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.

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.

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.

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.

```
"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."* .

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.

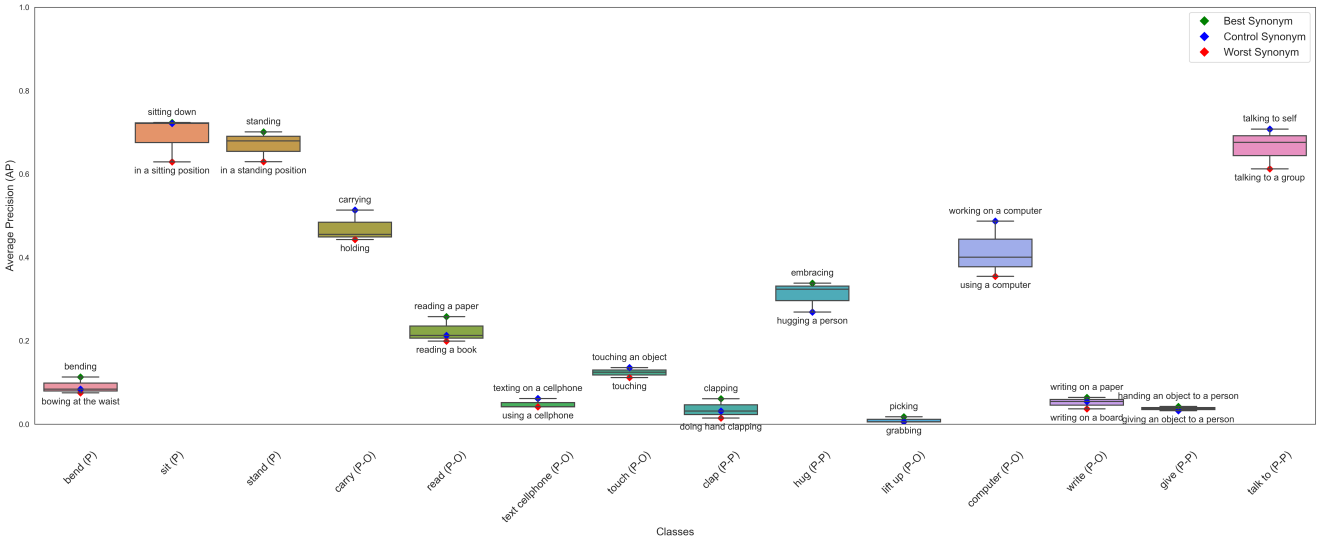

Figure 11. Performance when varying the class synonym in the text prompt. We display the mean and variance of results, as well as the best and worst synonym.