Step-by-Step Grounding and Box Refinement for Perception Test Grounded VideoQA

Gabriel Turbinca, Yuheng Zhou, Jialei Xu, Zizhuang Wei
Team **Huawei SpaceMind**

Abstract

We present Dialogue-Consensus-SAM, our Track 2 system for the 2026 Perception Test Challenge. GPT-5.6 Sol first identifies the objects named by the question and produces complete tracks. We then compare a second independent track, ask the model again on close-up crops, and use SAM 2.1 to improve box edges. These later steps may change box coordinates, but they never change the chosen object IDs, object count, frame list, or predicted lifetime. No model was fine-tuned and no test item was manually labelled. The final submission scored **0.6541 HOTA** on the official Test set.

1 Task and models

For each video and question, Grounded VideoQA requires a box track for every object that answers the question [1]. HOTA measures whether the right objects are found, whether their IDs stay consistent, and how well the boxes fit [2]. The Test set has 932 videos and 1,859 questions.

We use GPT-5.6 Sol through an API and the public `facebook/sam2.1-hiera-tiny` checkpoint [3]. Inference is zero-shot: no Perception Test split is used to update model weights, and validation or test examples are never placed in a prompt. Validation labels are used only to choose fixed box thresholds and size corrections. Each video/question pair is solved independently. The providers' public model documentation describes their released training data; the GPT training mixture is not disclosed.

2 Method

The pipeline has five concrete steps. The first step chooses the answer objects. The remaining steps only check or refine their boxes.

1. Build the main tracks. We send the question and uniformly sampled video frames to GPT-5.6 Sol. Sampling is at least 2 FPS and always includes the first and last physical frame. Temperature is zero. The model must return JSON with: the number of answer objects; a short description and fixed ID for each object; and, for every sampled time, a state (visible, covered, or outside the frame) and a normalized $[x_0, y_0, x_1, y_1]$ box. We map the result to source-frame IDs $0, 30, 60, \ldots$ plus the decoded last frame and keep a dense box list over each predicted active period. This exact frame grid prevents timestamp rounding from shifting later boxes.

2. Check the result with a second track. We run a second, independent GPT track on the same frame grid. If both runs predict the same object count, we match tracks by their mean box overlap on common active frames. On one frame, we average the two boxes only when their intersection over union (IoU) is at least 0.70 and their area ratio is between 0.7 and 1.4. Otherwise we keep the first box. Object IDs, lifetimes, frame IDs, and object count always come from Step 1. We then enlarge each accepted box by 1.02 around its centre to better match the annotation box convention.

3. Re-check difficult boxes on close-up video crops. We split active tracks into 3,672 Test segments. Around each current box we render a square crop four times as wide as the box and sample that crop sequence at at least 2 FPS. GPT-5.6 Sol is run three times with temperature 1 and the API's xhigh reasoning setting. The three boxes must agree: their minimum pairwise IoU must be at least 0.85. Their coordinate-wise median must also overlap the current box by at least 0.50 IoU. If both tests pass, we average the median box and current box equally. Otherwise nothing changes. This step refined 27,273 Test boxes.

4. Improve visible box edges with SAM 2.1. For each frame marked visible in Step 1, the current box is used as a SAM box prompt on the original frame. We turn the returned mask into a box, then expand it by 1.038 horizontally and 1.030 vertically because mask edges are usually tighter than human boxes. We replace the current box only when the two boxes have IoU at least 0.70 and area ratio in $[0.7, 1.4]$. SAM returned 30,440 non-empty masks and 29,743 passed this check.

5. Apply small, disjoint corrections. A size correction is applied to the 17,710 boxes untouched by Steps 3 and 4. The submitted version also checks far crop candidates against an older independent track. "Far" means candidate/current IoU below 0.50. The candidate must overlap the older track by at least 0.70

IoU on at least two visible frames of the same segment; the Step 4 SAM check is then repeated. This last rule changed 131 of 56,992 boxes across 16 questions. It never changed IDs, lifetimes, or frame IDs.

3 Results

Table 1 shows how the fixed rules combine on a 100-question validation subset. The crop and SAM steps give the largest measured gains. The last far-candidate rule was explored after the main pipeline and its confidence interval includes zero, so we do not claim a separate Test gain for it.

Table 1: Fixed100 validation HOTA.

System	HOTA	Change
Main + second-track check	0.66262	—
+ three-sample crop check	0.66989	+0.00727
+ calibrated SAM 2.1	0.68057	+0.01069
+ size correction	0.68090	+0.00032
+ far check and re-SAM	0.68187	+0.00097

Table 2: Final Test execution statistics.

Item	Count
Videos / questions	932 / 1,859
Output tracks / boxes	3,226 / 56,992
Crop segments / refined boxes	3,672 / 27,273
SAM masks returned / accepted	30,440 / 29,743
Size-only corrections	17,710
Far-branch changes	131 (0.23%)

Earlier official versions scored 0.59684, 0.60770, and 0.61645 HOTA. The complete submitted system scored **0.6541**. Its output covers every Test question and contains no empty, invalid, or out-of-range track. The version without the final 131 replacements was not scored officially, so the official result cannot tell us whether that last rule helped or hurt by itself.

4 Conclusion

The most useful design choice was to separate two jobs. GPT first decides which objects answer the question and keeps their IDs stable. Independent GPT crops and SAM then improve coordinates only when simple box-agreement tests pass. This improves box quality without letting a local correction silently change the answer object. The main remaining errors are wrong object choice and ID drift over long periods; local box refinement cannot solve those errors.

References

[1] V. Pătrăucean et al. Perception Test: A Diagnostic Benchmark for Multimodal Video Models. *NeurIPS*, 2023.

[2] J. Luiten et al. HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking. *IJCV*, 2021.

[3] N. Ravi et al. SAM 2: Segment Anything in Images and Videos. *ICLR*, 2025.