

# Towards More Unified In-context Visual Understanding

## Supplementary Material

Hyperparameter	GPT-2	Ours
Architecture	transformer decoder	transformer decoder
Vocabulary size	50257	51290 / 52295
Max positions	1024	2060
Hidden size	768	768
Hidden layers	12	12
Attention heads	12	12
Number of MoEs	-	6
Number of experts	-	8
Layer norm epsilon	1e-05	1e-12
Attention probs dropout prob	0.1	0.1
Hidden dropout prob	0.1	0.1

Table 1. Hyperparameters for our GPT-2 baseline and proposed model. Note for the vocabulary size, we experiment with two settings for whether to add the special category and bbox tokens.

Examples	B@4 $\uparrow$			CIDEr $\uparrow$		
	1	2	3	1	2	3
$\mathcal{L}_{out}$	<b>5.4</b>	1.8	1.6	<b>95.9</b>	52.5	51.2
w/ $\mathcal{L}_{in}$	4.5	1.9	1.9	85.0	<b>54.7</b>	54.5
w/ $0.5\mathcal{L}_{in}$	5.3	<b>2.0</b>	<b>2.0</b>	86.6	54.6	<b>55.6</b>

Table 2. Loss analysis on class-aware in-context captioning task.

Examples	CA-ICL segmentation		CA-ICL captioning		
	MIoU $\uparrow$	MAE $\downarrow$	B@4 $\uparrow$	CIDEr $\uparrow$	mAP $\uparrow$
0	45.70	0.094	2.8	65.7	0.2
w/o category information for CA-ICL captioning					
1	56.17	0.167	6.6	95.6	1.5
2	59.21	0.132	1.7	45.5	1.7
3	60.85	0.128	0.8	32.6	1.6
w category information for CA-ICL captioning					
1	58.04	0.110	5.3	86.9	10.9
2	61.65	0.101	2.3	60.9	0.8
3	62.33	0.098	2.3	62.2	1.5

Table 3. Analysis on in-context samples and category information. We report the metrics utilized in our main experiments for the two CA-ICL tasks.

## 1. Model Architecture and Configuration

We provide more details on the model architecture compared with GPT-2 small as shown in Table 1, where each

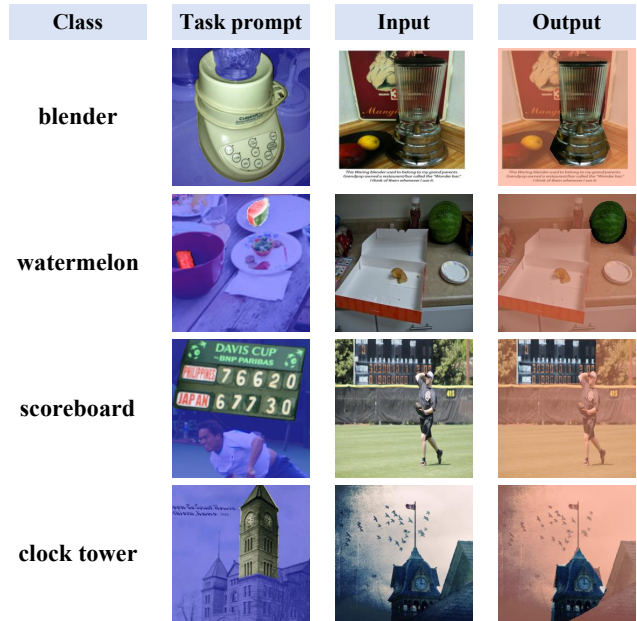


Figure 1. Out-of-domain test in CA-ICL segmentation. We employ the non-overlapping classes of the LVIS dataset to create a per-class mask pool, following the same approach used with the MS-COCO dataset.

dense decoder layer is the same and the even-numbered layer is replaced with the sparse decoder layer as discussed in Section 3.3 of our main submission. Another notable difference is the vocabulary size. We employ two different settings: one that includes special category and bbox tokens, and another without them. Compared with GPT-2, we adopt a smaller layer norm epsilon of 1e-12 to ensure stable training.

## 2. Additional Quantative Analysis

**Loss and padding analysis for CA-ICL captioning.** In this section, we delve into the effects of loss functions and padding strategies on captioning performance. We employ the baseline loss  $\mathcal{L}_{out}$ , delineated in Section 3.3 of our main submission, aligning the length of caption tokens with image tokens via padding. We examine different CE loss weights for the input image tokens in each context (denoted as  $\mathcal{L}_{in}$ ). This setting is based on the intuition that image captioning task may benefit from an increased focus on visual content because of the unbalanced sequence length between the image and text tokens. As indicated in Table 2, the experiment results reveal that using only  $\mathcal{L}_{out}$  surpasses other configurations with  $\mathcal{L}_{in}$  in a one-shot setting. In con-










Class	Task prompt	Input	Output
wool	 wool of a lamb.		 wool being shaved off a sheep.
wrist	 a blue band on a wrist.		 the wrist of a man.
yard	 train yard full of trains.		 two dogs in backyard.
wine	 glass of red wine.		 bottle of wine on the counter.

