Thinking in Frames: How Visual Context and Test-Time Scaling Empower Video Reasoning

Chengzu Li*[1 2] Zanyi Wang*[3] Jiaang Li*[2] Yi Xu[1] Han Zhou[1] Huanyu Zhang[4]
Ruichuan An[5] Dengyang Jiang[6] Zhaochong An[2] Ivan Vulić[1] Serge Belongie[2] Anna Korhonen[1]

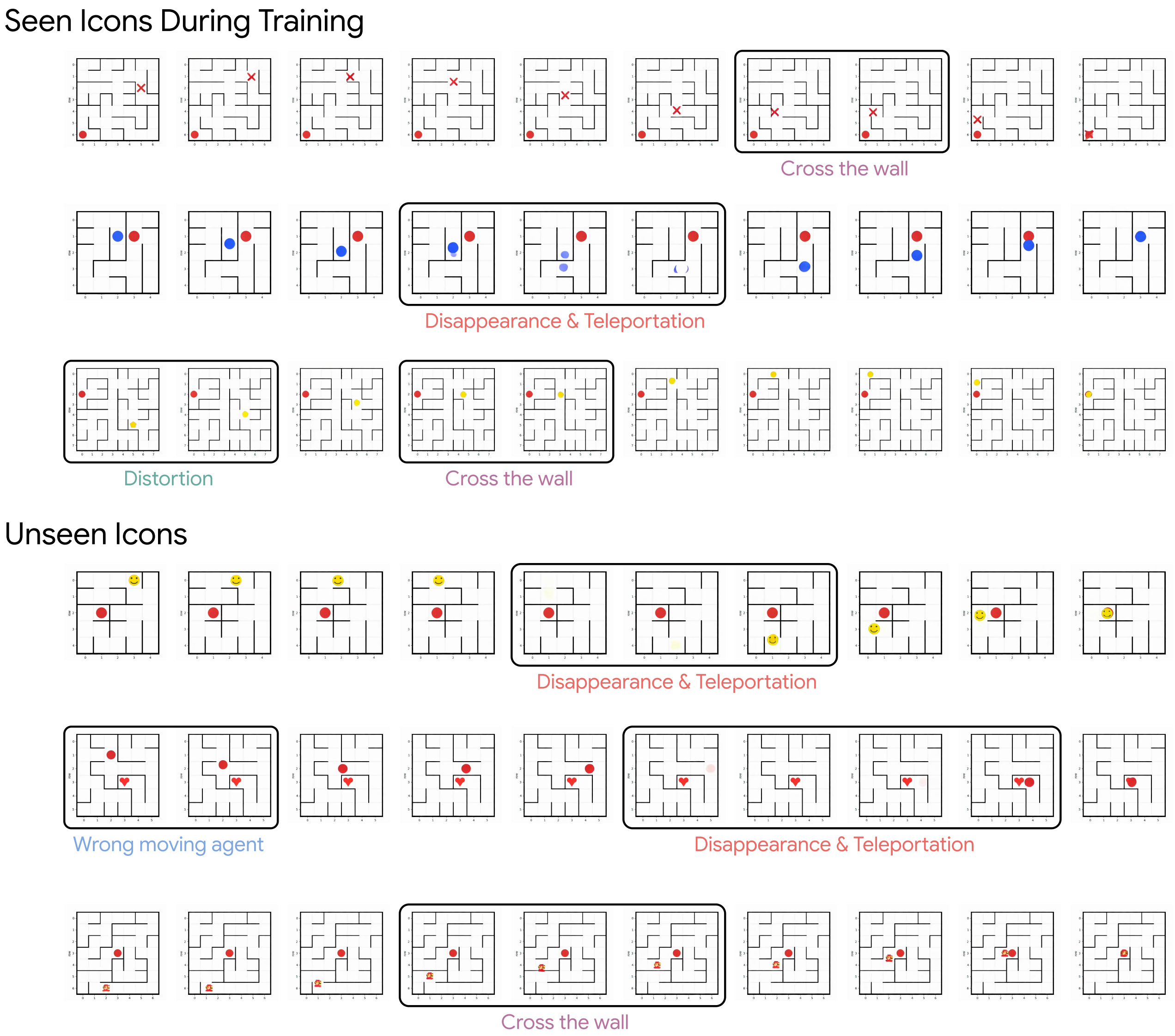
Seen Icons During Training
Cross the wall
Disappearance & Teleportation
Distortion
Cross the wall
Unseen Icons
Disappearance & Teleportation
Wrong moving agent
Disappearance & Teleportation
Cross the wall

Figure 1. Wrong examples of MAZENAVIGATION generated by fine-tuned Wan 2.2 TI2V 5B.

