performance. This suggests that conditioning the prompt to an individual helps to extract person-centric information.

3.4. VQA Results

To investigate the potential of LLMs and in-context learning for contextual cues extraction, we evaluate the BLIP-2 VQA model on the ChildPlay dataset, and compare it against ITM based VLM models (Figure 5). Note that the results are aggregated across all text prompts. As mentioned in Section 3.1, the BLIP-2 VQA model is much slower to run compared to the ITM based models which is why we use the smaller ChildPlay dataset. We also use a smaller set of templates and synonyms in the text prompt (Fig. 9 in supplementary) to reduce computation time.

Benefit of LLM. Comparing the performance of BLIP-2 against BLIP-2 VQA (BLIP-2 and vqa in the figure), we see that BLIP-2 VQA does much better for the 'child' class, but on par of worse for the other classes. This suggests that the LLM in the BLIP-2 VQA model is not necessarily providing better results. However, as mentioned previously, this model uses a smaller set of templates and synonyms in the text prompt for computational reasons so may benefit from using a larger set.

In-Context Learning. We see that the BLIP-2 vqa model with ICL improves for all classes except the 'child' class compared to no ICL. This is in contrast to the observations in the original paper where the architecture is intro-

duced [23], and suggests the potential of leveraging ICL for contextual cues extraction.

4. Text-Improved Gaze Following

In the second stage, we apply insights from Section 3 and leverage BLIP-2 along with the red ellipse visual prompting approach, and the *Ensemble* text prompting approach to extract contextual cues. We then evaluate the impact of incorporating these cues into a gaze following model.

4.1. Method

We employ the static version of the recently proposed MTGS [14] model. This model is a transformer-based architecture designed for multi-person gaze following and social gaze prediction. Given an input image and head crops of people in the scene, it first produces two types of tokens: image tokens ($\mathrm{x}_{\mathrm{image}} \in \mathbb{R}^{N \times D}$), similar to those in a standard Vision Transformer (ViT) architecture [9], and person gaze tokens ($\mathrm{x}_{\mathrm{gaze}} \in \mathbb{R}^{P \times D}$), where P represents the number of people in the scene. Person tokens are generated using head crops, a gaze backbone, and a subsequent linear projection layer. This formulation naturally supports incorporating contextual cues for each person, as the information can be fused with the corresponding person token.

Given the success of additive fusion in the case of position embeddings for transformers [9], and early fusion of body pose and depth information for gaze following models [13], we aim to incorporate contextual information derived from VLMs in an early fusion and additive manner. To this end, as illustrated in Fig. 6, we use a linear projection layer (Φ) to project the vector of predicted scores ($S_{vlm} \in \mathbb{R}^{P \times K}$, K is the number of classes) and generate person context tokens matching the dimensions of the person gaze tokens ($\Phi(S_{vlm}) \in \mathbb{R}^{P \times D}$). We then apply the add operation to combine the person context tokens from the VLMs with the corresponding person gaze tokens. Following this, the enriched person gaze tokens, now with added contextual cues, and the image tokens are fed into MTGS, where, people and scene tokens interact through self and cross-attention modules across multiple blocks.

$$\mathrm{x}_{\mathrm{out}} = \mathrm{MTGS}\left(\left[\mathrm{x}_{\mathrm{gaze}} + \Phi\left(\mathrm{S}_{\mathrm{vlm}}\right), \mathrm{x}_{\mathrm{image}}\right]\right) \quad (3)$$

Finally, a prediction module takes the updated tokens ($\mathrm{x}_{\mathrm{out}}$) and predicts the visual attention heatmap for each person, as well as pair-wise social gaze labels. For a more comprehensive understanding of the architecture, we direct readers to the original paper [14].

4.2. Experiments

Datasets. We leverage two gaze following datasets:

GazeFollow [32] is a large-scale static dataset for gaze following, featuring 122K images. Most images are annotated for a single person with their head bounding box and gaze target point. The test set contains annotations by multiple annotators. Despite lower quality images and annotations, it's diversity makes it a good dataset for pre-training.

