CORA: Consistency-Guided Semi-Supervised Framework for Reasoning Segmentation
Supplementary Material

Prantik Howlader[†] Hoang Nguyen-Canh[†] Srijan Das[‡] Jingyi Xu[†] Hieu Le[‡] Dimitris Samaras[†]
[†]Stony Brook University [‡]UNC-Charlotte
{phowlader,hcnguyen,jingyixu,samaras}@cs.stonybrook.edu {sdas24,hle40}@charlotte.edu

Summary: We provide additional analyses and results of our method, including:
- Impact of conditional visual instructions on limited labeled data
- Impact of Per-Pixel Confidence Thresholding and Consistency-Driven Weighting on Supervision from Unlabeled Images.
- Analysis of σ of the overall loss.
- Obtaining objects from semantic segmentation masks.
- Training LISA on conditional visual prompt instruction set.
- Qualitative Results.
- Conditional visual instruction dataset on Cityscape.
- Details on the Instruction Set for the PanNuke Dataset.

Unless stated otherwise, all experiments are conducted on the $\frac{1}{30}$ split of the Cityscapes dataset. We employ UniMatch [7] as the baseline semi-supervised method to generate pseudo-labels for unlabeled images.

1. Impact of conditional visual instructions on limited labeled data.

	(only labeled images)		
Conditional Visual instructions	1/30	1/16	1/8
	47.8	49.3	52.7
✓	**50.1**	**51.1**	**54.2**

Table 1. Impact of conditional visual instructions in training on only limited labeled images.

We analyze the effect of conditional visual instructions when training on a limited number of labeled images in Table 1. As a baseline, we follow LISA [4], generating referring and attribute-based reasoning segmentation instructions from labeled images to train CORA. Incorporating conditional visual instructions consistently improves performance across all data splits, with the most significant gain of $+2.3\%$ observed in the $\frac{1}{30}$ labeled data split. This demonstrates that conditional visual instructions are particularly beneficial in low-data regimes, as they not only increase the number of training instructions but also enhance CORA's ability to learn spatial context within images.

2. Impact of Per-Pixel Confidence Thresholding and Consistency-Driven Weighting on Supervision from Unlabeled Images.

We analyze the effect of per-pixel semi-supervised segmentation confidence (σ^{sseg}) and output consistency–based weighting (λ_i) in Table 2. In the first row, we apply a confidence threshold of $\sigma^{\text{sseg}} \geq 0.9$, following prior semi-supervised methods [1–3, 5, 7] that rely on high-confidence predictions to refine pseudo-labels. We observe that our output consistency–based weighting (λ_i) achieves the highest improvement across all data partitions. This improvement stems from the fact that while

Per-pixel adaptive weight	1/30	1/16	1/8
σ^{sseg}	54.4	56.3	58.8
λ_i	**56.9**	**58.2**	**60.5**

Table 2. Comparison of pseudo-label thresholding strategy based on semi-supervised segmentation confidence ($\sigma^{\text{sseg}} \geq 0.9$) against output consistency–based weighting (λ_i).

confidence thresholding is limited to the visual modality, output consistency incorporates both the image and its corresponding textual query, allowing for a more accurate estimation of the VLM confidence.

3. Analysis of σ of the overall loss

We analyze how CORA performs with different values of σ, which is used to control the contribution of unsupervised segmentation loss in the overall loss (Equation 9 of main paper). The results are provided in Table 3. We observe $\sigma = 0.001$ achieves the best performance and is used in all our experiments.

α	1/16	1/8
0.1	55.5	56.7
0.01	56.1	57.9
0.001	**56.9**	**58.2**
0.0001	55.3	57.6

Table 3. Analysis of (σ) of overall loss (1/30 and 1/16 partition protocols of Cityscapes Dataset)

4. Obtaining objects from semantic segmentation masks

For both Cityscapes and PanNuke datasets, we use semantic segmentation masks. To isolate individual objects within an image, we first decompose each category mask into separate connected components. These segmented objects are then utilized to construct the visual instruction prompt training set, ensuring that prompts are generated based on distinct object instances.

