

Supplementary Materials: RCA: Region Conditioned Adaptation for Visual Abductive Reasoning

Anonymous Authors

A EXTRA ABLATION EXPERIMENTS

A.1 Comparision with the CPT-CLIP on val-set

We compare RCA with the CPT on the Sherlock val set in Table 1. For the CPT baseline, we implement it under two specific settings: the “full fine-tune” and the “adapter⁺” + dual-contrastive loss” settings. Hereby, the up arrow \uparrow (or down arrow \downarrow) indicates the higher (or lower), the better.

We observe that our RCA outperforms the current SOTA method CPT on the validation set under all evaluation metrics. Moreover, we observe that the CPT can also benefit from the “Adapter⁺ + Dual-Contrastive Loss” tuning. This indicates the new adapter and loss tuning are generalizable. To conclude, on the validation set, this performance improvement is consistent with the test set in the main paper §4.2, indicating its robustness.

A.2 Comparison to Tiny Attention

Tiny Attention [2] was initially proposed for adapting attention heads of language models for downstream tasks. It shares a spirit similar to our RCA; we implement it for Vision-Language domains and compare it with our method.

As in Table 2, we compare the performances of our RCA with different settings of “TinyAtten + Adapter_M / Adapters_(A & M)” (Figure 1) against the RCA on frozen CLIP ViT-B-16. We note that “TinyAtten + Adapters_(A&M)” performs worse than the RCA with more tuned parameters and FLOPs. The reason might be Map Adapter only re-weights the attention map and does not change the “value” bases. Overall, the RCA is a more effective and efficient adapter than its counterparts.

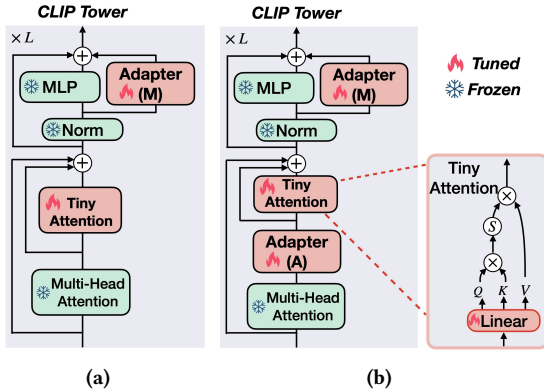


Figure 1: Tiny Attention Counterparts. (a): TinyAtten + Adapter_M; (b): TinyAtten + Adapter_A&M

A.3 Influence of Image Resolutions.

We test RCA with input combo images of different resolutions. We want to study whether the RCA can benefit from more tokens.

As in Table 3, it is straightforward to find an increment of computations when resolutions become larger (i.e., FLOPs 12.85G \rightarrow 41.48G \rightarrow 90.10G). However, the performance boost is not linear to the resolutions, reaching a saturate performance at the resolution of 448times224. This might lie in that the CLIP is pre-trained at 224 \times 224 resolution on the upstream dataset; thereby, downstream tuning is better to process images (one 448 \times 224 combo image = two 224 \times 224 images) at similar settings. Considering the trade-off of FLOPs and performance, we pick the resolution of 448 \times 224 for the CLIP ViT-B-16 backbone, in other ablation experiments.

A.4 Effects of Backbones

We further test the RCA with different backbones, namely CLIP ViT-B16 and CLIP ViT-L14 (336) on validation set.

We observe that a larger backbone contains more encoders, thereby increasing both tuned parameters (i.e., 43.28M \rightarrow 89.63M) and FLOPs (i.e., 41.48G \rightarrow 408.00G). This computation cost paid off, as performance under all evaluation metrics increased significantly (see Table 4).

A.5 Influence of Dimension d in Adapters.

We give qualitative results for setting adapters with different dimensions d . Table 5 is consistent with Figure 6 in §4.4 of the main paper, with exact values. We highlight the best/second-best values with bold and underlined denotations.

We observe that $d = D/4$ is the optimal setting among different evaluation metrics and achieves a good balance of performance and computations.

B MORE QUALITATIVE RESULTS OF RCA

We present more visual examples of retrieving the most likely inferences (hypothesis) with our RCA method in Figure 2.