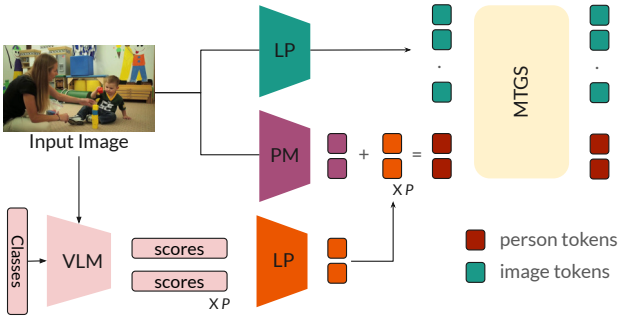

Figure 6. An overview of Text-Improved Gaze Following: Given an image containing P persons, image tokens and person tokens are generated via a Linear Projection (LP) and a person module (PM) respectively. To incorporate VLM contextual information, we use a VLM to obtain P score vectors, each with the dimension as the number of classes (K). We then linearly project these vectors and perform early fusion by adding them to the corresponding person tokens. Scene and updated person tokens are subsequently passed to MTGS [14] to model person and scene interactions using self and cross-attention modules across multiple blocks.

ChildPlay [39] is a recent video dataset for gaze following, featuring children playing and interacting with other children and adults. It is annotated with the head bounding box, gaze point and gaze label (consisting of 7 non-overlapping classes) of people in the scene.

Following [14], we pre-process both datasets to extract pair-wise social gaze labels for two tasks:

- *LAH:* It stands for looking at humans and occurs when a person looks at another person's head. In terms of positive/negative pair statistics, it is 27k/493k for GazeFollow, and 59k/682k for ChildPlay.
- *LAEO:* It stands for looking at each other and occurs when two people engage in mutual gaze. In terms of positive/negative pair statistics, it is 0/0 for GazeFollow, and 7k/351k for ChildPlay.

Contextual Cues. We define three sets of contextual cues:

- *AVA+CP:* These are the set of 24 cues defined in Section 3 for AVA and ChildPlay that were used for evaluating different VLMs and prompting strategies.
- *HICO:* The HICO dataset [6] is human-object interaction dataset that defines a list of 117 interaction verbs. We leverage these verbs as contextual cues.
- *SWIG:* The SWIG-HOI dataset [43] is a large-scale human-object interaction dataset that defines 406 verbs. We leverage these verbs as contextual cues.

We provide the manually curated synonyms and templates used for generating different text prompts for AVA+CP in Figures 8,9 of the supplementary. For HICO and SWIG, we use the same set of templates, but generate 4 synonyms for

Model	AUC↑	Avg. Dist↓	Min. Dist↓	$F1_{LAH}$↑
Fang [10]	0.922	0.124	0.067	-
Tonini [40]	0.927	0.141	-	-
Jin [18]	0.920	0.118	0.063	-
Bao [4]	0.928	0.122	-	-
Hu [16]	0.923	0.128	0.069	-
Tafasca [39]	0.936	0.125	0.064	-
Chong [7]	0.921	0.137	0.077	-
Gupta [13]	0.933	0.134	0.071	-
Jin [17]	0.919	0.126	0.076	-
MTGS [14]	0.929	0.118	0.062	0.639
MTGS + AVA + CP	0.936	0.118	0.061	**0.643**
MTGS + HICO	0.934	**0.116**	**0.060**	0.639
MTGS + SWIG	0.933	0.119	0.061	0.619

Table 2. Results for incorporating VLM context with different sets of classes on the GazeFollow dataset. AVA+CP has 24 classes, HICO has 117 classes and SWIG has 406 classes. Best results are given in bold, second best results are underlined.

each cue using ChatGPT [26].

Training and Validation. Following [14], we train the model for 20 epochs on GazeFollow using a learning rate of 1e-4 and the AdamW [25] optimizer. We supervise using the standard MSE loss for gaze heatmap prediction, and binary cross entropy loss for LAH prediction. For validation, we use the split proposed by [39].

Metrics. We use the standard gaze following metrics:

- *AUC*: the predicted heatmap is compared against a binary GT map with value 1 at annotated gaze point positions, to compute the area under the ROC curve.
- *Distance (Dist.)*: the arg max of the heatmap provides the gaze point. We can then compute the L2 distance between the predicted and GT gaze point on a 1×1 square. We compute Minimum (Min.) and Average (Avg.) distance against all annotations.