5. Training LISA on conditional visual prompt instruction set

We follow SegLLM [6] and replace the mask and bounding box encoding tokens of the reference instance with the word "mask" to simplify query formulation. For example, in the Cityscapes dataset, a query such as "Segment the car to the right of <mask> <box>." is reformulated as "Segment the car to the right of the mask." This transformation standardizes the input format, allowing the model to learn object relationships without relying on explicit bounding box tokens, improving generalization across different scene contexts.

6. Qualitative results

In Fig. 1, we compare the segmentation results of CORA with SegLLM [6]. Incorporating conditional visual instructions for labeled images and leveraging multi-modal LLM uncertainty to refine pseudo-labels for unlabeled images significantly improves segmentation performance on reasoning segmentation tasks in histopathology dataset (PanNuke).

7. Conditional Visual Instruction set for Cityscape dataset

In this section, we provide an in-depth analysis of the conditional visual instruction set on Cityscapes. We first present some statistics on the instruction set, followed by a detailed explanation of the instruction generation pipeline. Finally, we showcase sample images as demonstrations.

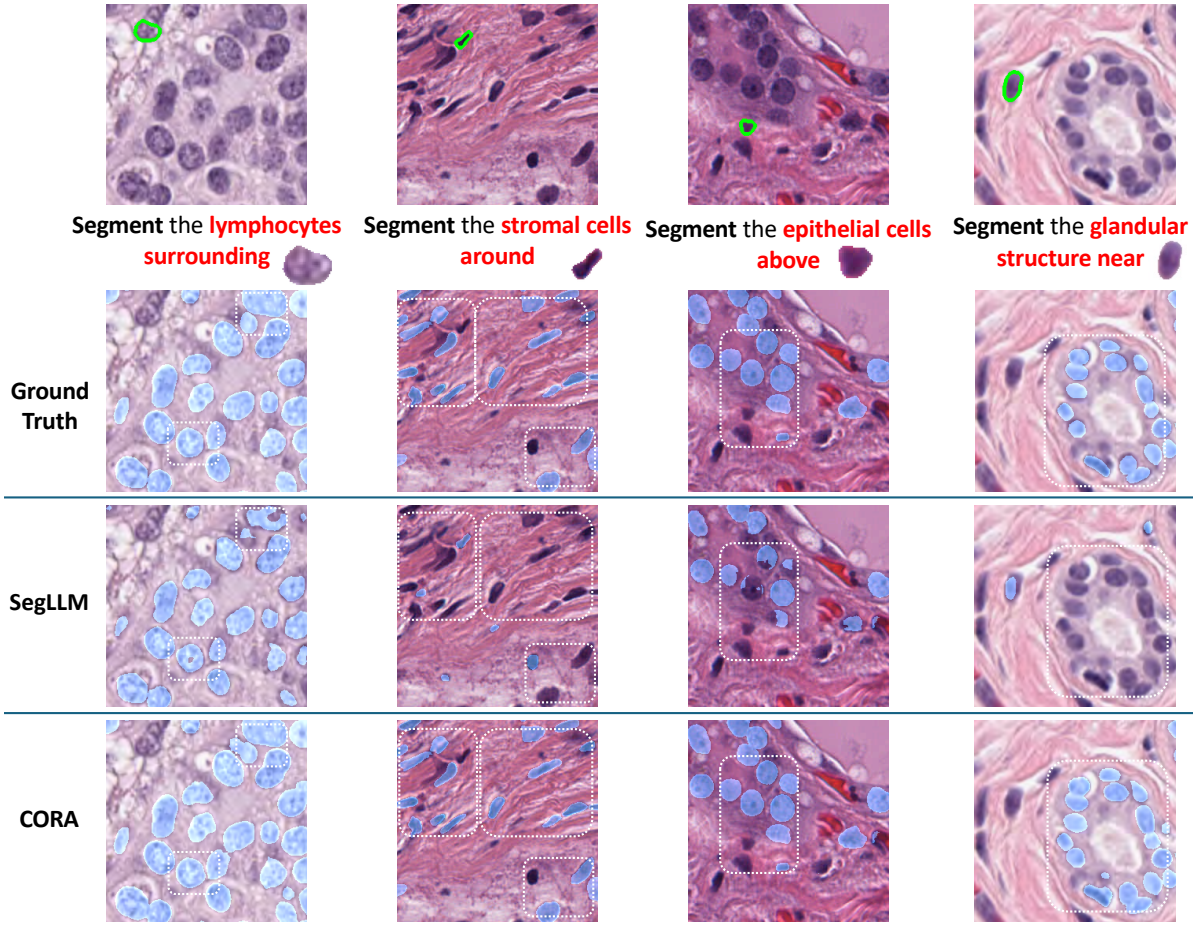


Figure 1. **Qualitative Results on PanNuke dataset:** Segmentations of SegLLM and CORA in PanNuke Dataset. White boxes show the areas where our model performs better segmentation than SegLLM.