We observe that the RCA could effectively retrieve human-like inference, with regional clues of different sizes, from tiny clues (e.g., “beer can” in Example 3, “necklace ring” in Example 12) to large ones (e.g., “captain” in Example 9)

REFERENCES

- [1] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2021. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797* (2021).
- [2] Hongyu Zhao, Hao Tan, and Hongyuan Mei. 2022. Tiny-Attention Adapter: Contexts Are More Important Than the Number of Parameters. In *Conference on Empirical Methods in Natural Language Processing*.

Table 1: Comparison with CPT using the Sherlock Validation Set.

<i>Val-Set</i> Model	Retrieval			<i>Localization</i> GT/Auto-Box (↑)	Comparison Human Acc (↑)
	im→txt (↓)	txt→im (↓)	P@1 _{i→t} (↑)		
CPT [1] (full fine-tune)	19.03	20.66	31.10	85.05 / 38.37	25.07
↳ Adapter ⁺ + Dual-Contrast Loss	17.99 (-1.04)	19.71 (-0.95)	31.94 (+0.84)	86.22 / 39.98 (+1.17 / 1.61)	25.53 (+0.46)
Our RCA (R-CTX)	16.30	17.92	33.09	86.10 / 40.80	25.83
↳ Mixed Prompts	15.16 (-1.14)	16.96 (-0.96)	34.57 (+1.48)	87.96 / 41.60 (+1.86 / 0.80)	25.64 (-0.19)
↳ Dual-Contrast Loss	14.26 (-0.90)	16.44 (-0.52)	35.46 (+0.89)	88.23 / 41.91 (+0.27 / 0.31)	26.80 (+1.16)

Table 2: Comparison of Map Augmented Adapter and Tiny Attention Adapter

<i>Val-Set</i> Attention Adapter	FLOPs (G↓)	Parameters Tuned (M↓)	Retrieval			<i>Localization</i> GT/Auto-Box (↑)	Comparison Human Acc (↑)
			im→txt (↓)	txt→im (↓)	P@1 _{i→t} (↑)		
Tiny Atten + Adapter_M	41.82	42.26	14.91	17.17	34.33	87.87 / 42.18	26.03
Tiny Atten + Adapter_(A & M)	43.33	47.39	14.68	16.74	34.90	87.68 / 41.91	25.46
Our RCA	41.48	42.26	14.26	16.44	35.46	88.23 / 41.91	26.80

Table 3: Impact of input image resolution.

<i>Val-Set</i> Resolution	FLOPs (G↓)	Retrieval P@1 _{i→t} (↑)	<i>Localization</i> GT/Auto-Box (↑)	Comparison Human Acc (↑)
224×112	12.84	33.12	86.90 / 41.74	25.38
448×224	41.48	35.46	88.23 / 41.91	26.80
672×336	90.10	34.93	88.30 / 42.44	26.77

Table 4: Comparison of ViT-B-16 and ViT-L-14 backbones

<i>Val-Set</i> Model Backbone		FLOPs (G)	Params Tuned (M)	Retrieval			<i>Localization</i> GT/Auto-Box (↑)	Comparison Human Acc (↑)
				im→txt (↓)	txt→im (↓)	P@1 _{i→t} (↑)		
RCA	ViT-B16	41.48	42.26	14.26	16.44	35.46	88.23 / 41.91	26.80
	ViT-L14 (336)	408.00	89.63	10.85	12.64	39.40	89.70 / 44.20	32.53

Table 5: Comparison of different bottleneck dimension on the performance.

<i>Val-Set</i> Dim of Adapter (<i>d</i>)	FLOPs (G)	Parameters Tuned (M↓)	Retrieval			<i>Localization</i> GT/Auto-Box (↑)	Comparison Human Acc (↑)
			im→txt (↓)	txt→im (↓)	P@1 _{i→t} (↑)		
<i>D</i> /32	37.19	28.83	16.55	19.04	33.04	86.91 / 40.68	24.99
<i>D</i> /16	37.80	30.75	15.43	17.77	33.74	87.20 / 41.40	25.79
<i>D</i> /8	39.03	34.59	14.68	17.01	34.71	87.68 / 42.07	<u>27.35</u>
<i>D</i> /4	41.48	42.26	14.26	16.44	<u>35.46</u>	88.23 / 41.91	26.80
<i>D</i> /2	46.38	57.61	<u>14.43</u>	16.55	35.48	88.25 / 42.05	27.08
<i>D</i> /1	56.18	88.32	14.54	<u>16.45</u>	35.76	88.08 / 41.80	28.25



Human Inference:
"Someone has already opened the top of the Heineken can for consumption."