In addition, we compute F1 scores for LAH ($F1_{LAH}$) and LAEO ($F1_{LAEO}$). For LAH, we check if the predicted gaze point falls inside the target person's head bounding box. For LAEO, we check the reverse as well. [1]

4.3. Results

GazeFollow. We provide results for incorporating VLM context on the GazeFollow dataset in Table 2. We observe that performance does not change much for the distance score. In contrast, for LAH, we observe a slight improvement with the addition of AVA+CP cues, and a degradation with the addition of SWIG cues. However, the GazeFollow test set is very small (approx. 5k instances), and often contains simple scenes with a single salient target such as the held object. Also, annotations on GazeFollow are not always reliable as mentioned in Section 4.2. Hence, analyzing results on GazeFollow alone is not sufficient.

[1]The predicted LAH scores can also be used for these tasks but were shown to have slightly lower performance [14].

Method	Dist.↓	$F1_{LAH}$ ↑	$F1_{LAEO}$ ↑
Tafasca [39]	**0.115**	-	-
Gupta [13]	0.142	-	-
MTGS [14]	0.122	0.588	0.376
MTGS + AVA + CP	0.129	0.586	0.371
MTGS + HICO	0.119	**0.601**	0.407
MTGS + SWIG	0.117	0.600	**0.426**

Table 3. Cross-dataset results for the models trained on GazeFollow and evaluated on the ChildPlay dataset. Best results are given in bold, second best results are underlined.

Method	AUC↑	Avg. Dist↓	Min. Dist↓	$F1_{LAH}$↑
Multi Fusion	0.932	0.119	0.062	0.633
Early Fusion	0.936	**0.118**	**0.061**	**0.643**

Table 4. Ablation on early vs multi-stage fusion of VLM context using the AVA+ChildPlay classes on the GazeFollow dataset. Best results are given in bold.

ChildPlay. To further investigate the properties of our models, we perform cross-dataset evaluation on ChildPlay. The ChildPlay test set has a large number of instances (approx. 20k), and contains challenging scenes with multiple salient targets (ex. toys, other children/adults), making it an interesting benchmark. We observe that incorporating the AVA+CP classes results in a drop in performance for the distance score. However, with the larger set of HICO and SWIG classes, there is an improvement in performance for distance, LAH and LAEO. In particular, incorporating the SWIG classes gives the most improvements, with gaze following results comparable to the state of the art [39] and contrasts with our observations on GazeFollow. This suggests that incorporating gaze contextual cues can result in more robust performance with better generalization.

Ablation: Early Fusion vs Multi-Stage Fusion. We perform an ablation with two different fusion mechanisms for incorporating VLM contextual information in MTGS. The first is early fusion, and follows the approach described in Section 4.1. The second is a multi-stage fusion approach, where the VLM context is fused with the person tokens at every block of the architecture (4 times). We observe that the early fusion approach slightly outperforms the multi-stage fusion approach, especially for LAH, so we followed the early fusion approach for all our experiments.

Qualitative results. We provide qualitative results for MTGS, with and without the use of contextual cues in Figure 7. We observe that incorporating contextual cues can improve performance, helping identify the gaze target in challenging situations with multiple salient people and objects. For instance, in row 1, person 2 has a high score for *carrying*, which might indicate that this person is looking towards their hand. In row 2, person 3 has a high score for