7.1. Dataset statistic

Figure 2a illustrates the distribution of instructions for each class in the training subset of the generated conditional visual instruction dataset. Due to the inherent class imbalance present in the Cityscapes dataset, some classes, such as"pole," "traffic sign," and "car," occur frequently and thus have higher instruction counts. Conversely, less common classes, such as "train" and "bus," have significantly fewer instructions generated, reflecting their lower frequency in the dataset. Consequently, this distribution directly affects the variety and frequency of generated instructions, resulting in a natural emphasis on more prevalent object categories.

The average length of visual instructions remains relatively consistent across all classes, indicating a uniform level of complexity in the queries regardless of the target class as shown in Figure 2b

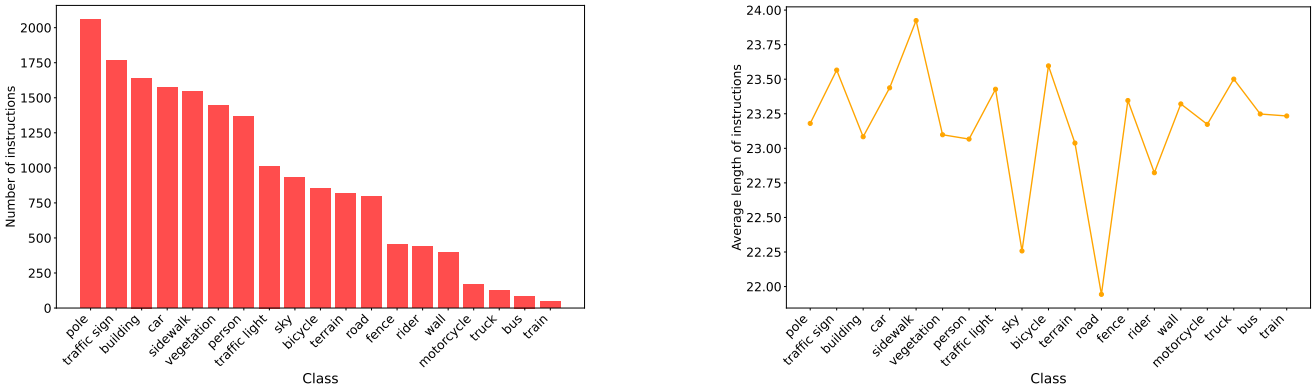
In the following section, we provide detailed explanations of the data generation pipeline initially outlined in Section 3.4 and Figure 4 of the main paper.

7.2. Detailed Prompt for Creating Conditional-relationship Visual Instruction Set

Our system prompts to generate conditional visual instruction is provided in Table 5

We begin by extracting semantic segmentation labels for each class, splitting these labels into distinct connected clusters, and assigning unique region IDs. Subsequently, we randomly sample anchor and target objects from the generated segmentation clusters. The anchor object serves as a reference point to define the position of the target object based on their relative spatial relationship

To give GPT-4o more information about spatial information of the objects, we provide GPT-4o with the target and anchor



(a) (b)

Figure 2. Statistics of the conditional visual instruction training set derived from the Cityscapes dataset: (a) Number of instructions referring to each class as the target label; (b) Average length of instructions for each class;

labels and their respective segmentation masks, represented as bounding polygons. Due to the complexity of the desired instructions and to avoid hallucination, we instruct GPT-4o to generate these descriptions incrementally, following these steps:

- First, GPT-4o generates visual descriptions of the anchor and target objects using their labels and mask contours.
- Finally, GPT-4o integrates the labels, mask contours, and visual descriptions of the anchor and target objects to construct the conditional visual instruction set, using the camera as the frame of reference to define spatial relations.