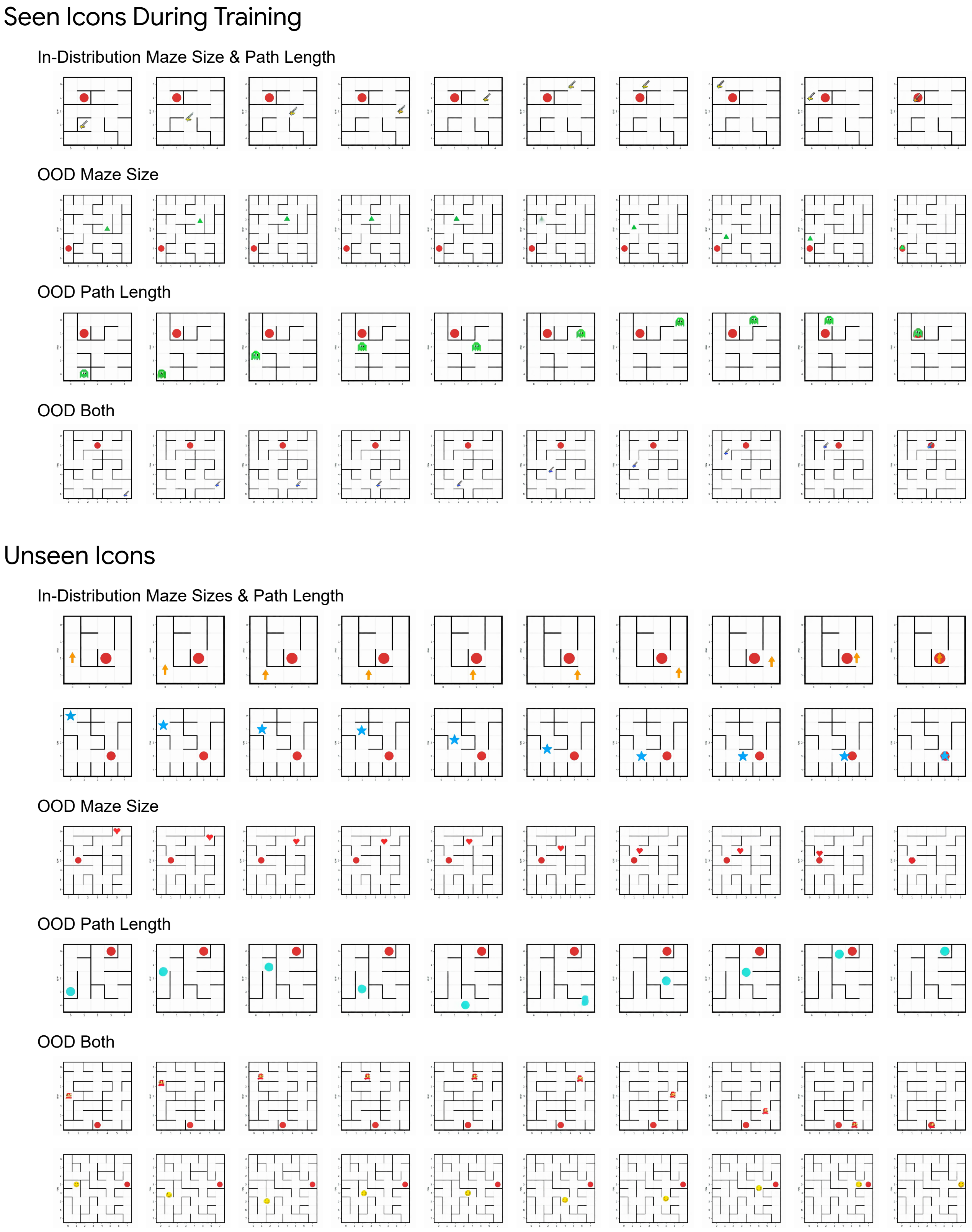
Seen Icons During Training
In-Distribution Maze Size & Path Length
OOD Maze Size
OOD Path Length
OOD Both
Unseen Icons
In-Distribution Maze Sizes & Path Length
OOD Maze Size
OOD Path Length
OOD Both

Figure 2. Correct examples of MAZENAVIGATION generated by fine-tuned Wan 2.2 TI2V 5B.

