Focus Ambiguity in Visual Questions: A Disambiguation Problem, Not Instance Segmentation - Supplementary Materials -

This document supplements the main paper with more in- formation about:

1. Synthetic Training Data for Focus Parsing
2. Qualitative Examples of the Proposed Metrics
3. IoP–IoU Mismatch: High Sufficiency, Low Overlap
4. Incomplete CTR Annotations in VQ-FocusAmbiguity

1. Synthetic Training Data for Focus Parsing

The Focus Parsing module is trained using synthetically generated question–focus description pairs following the compositional templates described in the main paper. Table 1 shows representative examples.

For each question, the Specified phrase preserves all relational and attribute modifiers to reflect the full focus description implied by the question, whereas the Base phrase removes modifiers and retains only the core noun phrase. This design enables precise grounding when possible, while providing a simpler fallback prompt when the fully specified expression fails to produce candidate masks.

2. Qualitative Examples of the Proposed Metrics

We propose two complementary metrics for evaluating disambiguation: AP_{IoP}, which measures sufficiency detection, and $\mathcal{E}_{Leak}$, which measures interpretive exclusivity.

Figure 1 presents representative predictions from our baseline model spanning low to high performance under both metrics. These examples illustrate how high AP_{IoP} corresponds to successful localization within valid CTRs, while low $\mathcal{E}_{Leak}$ reflects concentrated predictions that avoid spreading across competing interpretations.

3. IoP–IoU Mismatch: High Sufficiency, Low Overlap

To analyze the difference between IoU-based evaluation and our sufficiency-based criterion, we compare each prediction's maximum IoU and maximum IoP with any CTR inVQ-Disambiguation in the main paper and observe numerous cases where predictions achieve low IoU yet high IoP. Such predictions correspond to valid disambiguating cues that lie within a CTR but do not reconstruct its full spatial extent. Representative examples are shown in Figure 2, illustrating how IoU under-credits sufficient cues.

4. Incomplete CTR Annotations in VQ-FocusAmbiguity

We identify that 33% of ambiguous samples in VQ-FocusAmbiguity contain incomplete CTR annotations. In such cases, ground-truth masks do not exhaustively capture all valid disambiguating regions.

Figure 3 presents representative examples. Predictions that successfully resolve ambiguity receive artificially low scores due to missing annotation coverage. This misalignment motivates our refined dataset construction and evaluation protocol under the disambiguation formulation.

Question	**Specified** Phrase	**Base** Phrase
What is written on the paper under the mug?	the paper under the mug	the paper
What color is the white truck?	the white truck	the truck
Is the tennis racket being held?	the tennis racket	the tennis racket

Table 1. Representative synthetic question–focus description pairs used to train the Focus Parsing module. The *Specified* phrase preserves relational or attribute modifiers, while the *Base* phrase retains only the core noun phrase.

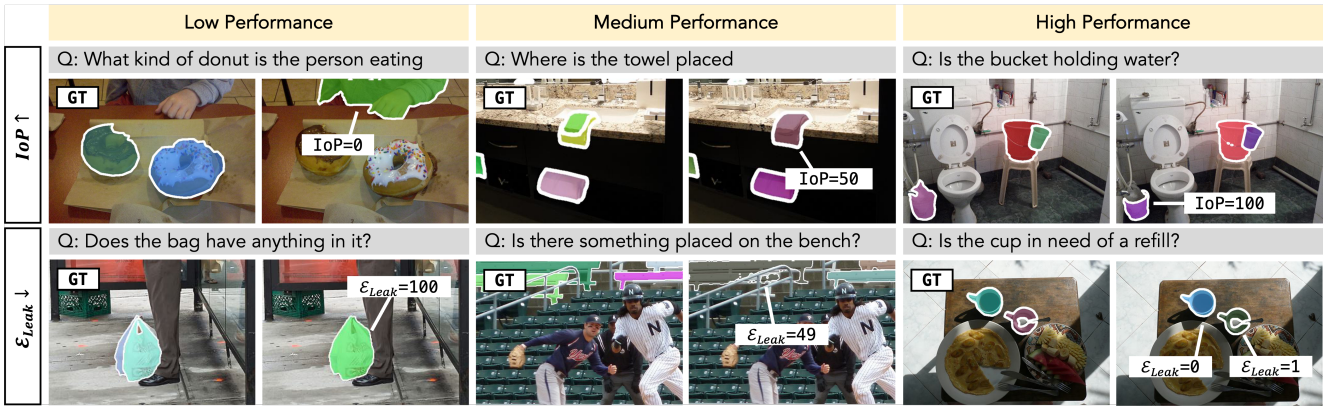

Figure 1. Representative predictions from our baseline model illustrating behavior under AP_{IoP} and $\mathcal{E}_{Leak}$. Examples span low to high performance for both sufficiency detection and leakage control.

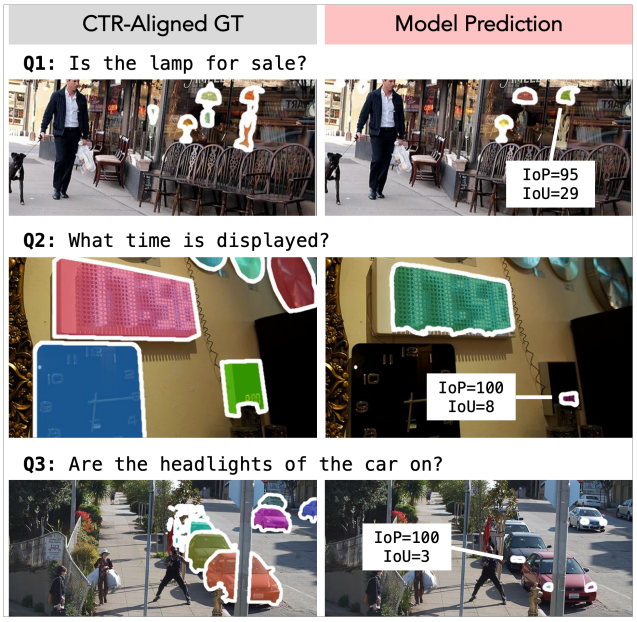

Figure 2. Examples of the baseline model predictions that achieve high IoP scores while rated low in IoU.

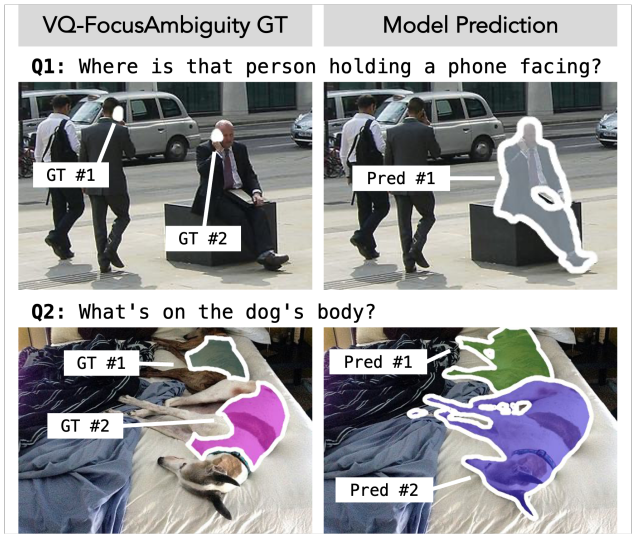

Figure 3. Examples of samples in VQ-FocusAmbiguity without complete CTRs, leading to unfaithful evaluation for disambiguation.