7.3. Detailed Prompt for Creating Attribute-based Instruction Set:

Table 4 provides the complete system prompt used to generate attribute-based visual instruction sets. We first assign unique regions ID for the segmentation cluster as in 7.2 and then sample random target objects. We then provide GPT-4o the image, target object metadata (label ID, label name, region ID, and polygons), and all other regions present in the image — and instructs GPT-4o to produce a three-turn question–answer conversation describing the target object.

The prompt enforces a structured format where:

- Each question asks for a visual description of the target object using its numeric region ID, without mentioning segmentation, polygons, or explicit labels.
- Each answer provides a concise but vivid description highlighting distinctive visual attributes and context that set the target apart from all other objects in the image.
- Descriptions avoid references to the viewer or photographer, maintain independence across all three Q&A pairs, and treat persons or riders (region IDs starting with 11_or 12_) as "individuals."

This detailed prompt ensures that the generated conversations are consistent, self-contained, and provide clear object-level descriptions suitable for training and evaluating visual understanding models.

7.4. Some example from the dataset

Figure 3 shows examples from the conditional visual instruction dataset used in this paper. Each instruction guides the model to identify the target object (red mask) based on semantic information from the provided label description and spatial relationships relative to an anchor object (green mask). For instance, in the top-left image, the task involves segmenting a rider, with the instruction clearly specifying both the semantic identity *"rider"* and the spatial context with respect to the anchor object (in this case is a *"person"*. This formulation encourages models to leverage both semantic and spatial relation for precise segmentation.

8. Details on the Instruction Set for the PanNuke Dataset

For the PanNuke dataset, we create two types of instruction sets:

- Semantic Segmentation Instruction Set: Similar to Cityscapes, we map the classes into a semantic segmentation template. For example, one template in our instruction set is: "Segment the epithelial cells / normal cells / glandular structures."

system_prompt = ”’ You are an AI visual assistant capable of analyzing a single image and its associated JSON file. The JSON file contains:
1. imgHeight and imgWidth: The resolution of the image.
2. target_label_id: Numeric identifier for the target object category (e.g., 1 for “person,” 2 for “car,” etc.).
3. target_label_name: Category name of the target object.
4. target_region_id: Unique ID for the connected region of the target object.
5. target_polygon: A list of polygons covering the target object region, where each polygon is a list of (x, y) points.
6. all_other_region_id_polygons: A dictionary containing all other region IDs (unique ID for each connected region) in the image as keys and the corresponding list of polygons covering that region, where each polygon is a list of (x, y) points.
Note: the target object is represent by target_label_id, target_label_name, target_region_id, target_polygon
Note: all_other_region_id_polygons represent all other connected regions in the image except the target object and their corresponding polygons to cover their respective regions

Your Goal
Produce a 3-question-and-answer conversation. Each question should ask for a visual description of how a target object (identified by a target_region_id) distinguishing it from all other regions in the image (represented by the keys in all_other_region_id_polygons and the corresponding value as the list of polygons covering that region), and each answer should provide that description.
Each Q&A pair must follow these rules:
1. Format
Question (Q[number]):
- Ask about the description of the target object (target_region_id).
- Use only the numeric target_region_id; do not write words like “region,” “label,” or mention polygons/coordinates.
Answer (A[number]):
- Give a concise but vivid description of of the target object distinguishing it from all other objects in the image represented by the region ids and the corresponding list of polygons in all_other_region_id_polygons.
- Do not mention viewer or photographer or synomyms of ’viewer/ photographer’ in the answer
- Do not mention target_region_id in the answer
2. No Segmentation or Polygon Details
- Do not mention segmentation or polygon coordinates. - Avoid discussing labels explicitly.
3. Persons or Riders
- If the object’s region_id begins with 11_ or 12_, refer to them as ”individual(s),” ”people,” or ”rider(s).”
4. Independence
- Each Q&A pair must stand alone and not refer to previous pairs.

Additional Rules:
- Each answer focuses on the unique visual traits or context that distinguish the target object (target_region_id) from all the other objects (in all_other_region_id_polygons) without disclosing their polygons or explicit labels.
- All questions are only about the target object (target_region_id).
- In each question dont compare the target object with any other specific other object, rather with all other objects (in all_other_region_id_polygons)
”’

Table 4. System prompt used to generate target Attribute-based Instruction Set using GPT-4o

- Conditional-Relationship Visual Instruction Set: Since the PanNuke dataset contains only 2D spatial relationships, we can manually extract spatial information without relying on GPT-4o. An example instruction from the dataset is: “Segment the cancer cells above the dead cell <dead cell mask>.”