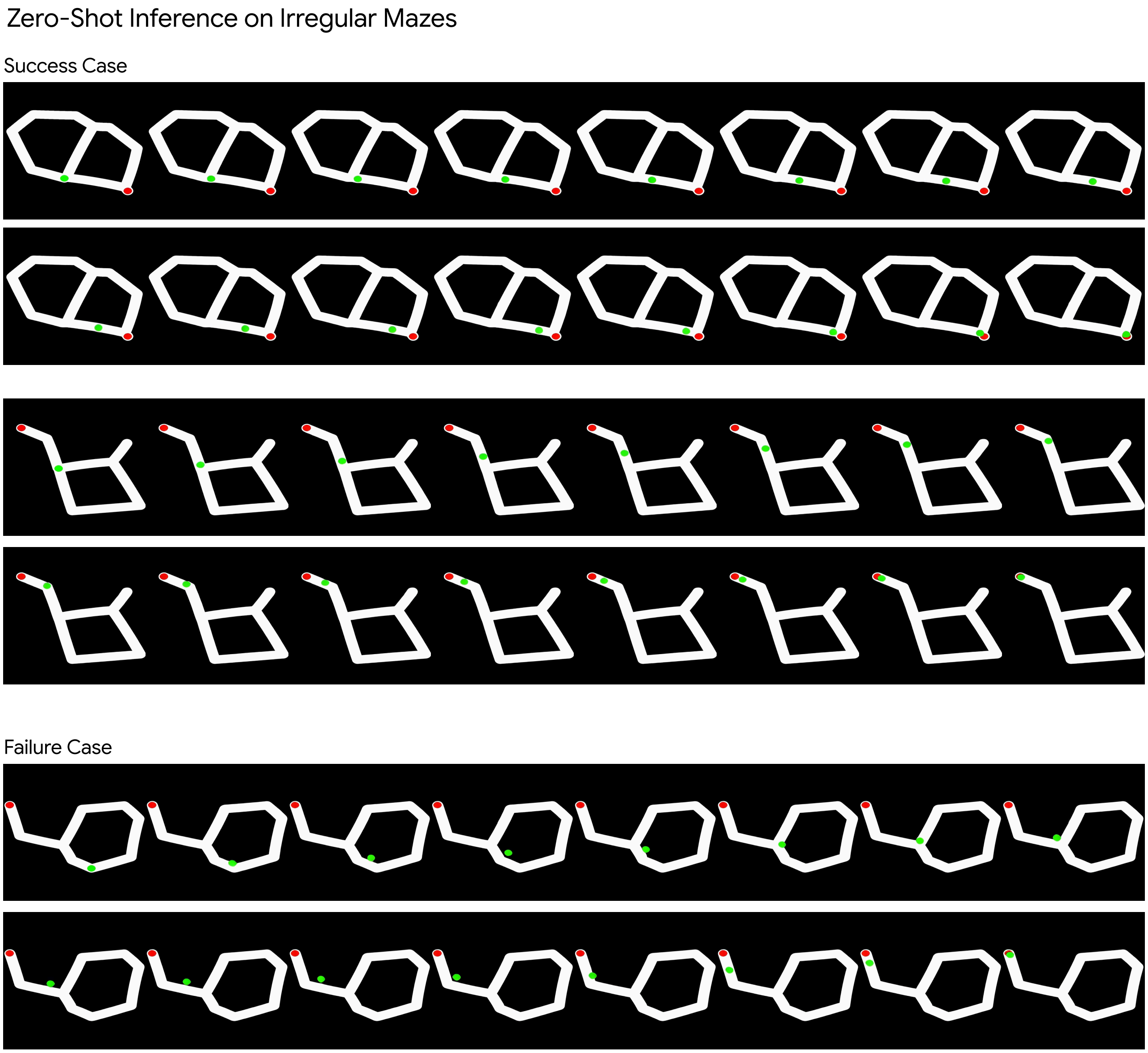

Figure 3. Visualizations of zero-shot inference on irregular maze with Wan 2.2 TI2V 5B trained on regular mazes. We observe that although not included in the training data, trained video generation model can adapt a certain level of planning capabilities to such irregular mazes with different background and meanwhile keeping the constraint (no change in the background, follow the path, do not cross the wall), and surprisingly can move diagonally.