AI Retrieved Inferences:
 1. "Someone has already opened the top of the Heineken can for consumption."
 2. "Someone was just enjoying a beer."
 3. "Someone has been drinking beer."
 4. "Someone in the room drank the beer."
 5. "Someone drank a beer."

(a) Example 3



Human Inference:
"You can get a meal here."

AI Retrieved Inferences:
 1. "That is the name of the restaurant."
 2. *"You can get a meal here."*
 3. "Its the store name."
 4. "The restaurant location is within the United States."
 5. "This is the name of the store."

(b) Example 4



Human Inference:
"Someone in the house won the trophy for a big achievement."

AI Retrieved Inferences:
 1. "Someone in the house won the trophy for a big achievement."
 2. "This is a trophy someone won in a contest."
 3. "Person holding the trophy is handing out of prize."
 4. "Someone who lives in this house won the trophy."
 5. "Someone won a competition."

(c) Example 5



Human Inference:
"It is the dog's birthday."

AI Retrieved Inferences:
 1. "The dog is the family pet."
 2. *"It is the dog's birthday."*
 3. "The dog is cherished as a part of the family."
 4. "The dog is hungry."
 5. "The person on the floor is the animal's master, the animal is a dog."

(d) Example 6



Human Inference:
"The men are farmers."

AI Retrieved Inferences:
 1. "The owner likes to wear overalls."
 2. "He has been working on the farm."
 3. "They work on a farm."
 4. "This man works on a farm."
 5. *"The men are farmers."*

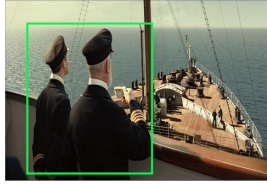
(e) Example 7



Human Inference:
"This is near Christmas."

AI Retrieved Inferences:
 1. "It is around the Christmas Holiday."
 2. *"This is near Christmas."*
 3. "It is near the Christmas holiday season."
 4. "It is around Christmas time."
 5. "It is near Christmas."

(f) Example 8



Human Inference:
"They are officers on a ship."

AI Retrieved Inferences:
 1. *"They are officers on a ship."*
 2. "They are looking at people drowning overboard."
 3. "The men are on a ship powered by the wind."
 4. "They are a captain or officer."
 5. "The people are on a ship."

(g) Example 9



Human Inference:
"The vase is ornate."

AI Retrieved Inferences:
 1. *"The vase is ornate."*
 2. "They are inspecting the vase."
 3. "It is holding flowers for viewing."
 4. "The flowers are for decoration."
 5. "This is use for flowers when there are guests."

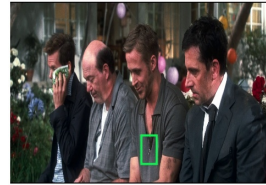
(h) Example 10



Human Inference:
"The person likes cats."

AI Retrieved Inferences:
 1. *"The person likes cats."*
 2. "This person loves cats as house pet."
 3. "Someone loves this cat."
 4. "This person loves this cat very much."
 5. "The resident owns a cat."

(i) Example 11



Human Inference:
"The man is married and wearing the ring around their neck instead of on their finger."

AI Retrieved Inferences:
 1. *"The man is married and wearing the ring around their neck instead of on their finger."*
 2. "The man likes to wear jewelry."
 3. "Its a necklace."
 4. "The man is religious."
 5. "The man like the present he has been given."

(j) Example 12



Human Inference:
"Drinks are sold in there."

AI Retrieved Inferences:
 1. "This is a pub."
 2. "This building is a bar."
 3. "The bull is here to eat."
 4. *"Drinks are sold in there."*
 5. "It is a bar."

(k) Example 13



Human Inference:
"She is a witch."

AI Retrieved Inferences:
 1. "This is a witch."
 2. *"She is a witch."*
 3. "The man is a wizard."
 4. "This child is playing a wizard."
 5. "The woman is sitting at the stage but looking down."

(l) Example 14

Figure 2: Qualitative results obtained by rpa. The machine retrieves the top-5 most likely inferences according to the box region. **Red sentence** indicates that the machine finds the same inference as a human expert.