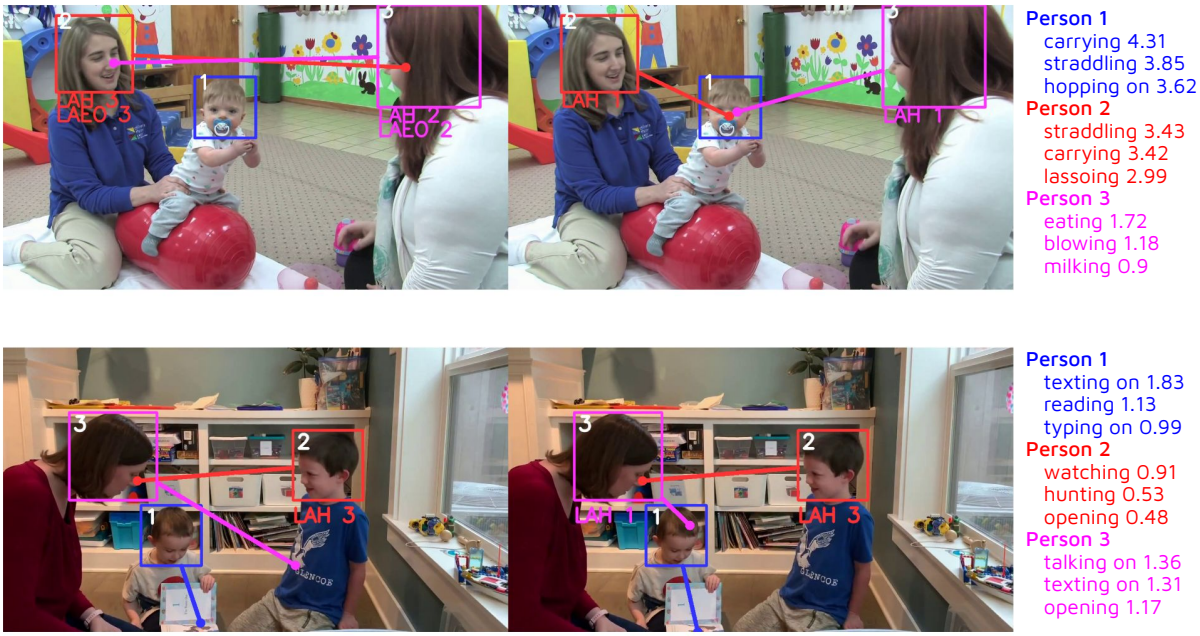

Figure 7. Qualitative results of MTGS [14] trained on GazeFollow and evaluated on ChildPlay. For each person, we display the predicted gaze point as well as social gaze task along with the associated person id. We provide results without contextual cues (left) and with contextual cues from the HICO classes (right). We also display the top three classes with the highest normalized score for each person.

talking on, which suggests social interaction such as LAH.

5. Discussion

Our observations in Section 4.3 suggest that incorporating a larger set of contextual cues can improve generalization performance for gaze following. As the set of cues becomes larger, it can capture more specific situations (ex. unlocking, sewing in SWIG) which are usually associated with certain gaze targets. It is worth noting that increasing the number of classes has a negligible impact on computation time. As mentioned in Section 3.1, the ITM approach processes the text prompts and images independently to obtain text and image embeddings. The final score is then a dot product of the two. Hence, all the text embeddings can be computed and saved at the start, and then used with any new image.

We also note that the set of HICO and SWIG classes utilized in our study are obtained from HOI datasets, hence, scores for the different cues could alternatively be obtained from HOI models. This is another interesting direction of investigation, but its main drawback is that the set of cues that can be considered is fixed depending on the chosen model. On the other hand, leveraging VLMs in a zero-shot manner allows us to consider any set of cues, including larger sets than the ones we considered (with a negligible impact on computation time), or more domain specific cues tailored for specific applications.

6. Conclusion

In this work, we explored the zero-shot capabilities of VLMs for extracting contextual cues related to a person's pose or interactions with objects and other people, and evaluated the impact of incorporating these cues into a gaze following model. We learned that VLMs can indeed extract contextual cues, and that considering the entire image with a red-circle drawn over the person of interest serves as the best visual prompt, and that ensembling scores from different textual prompts serves as the best text prompting strategy. We also observed that BLIP-2 is the overall best performing VLM, and that ICL can potentially bring further benefits. In the second part, we observed that incorporating the extracted cues into a gaze following model can provide better generalization performance, especially when considering a larger set of classes. In future work, we plan to investigate other VLMs and further explore prompting strategies such as ICL. We also plan to explore the option of predicting the different cues rather than providing them as input to the model.

Acknowledgement. This research was supported by the AI4Autism project (digital phenotyping of autism spectrum disorders in children, grant agreement number CRSII5_202235 / 1) of the Sinergia interdisciplinary program of the SNSF. It was also supported by Innosuisse, the Swiss innovation agency, through the NL-CH Eureka Innovation project ePartner4ALL (a personalized and blended care solution with virtual buddy for child health, number 57272.1 IP-ICT).

References

[1] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[3] Jacob Andreas. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, 2022.

[4] Jun Bao, Buyu Liu, and Jun Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14126–14135, 2022.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*, pages 1017–1025, 2015.

[7] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406, 2020.

[8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11390–11399, 2021.

[11] Daniel Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and vision computing*, 27(12):1775–1787, 2009.

[12] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5041–5050, 2022.

