Tag2Text: Guiding Vision-Language Model via Image Tagging

Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, Lei Zhang

https://github.com/xinyu1205/recognize-anything

Model Framework

- The core of Tag2Text lies in the introduction of **image tagging** supervised by the **annotation-free image tags** parsed from its paired text.
- **Image Tagging**: Training robust tagging model from large-scale image-text pairs. Textual label queries based on CLIP text encoder empower **open-set tagging**.
- **Image Captioning**: Tag2Text learns to generate text related to the image by leveraging the automatically parsed tags, resulting in **comprehensive and controllable texts with the guidance of recognized tags**.

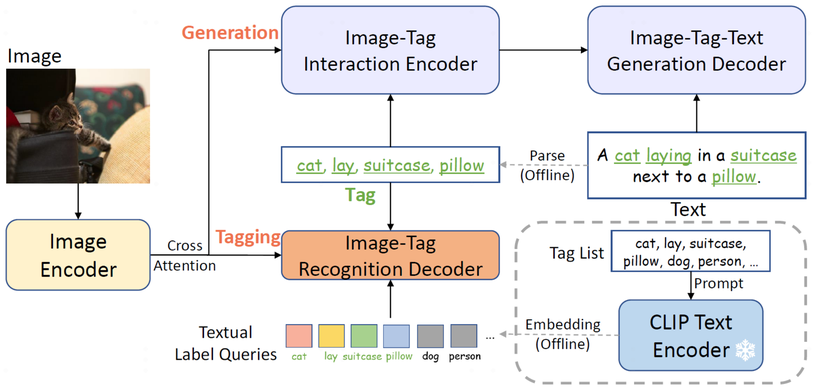
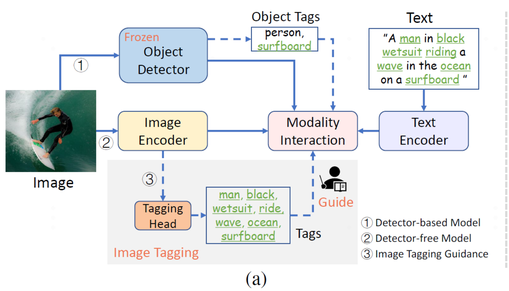


Image Tagging vs. Object Detection

- Prior works demonstrate the effectiveness of incorporating object tags into VL models based on an off-the-shelf detector.
- Since the **detector restricts the model's capacity and is time-consuming**, recent VL models normally avoid using a detector, resulting in poor utilization of valuable tags.
- We re-introduce tag guidance into detector-free VL models via image tagging with a simple tagging head. The tagging head is supervised by annotation-free image tags parsed from its paired text. **Our model achieves a superior tagging ability and effectively enhances vision-language tasks.**



(a)

Method	Object Detector	Image Tagging
No Manual Annotation	✗	✓
Enable End-to-End	✗	✓
Tag Categories	Objects	Objects, Scenes, Attributes, Actions
Additional Parameters	⩾ 42M	⩽ 5M
Running Time	∼ 153ms	∼40ms

(b)

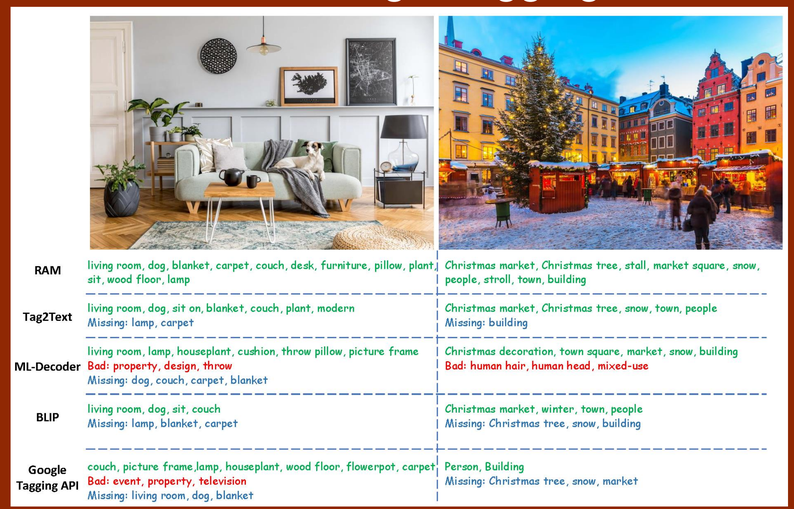
Highlight

We introduce Tag2text, a vision-language model guided by image tagging:

- Powerful Image Tagging.
- Comprehensive and Controllable Captioning.

