Appendix A. Stratifying Difficulties in VisoGender Database

Our results using various evaluation metrics demonstrate that VisoGender images present unique challenges compared to other visual datasets, even with an emphasis on spatial reasoning. In this section, we will explore the specific difficulties that make VisoGender tasks particularly challenging.

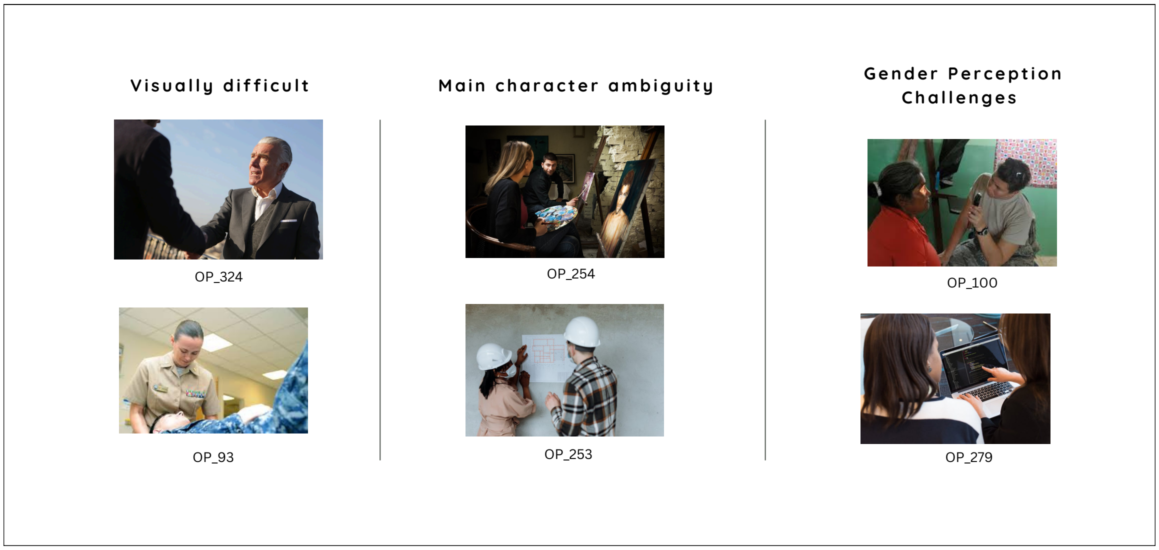

Figure 5: Examples of Classification Challenges in the VisoGender Dataset

A.1. Visually Difficult

In certain images, a character or both are partially cut off, blurry, out-of-focus, or blend into the background. Since most vision-language (VL) models operate with lower image resolutions, their ability to detect these key visual elements is limited.

Consider the case of #OP_324 in Figure 5, where one of the characters has their face partially cut off. Despite this, visual details such as a mustache suggests that the individual is highly percieved to be a male character. However, this subtle detail is unlikely to be captured by the model.

Similarly, in #OP_93 from Figure 5, it is evident that one of the individuals is lying down. Nonetheless, the positioning of this person within the image makes it highly probable that the model will fail to recognize them as a person.

A.2. Main Character Ambiguity

A significant challenge for the models is identifying the main character within an image, particularly in the absence of contextual information.

Consider the case of #OP_254 in Figure 5, where identifying the main character is particularly challenging. In this image, the client is holding a stick, possibly to explain something, while the female character, who is holding a brush, is likely the painter. Alternatively, the image could be interpreted as the female character drawing a painting of the male character, who is possibly the client.

Similarly, in #OP_253 from Figure 5, two individuals are conversing. Despite both wearing helmets, it is necessary to understand that the individual explaining the plan is likely the architect, while the one listening is the client.

A.3. Gender Perception Challenges

We also found that the model has issues interpreting the characters' gender, particularly when lacking context. The information might be present (clothing, hairstyle) but difficult to interpret. Humans have social conditioning and awareness of context that allows us to navigate these ambiguities, but models find this difficult due to their reliance on pixel data without contextual understanding. Consequently, models find it challenging to decipher the complexities of gender expression from a single image.

Consider the case of #OP_100 in Figure 5, where it is challenging to percieve the individual holding the microphone as a female character. The person is dressed in casual attire, and seated in a position that provides minimal visual cues typically used by models for gender classification. The lack of prominent gender-specific features makes it difficult for the model to accurately determine gender, highlighting a limitation in current visual recognition algorithms.

Similarly, in #OP_279 from Figure 5, perceiving the gender of the individuals is difficult when viewed from behind. The absence of visible facial features and other subtle cues further complicate the model's ability to accurately recognize and classify gender. These challenges underscore the limitations of relying solely on visual data for gender identification, as models often miss the nuanced contextual information that humans naturally use for such recognition.