References

[1] Alonso, I., Sabater, A., Ferstl, D., Montesano, L., Murillo, A.C.: Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8219–8228 (2021)

[2] Fan, J., Gao, B., Jin, H., Jiang, L.: Ucc: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9947–9956 (2022)

[3] Hu, H., Wei, F., Hu, H., Ye, Q., Cui, J., Wang, L.: Semi-supervised semantic segmentation via adaptive equalization learning. Advances in Neural Information Processing Systems **34**, 22106–22118 (2021)

[4] Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024)

[5] Liu, Y., Tian, Y., Chen, Y., Liu, F., Belagiannis, V., Carneiro, G.: Perturbed and strict mean teachers for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4258–4267 (2022)

[6] Wang, X., Zhang, S., Li, S., Kallidromitis, K., Li, K., Kato, Y., Kozuka, K., Darrell, T.: Segllm: Multi-round reasoning segmentation. arXiv preprint arXiv:2410.18923 (2024)

[7] Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7236–7246 (2023)

system_prompt = "' You are an AI visual assistant that can analyze one image and one JSON file to label conditional segmentation tasks.
The segmentation is semantic, but for each label, the mask is divided into separate regions based on connectivity. The continuous mask of multiple objects with the same label is considered a segmentation region.
The JSON file contains information for segmenting the image. It has three keys:
- **height** and **width**: the image resolution
- **objects**: the list of object clusters in the image

For each object cluster:
- **label_ids**: identifiers for the object
- **label_name**: the object type
- **region_ids**: identifiers for the object cluster
- **polygon**: list of polygons covering the object mask, each represented as a list of (x, y) coordinates. Note: each region_id corresponds to a list of polygons — consider all polygons for that region together.

Then, design a conversation between you and a person asking about segmenting the target object region in this image based on a given anchor object.

- First, select two objects with a clear spatial relationship.
- Write a description of the relationship between these objects.
- Create a question-and-answer pair based on the description.
- Each question must stand alone and not refer to previous questions.
- Use object IDs when referring to objects to avoid confusion.
- Focus on contrasting positions of objects (left, right, behind, above, near, etc.).

Template:
D[number]
<target region_id> (obj labels): Description of target object
<anchor region_id> (obj labels): Description of anchor object
<position relation>

Q[number] Segment object which/what/that <description of target object> <position relation> <anchor region_id> (obj labels)?

A[number] It is <target region_id> (obj label).

Rules:
- D[number], Q[number], and A[number] must start exactly as shown.
- Treat anchor region_id and target region_id as nouns.
- Do not use the words "*region*" or "*region id*"; use only the numeric ID.
- Write sentences that are natural, smooth, and grammatically correct.
- Do not mention "*coordinates*", "*polygons*", or "*label names*" explicitly.
- The description should focus on the usage and nature of the target object, not just its appearance.
- Do not include positional information in the description (position is specified later).
- The description must uniquely distinguish the target object from all other objects in the image.

Special handling for persons and riders:
- If the target object is a person (region id 11_) or a rider (region id 12_), describe them simply as "*individuals*".
- Apply this rule only when appropriate — ensure a fair sampling of other object classes as well.

Spatial relations:
- Use polygons to determine the 3D relationship between objects.
- Relations should reflect realistic spatial context — the anchor and target should be close enough for a meaningful reference.
- Use the viewer's perspective, not the orientation of the anchor object.
- For "*in front of*" / "*behind*", the closer object is "*in front of*" the farther object.

Additional requirements:
- Each label may have multiple region IDs — differentiate them carefully.
- Do not mention the target object's label name in the description.
- Answers must strictly follow the format with no additional information.
- Generate exactly 10 question-and-answer pairs for each query.
- Ensure diversity in label names, position relations, and descriptions.
- Anchor and target objects must have different region IDs.
- Include only questions that have definite, unambiguous answers.

"'

Table 5. System prompt used to generate Conditional-relationship Visual Instruction Set

Figure 3. Example instructions from the conditional visual instruction dataset. Red masks represent target objects, and green masks indicate anchor objects. Instructions specify segmentation tasks using semantic descriptions and spatial relationships relative to anchor objects.