Figure 2. Out-of-domain test in CA-ICL captioning. We employ the non-overlapping classes of the Visual Genome dataset to create a per-class pool, following the same approach as discussed in Section 4.1 of our main submission.

trast, a composite loss of  $0.5\mathcal{L}_{in} + \mathcal{L}_{out}$  achieves superior results in two- and three-shot scenarios. Consequently, we adopt the  $0.5\mathcal{L}_{in} + \mathcal{L}_{out}$  loss for individual captioning tasks, while utilizing a consistent  $\mathcal{L}_{out}$  during co-training sessions.

**In-context effectiveness analysis.** We study the impact of increasing the number of in-context pair examples and whether to add category information in CA-ICL captioning in the task prompt on the outcomes. We trained the model using three in-context samples and inference with 1 to 3 samples. Additionally, we trained one model with explicit class information input instead of in-context samples for comparison. As presented in Table 3, with only class information, the model performs pool on both tasks for the CA-ICL segmentation task, which indicates the effectiveness of in-context samples as they are given more information than the simple class label. The inclusion of more examples consistently improves the segmentation performance. However, for the CA-ICL captioning task, the performance does not exhibit a steady increase, even more serious if category information is not provided. The possible reason is that using more in-text samples for the segmentation task can provide more segmentation clues coming from different views, and appearances of different image samples for the same category of target object. But for the caption task, one caption is already enough to denote the target object while multiple description styles from different samples will introduce more style ambiguity. From the perspective of performance, we report the best results of the model with class information for the captioning task. The problem of captioning is left to further study.

**Time cost of different models.** We calculate the fps metric for the 1 in-context example setting to analyze the time cost. The inference speed of our model using 0, 1, 2, 3 in-context examples is 2.8, 2.4, 1.9, 1.4 fps, respectively. While for SegGPT and OpenFlamingo, the fps is 7.7 and 0.3 img/s. Our model is capable of using in interactive applications.

### 3. Additional Qualitative Results

**In-context reasoning.** To illustrate the in-context reasoning ability of our model, we provide qualitative results on the two CA-ICL tasks. As illustrated in Figure 3, the model shows excellent semantic understanding for both in-door and out-door scenarios. Given suitable input prompts, the model demonstrates exceptional reasoning capabilities in segmenting instances that belong to the same category as the in-context samples. Figure 4 showcases the model’s ability to generate accurate captions with locations that precisely identify the region of the desired category, demonstrating its strong reasoning capabilities as well.

**Out-of-domain tests.** To evaluate the efficacy of semantic clues and the model’s capabilities, we conducted out-of-domain tests on two distinct tasks. As illustrated in Figure 1, the model demonstrates proficiency in utilizing cues from in-context examples featuring categories not encountered during training, thereby achieving dependable segmentation results. Additionally, for the captioning task, we utilized a per-category pool derived from the Visual Genome dataset, specifically selecting category data that do not coincide with the training categories. The results shown in Figure 2 further revealed the model’s ability to generalize effectively to unfamiliar categories.

### 4. Limitation and future work

Because of the long-tailed class and object scale distribution of the training dataset, the model does not perform well with multiple small objects or uncommon classes like traffic light. Some typical failure cases are presented in Figure 5 and Figure 6. We think a more balanced data distribution may be beneficial for the situation. For example, utilizing Copy-Paste strategies [1, 2] to expand the per-category instance pool. For improving captioning, a potential solution involves resampling the data or other data balancing strategies. Another limitation is that the model only supports one class per forward. Currently, the proposed model can support multiple categories by multiple times inference. The color mapping strategy utilized in SegGPT might be helpful.

The proposed method can accommodate a more diverse range of in-context learning tasks beyond the scope of class-aware tasks. As we discussed in Section ??, the multi-modal input will be quantized and mapped into the unified representation space. Therefore, all modal inputs quan-

tized by modality-specific quantizers can be modeled using our framework, regardless of the task. Next, we plan to extend M<sup>2</sup>oEGPT to accommodate even more modalities (*e.g.*, web page, 3D vision, heat map, tables) and tasks (*e.g.*, image generation and editing, inpainting, and grounding), also support high-resolution image and longer output, broadening the system’s applicability such that it becomes more general.

## References

- [1] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1301–1310, 2017. [2](#)
- [2] Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, et al. X-paste: Revisit copy-paste at scale with clip and stablediffusion. *arXiv preprint arXiv:2212.03863*, 2022. [2](#)











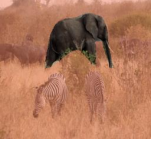

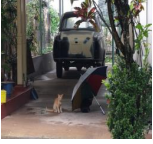


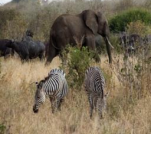


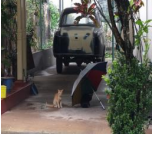





























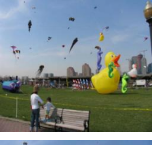




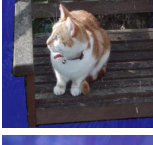
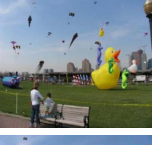

