Training-Free Zero-Shot Semantic Segmentation with LLM Refinement —SUPPLEMENTAL MATERIALS—

Yuantian Huang[1,2]
huang_yuantian@cyberagent.co.jp
Satoshi Iizuka[2]
iizuka@cs.tsukuba.ac.jp
Kazuhiro Fukui[2]
kfukui@cs.tsukuba.ac.jp

[1] CyberAgent, Inc. Tokyo, Japan
[2] University of Tsukuba
Tsukuba, Japan

1 Additional Examples in Pre-Refinement

During the pre-refinement process, additional examples are optional and could be automatically generated if an annotated segmentation dataset is available. Our experiments indicate that overall accuracy would improve with these additions, as shown in Table 1. Our paper does not include additional experiment examples because our proposed method focuses on a zero-shot scenario.

The system prompt for GPT-4 to create examples and explanations based on the following template:

Task Description:
- You will provide detailed explanations for example inputs and outputs within the context of the task.

Please adhere to the following rules:
- Exclude terms that appear in both lists.
- Detail the relevance of unmatched terms from input to output, focusing on indirect relationships.
- Identify and explain terms common to all output lists but rarely present in input lists; include these at the end of the output labeled 'Recommend Include Labels'.
- Each explanation should be concise, around 50 words.

Output Format:
- '1. Input... Output... Explanation... n. Input... Output... Explanation... Recommend Include Labels: label1, labeln, ...'

Table 1: **Quantitative Comparison** of the impact of additional examples.

Methods	DRAM	PASCAL VOC 2012	Cityscapes	COCO-81
w/o examples	62.01	63.57	34.36	37.45
w/ examples	**62.42**	**64.03**	**35.42**	**38.66**

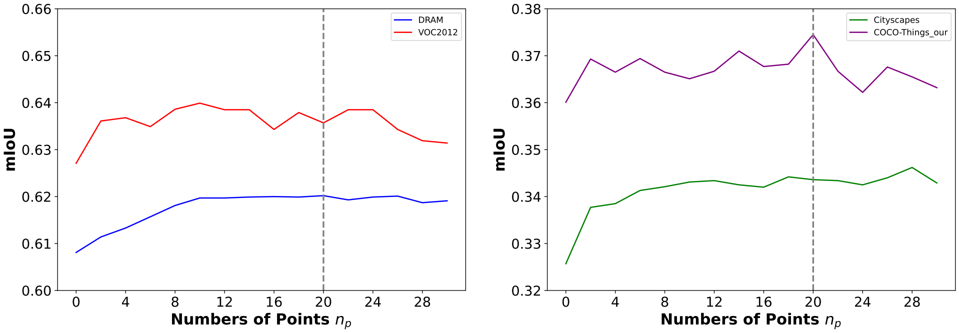

Figure 1: **Analysis of Points Prompt.**

The generated examples and explanations would be attached to the original prompt as an additional prompt for the pre-refinement process.

2 Analysis of Points Prompt

The relations regarding how varying the number of points prompts used affects overall accuracy are shown in Figure 1. Experiments suggest that using $n_p = 20$ points as prompts may be optimal.