[14] Anshul Gupta, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard, and Jean-Marc Odobez. A novel framework for multi-person temporal gaze following and social gaze prediction. In *Arxiv*, 2024.

[15] Zhengxi Hu, Dingye Yang, Shilei Cheng, Lei Zhou, Shichao Wu, and Jingtai Liu. We know where they are looking at from the rgb-d camera: Gaze following in 3d. *IEEE Transactions on Instrumentation and Measurement*, 2022.

[16] Zhengxi Hu, Kunxu Zhao, Bohan Zhou, Hang Guo, Shichao Wu, Yuxue Yang, and Jingtai Liu. Gaze target estimation inspired by interactive attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8524–8536, 2022.

[17] Tianlei Jin, Zheyuan Lin, Shiqiang Zhu, Wen Wang, and Shunda Hu. Multi-person gaze-following with numerical coordinate regression. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021.

[18] Tianlei Jin, Qizhi Yu, Shiqiang Zhu, Zheyuan Lin, Jie Ren, Yuanhai Zhou, and Wei Song. Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence*, 113:104924, 2022.

[19] Roland Johansson, Goran Westling, Anders Backstrom, and Randall Flanagan. Eye-Hand Coordination in Object Manipulation. *Journal of Neuroscience*, 21(17):6917–6932, 2001.

[20] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022.

[21] Jing Li, Zejin Chen, Yihao Zhong, Hak-Keung Lam, Junxia Han, Gaoxiang Ouyang, Xiaoli Li, and Honghai Liu. Appearance-based gaze estimation for asd diagnosis. *IEEE Transactions on Cybernetics*, 52(7):6504–6517, 2022.

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[24] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018.

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[26] OpenAI. ChatGPT (February 25 version). https://chat.openai.com/chat, 2024.

[27] Kazuhiro Otsuka. Conversation scene analysis [social sciences]. *IEEE Signal Processing Magazine*, 28(4):127–131, 2011.

[28] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Inter. Conf. on Multimodal Interfaces*, pages 191–198, 2005.

[29] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[32] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1435–1443, 2017.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[34] Samira Sheikhi and Jean-Marc Odobez. Combining dynamic head pose–gaze mapping with the robot conversational state for attention recognition in human–robot interactions. *Pattern Recognition Letters*, 66:81–90, 2015.

[35] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11987–11997, 2023.

[36] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. 2023.

[37] Rainer Stiefelhagen, Michael Finke, Jie Yang, and Alex Waibel. From gaze to focus of attention. In *International Conference on Advances in Visual Information Systems*, pages 765–772. Springer, 1999.

[38] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Trans. Neural Networks*, 13(4):928–938, 2002.

[39] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20935–20946, 2023.

[40] Francesco Tonini, Cigdem Beyan, and Elisa Ricci. Multimodal across domains gaze target detection. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 420–431, 2022.

[41] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[42] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[43] Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

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

[45] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*, 2023.

[46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

Exploring the Zero-Shot Capabilities of Vision-Language Models for Improving Gaze Following

Supplementary Material

7. Appendix

7.1. Example of visual prompts

As described in Section 3.1, we investigate different visual prompting approaches to focus on a specific individual in the scene. An example of each prompt is provided in Fig. 10. These techniques are implemented on either the whole image or specifically on the cropped image of the target person. In total, this leads to eight distinct visual prompting strategies.

7.2. Details of the Childplay dataset

In Table 5, we detail the number of annotated negative and positive samples for each class in the ChildPlay dataset.

Classes	negative	positive
looking at hand	36	35
reaching	36	34
sitting	60	52
child	59	58
manipulation	59	59
speaking	31	30

Table 5. Classes and statistics of the ChildPlay dataset annotation.

7.3. Details of Text Prompts

ITM. Fig. 8, lists different text prompt variations as described in Section 3.3 for the ITM approach. A final prompt is a combination of {template},{person}, {photo} and {synonym} such as *"this individual is grabbing"* or *"a snapshot of a human handling"*.