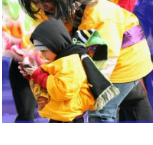




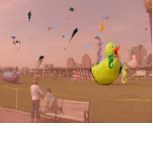
Class	Task prompt	Input	Output	Class	Task prompt	Input	Output
cow				cat			
							
							
apple				bowl			
spoon				cup			
bottle				book			
bowl				banana			
bike				kite			
horse				bench			
person				bird			

Figure 3. Results of CA-ICL segmentation. Our model demonstrates robustness across in-context prompts, effectively handling objects from diverse classes and accommodating variations in size and quantity. For better visualization, we overlay the mask onto the corresponding image. In this setup, the blue area indicates the mask for in-context prompts, while the red area represents the output mask.

Class	Task prompt	Input	Output	Class	Task prompt	Input	Output
bowl	 The <b>bowl</b> contains food.		 purple <b>cabbage</b> in a <b>bowl</b> .	cat	 an inside <b>cat</b> laying down.		 the <b>cat</b> is on a desk.
cup	 A <b>cup</b> on the table		 small cup of <b>sauce</b> .	backpack	 man carrying <b>backpack</b> .		 a red and black <b>backpack</b> .
fork	 a silver <b>fork</b> lying on the side of the plate.		 a <b>fork</b> is on the table.	chair	 empty black <b>chairs</b> .		 a pair of gloves on the back of an office <b>chair</b> .
sandwich	 <b>sandwich</b> is next to fruits.		 a <b>sandwich</b> on the plate.	laptop	 a black <b>laptop</b> on top of the desk.		 a computer <b>laptop</b> in the foreground.
book	 blue <b>book</b> in hands of kid.		 a <b>book</b> on the table.	train	 <b>train</b> going down the track.		 <b>train</b> on the tracks.
spoon	 <b>spoon</b> next to coffee cup.		 a large metal <b>spoon</b> .	clock	 a blue half <b>clock</b> .		 the <b>clock</b> positioned towards the left on the pole.
laptop	 built in keyboard of <b>laptop</b> .		 <b>laptop</b> sitting on a table.	motorcycle	 two people on a <b>motorcycle</b> .		 red and black <b>motorcycle</b> .
cup	 a small <b>cup</b> of mayo.		 tall glass <b>cup</b> with a black base.	bench	 horizontal gray slats on <b>bench</b> .		 <b>benches</b> in the background.
horse	 a brown <b>horse</b> eating grass by the water.		 gray <b>horse</b> grazing on grass.	dog	 a black <b>dog</b> jumping.		 a <b>dog</b> on the bus.
zebra	 <b>zebras</b> drinking in water		 <b>zebra</b> with black and white stripes.	bus	 a <b>bus</b> is on its regular route.		 a <b>bus</b> on the street.
bowl	 <b>bowl</b> with black rim.		 the <b>bowl</b> is clear.	car	 a white <b>car</b> in a parking lot.		 <b>car</b> driving on the street.

Figure 4. Qualitative results of CA-ICL captioning. Our model shows great semantic reasoning, accurately interpreting clues within in-context samples. It generates relatively precise bounding boxes and descriptions that correspond well with the desired objects in the images. To enhance visual clarity, we illustrate bounding boxes in in-context samples using red squares  $\square$ , and the predicted bounding boxes are marked in green  $\square$ . Additionally, we emphasize category information in the captions by using red text.




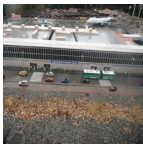


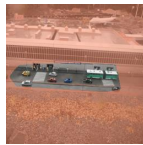

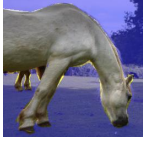

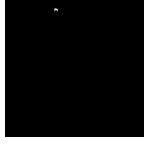




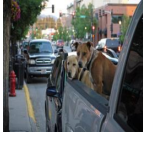
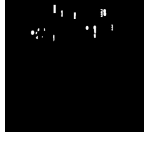

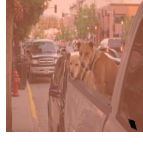

Class	Task prompt	Input	Ground truth	Output	Visualization	SegGPT
car						
horse						
traffic light						

Figure 5. Typical failure case for CA-ICL segmentation. The model faces challenges in processing input images with numerous small instances and also performs worse with categories that are infrequently represented in the training data. Zoom in for a better view.


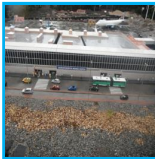










Class	Task prompt	Input	Ground truth	Output
car	 the car is blue.		 cars are driving down the street.	 car going down a street.
cow	 white cow in green field.		 group of black and white cows.	 three cows in a pasture.
dog	 a dog laying down.		 a large black and white dog.	 the one brown dog is laying down in the grass.

Figure 6. Typical failure case for CA-ICL captioning. When facing multiple instances or small instances, the model may predict inaccurate region location or wrong caption as highlighted in yellow. Zoom in for a better view.