Local2Global query Alignment: A few Qualitative Examples

To provide a comprehensive understanding of our approach, we present a selection of qualitative examples in this section. Our analysis begins with Figure 2, which shows scenarios where our model excels, alongside instances where it still faces challenges. Complementing this, Figure 3 offers a critical comparative evaluation. This figure pits our model against both our baseline (without alignment) and the well-regarded GenVIS method [18], providing compelling visual evidence of our improvements.

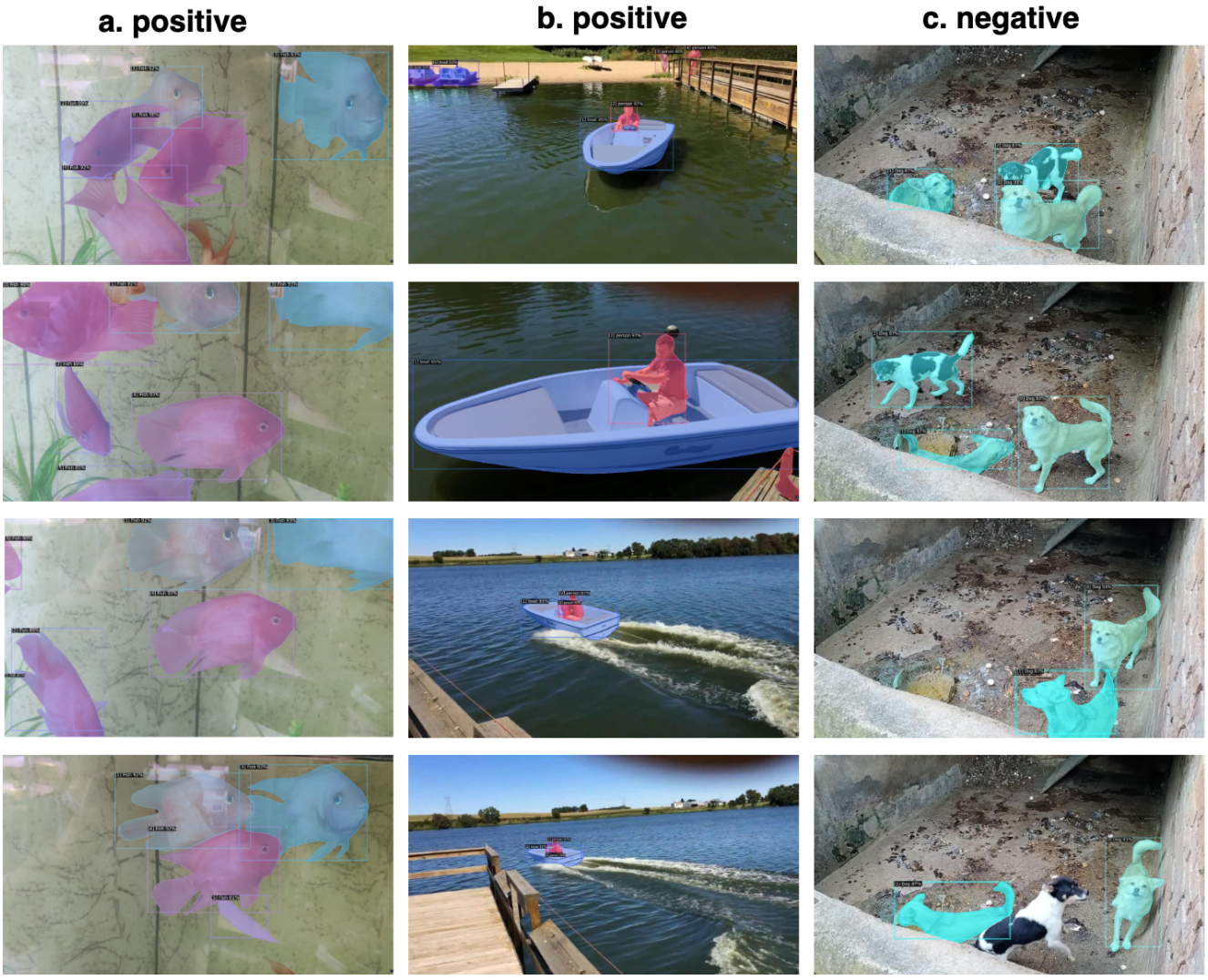

Figure 2. Qualitative examples of our model. In (a) our Local2Global framework successfully segments and tracks a large number of fish swimming in an aquarium. While in (b) displays successful segmentation and tracking of a person driving a speedboat as it moves away from the camera. (c) illustrates a negative example, showing intial tracking of three dogs, where one disappears and then re-appears later in the video. In this case, Local2Global fails to re-identify and track the reappearing dog, highlighting the need for stronger long-term temporal modeling.

Figure 3. Qualitative evaluation. Columns from left to right are **Ours without L2G-aligner**(baseline), **GenVIS**, and **Ours** (with L2G-aligner). Three examples are shown, and rows, top down show the time shift. In **a)**, a human hand appears in the bottom row. Only our method tracks the parrot being grabbed, without disrupting other tracked IDs. In both **b)** and **c)**, we purposefully select cases where mutual occlusion happens (top and bottom row, the occlusion frames are not shown due to space). As can be seen, our model tracks the right IDs for airplanes in b) and motorbike racers in c), while other methods swap the tracking IDs after the mutual occlusion. Notably, without L2G-aligner, the predictions across frames are not stable. More results can be found in supplementary material.