VQA. For the VQA approach, for computational reasons, we consider a single template in the form of a question, and reduce the number of synonyms for the classes. We provide the template and synonyms in Fig. 9.

```
"template":   [ "this [person] is [class_synonym].",
                "a [person] is [class_synonym].",
                "a [person] [class_synonym].",
                "[class_synonym].",
                "a [name_photo] of a [person] [class_synonym]."]

"person":     [ "person", "individual", "human"]

"photo": [ "photo", "picture", "image", "snapshot", "shot", "pic"]

"synonym":
    "looking_hand": ["looking at hand" ,"examining hand", ...]
    "reaching":   ["reaching", "grabbing", "catching", "picking up", ...]
    "sitting":    ["sitting","seated", "resting", ...]
    "child":      ["a kid","a child", "a youth", ...]
    "manipulation": ["handling", "manipulating", "touching", ...]
    "speaking":   ["speaking","talking", "narrating", ...]
```

Figure 8. List of the different prompts variations used as described in section 3.3. A final prompt is a combination of {template},{person}, {photo} and {synonym} such as *"this individual is grabbing"* or *"a snapshot of a human handling"*.

```
"template":    [ "Is this [person] [class]? Answer yes or no."],

"person":      [ "person", "individual", "human"]

"synonym":
    "reaching":    ["reaching", "grabbing", "catching", "picking up"]
    "sitting":     ["sitting","seated", "resting"]
    "child":       ["a kid","a child", "a youth"]
    "manipulation":["handling", "manipulating", "touching"]
    "speaking":    ["speaking","talking", "narrating"]
```

Figure 9. List of the different prompt variations used for VQA model. A final prompt is a combination of {template},{person}, and {synonym} such as *"Is this individual grabbing? Answer yes or no."* .

7.4. Impact of class synonyms

In Fig. 11, we provide the results for varying the class synonym in the text prompt. We observe that performance can change depending on the used synonym by a large margin.

Figure 10. Different visual prompts are used to focus on the person of interest. Row-wise, the image-based and person cropped-based variants are displayed. Column-wise, various visual prompts such as ellipse, blur, and gray are presented.

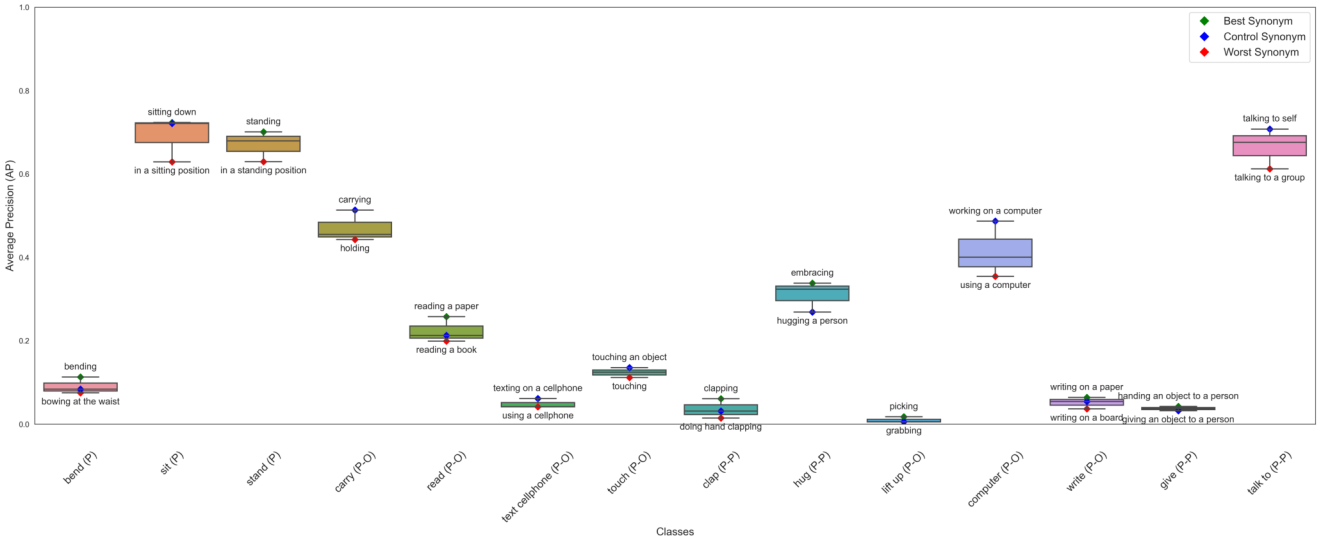

Figure 11. Performance when varying the class synonym in the text prompt. We display the mean and variance of results, as well as the best and worst synonym.