From Tag2Text to Recognize Anything Model (RAM):

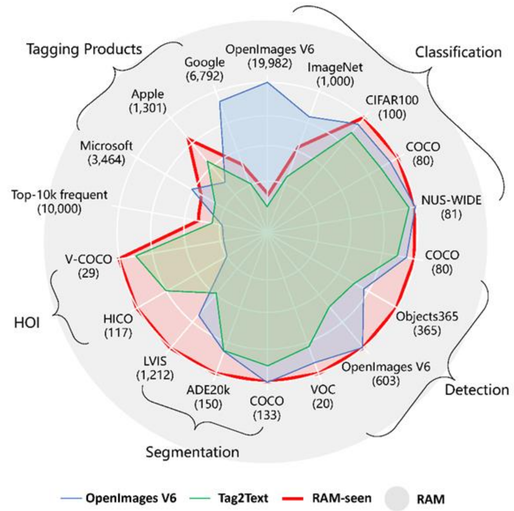
- More Powerful Image Tagging.



Codes, demos and models at

More Valuable Tags

- Our models offer **more comprehensive and commonly used tags**, including objects, scenes, attributes, and actions.
- Tag2Text recognizes **3,400+** fixed tags.
- RAM upgrades the number to **6,400+**, and develop **open-set capability**.



Zero-Shot Image Recognition

- Tag2Text showcases **superior zero-shot image recognition capabilities**, surpassing other vision-language models with significantly larger training dataset.

Methods	Pre-train #Images	Evaluation Paradigm	OPPO F1	OPPO Precision	OPPO Recall	OpenImages F1	OpenImages Precision	OpenImages Recall	COCO F1	COCO Precision	COCO Recall
CLIP (Radford et al., 2021)	400M	Alignment	63.4	76.6	54.1	63.0	77.9	52.9	48.2	64.0	38.7
DiHT (Radenovic et al., 2023)	438M	Alignment	66.8	75.3	60.0	66.3	77.0	65.3	48.9	51.4	46.7
BLIP (Li et al., 2022)	129M	Alignment	65.7	76.7	57.5	64.8	78.6	55.1	54.3	65.2	46.5
BLIP (Li et al., 2022)	129M	Captioning	58.6	79.1	46.6	56.6	73.7	45.9	55.7	93.0	39.8
BLIP-2 (Li et al., 2023)	129M	Captioning	58.2	72.8	48.5	58.1	74.2	47.8	59.1	**95.5**	42.8
Tag2Text (Ours)	14M	Captioning	65.9	**82.4**	54.9	62.7	76.7	53.0	62.7	93.2	47.2
Tag2Text (Ours)	4M	Tagging	75.7	76.6	74.8	71.8	79.7	65.3	**72.6**	80.5	**66.1**
Tag2Text (Ours)	14M	Tagging	**78.6**	77.9	**79.4**	**72.7**	**80.1**	**66.6**	71.5	80.1	64.5

- RAM further expands training image tags through an **automatic data engine** (Parsing + Generating + Cleaning).
- RAM's zero-shot generalization to OpenImages-common is superior to ML-Decoder's full supervision.

Methods	Tags‡	Multi-label Classification: OPPO -common	Multi-label Classification: OpenImages -common	Multi-label Classification: OpenImages -rare	Detection: COCO-80	Segmentation: COCO-133	Segmentation: ADE20k	Segmentation: ADE20k -clean
ML-Decoder [23]	33.9M	82.4†	85.8	**79.5**	72.8†	✗	✗	✗
MKT [8]	0.6M	78.2	77.8	63.5	62.9	51.0	37.1	38.4
Tag2Text-4M [10]	11.4M	83.0	82.9	✗	78.3†	66.9†	✗	✗
Tag2Text-14M [10]	33.6M	85.4	83.4	✗	78.2†	67.1†	✗	✗
RAM-4M	39.3M	85.6	86.0	66.7	79.0	68.3	51.5	53.2
RAM-14M	119.9M	**86.9**	**86.5**	69.2	**80.6**	**69.4**	**55.4**	**56.9**

Green means fully supervised learning; Blue means zero-shot performance.