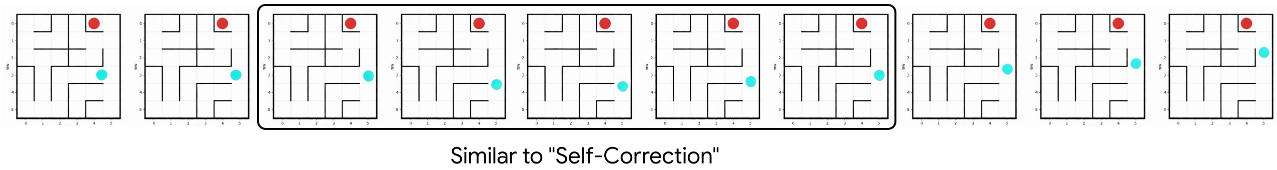

Figure 4. Example of trajectory similar to "self-correction" in MAZENAVIGATION when provided with more inference frame budget.

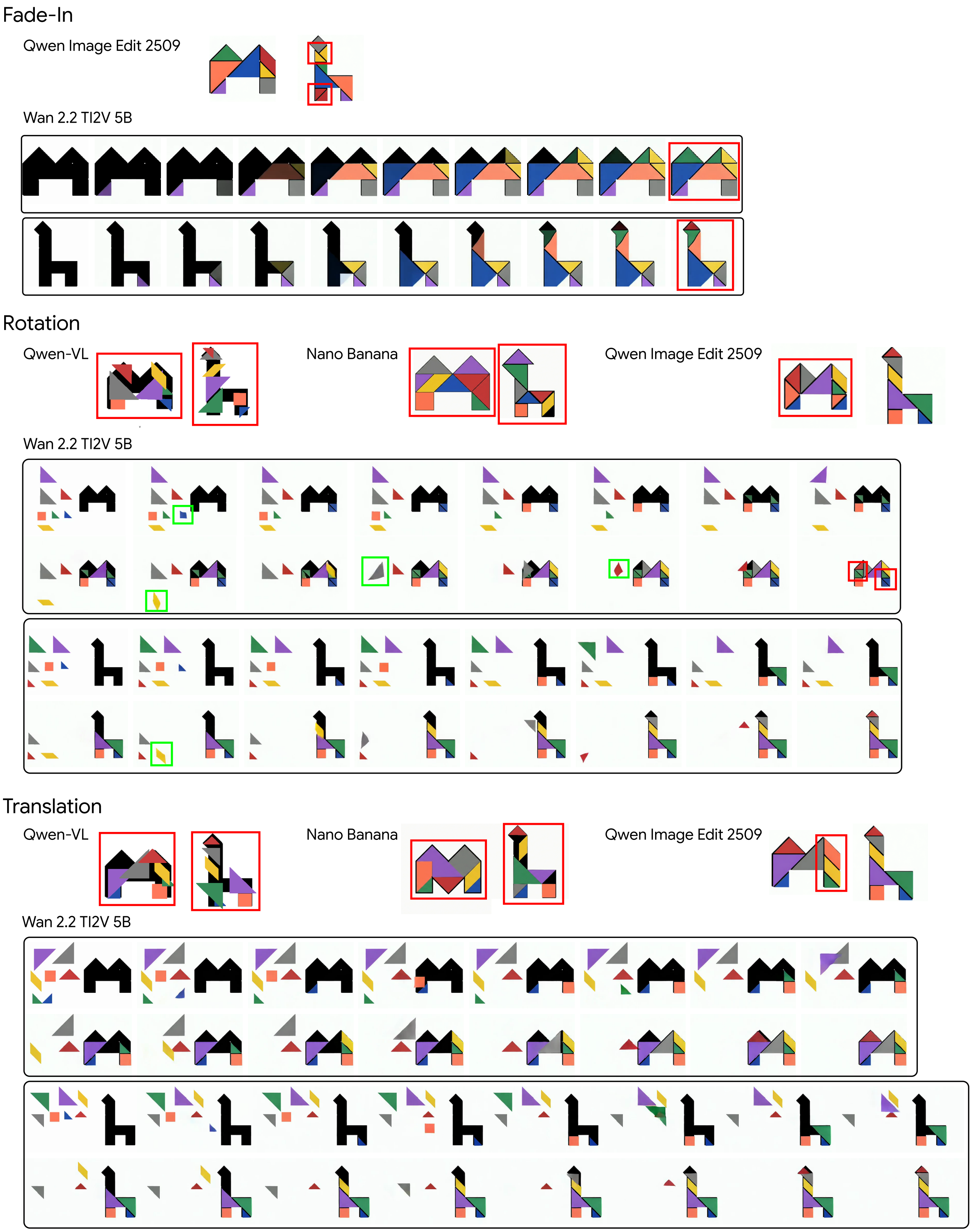

Figure 5. Examples of TANGRAMPUZZLE generated by different systems. Red bounding boxes indicates wrong predictions, and green bounding box indicates distortion throughout the process.