Generative Prompt Model for Weakly Supervised Object Localization

Yuzhong Zhao[1], Qixiang Ye[1], Weijia Wu[2], Chunhua Shen[2], Fang Wan[1*]

[1] University of Chinese Academy of Sciences [2] Zhejiang University

Dataset	Ann.	# Images	How to collect	t (s/img)
CUB-200-2011 [21]	$\mathcal{I}$	11,788	Manual	1.5
Imagenet [17]	$\mathcal{I}$	14,197,122	Manual	1.5
JFT-3B [7]	$\mathcal{I}^\dagger$	3,000,000,000	Semi-automatic	≈ 0
CC12M [2]	$\mathcal{T}^\dagger$	12,000,000	Web crawler	≈ 0
WIT [16]	$\mathcal{T}^\dagger$	400,000,000	Web crawler	≈ 0
LAION-400M [19]	$\mathcal{T}^\dagger$	400,000,000	Web crawler	≈ 0
LAION-5B [18]	$\mathcal{T}^\dagger$	5,850,000,000	Web crawler	≈ 0
Cityscapes [6]	$\mathcal{B}$	25,000	Manual	37.5
COCO [11]	$\mathcal{B}$	328,000	Manual	37.5

Table 1: **The size and data collecting approaches of some commonly used datasets**. $\mathcal{I}, \mathcal{T}, \mathcal{B}$ denotes the image category labels, text descriptions and bounding box annotations respectively. t denotes the average annotation time per image. † indicates the annotation is noisy.

A. Annotation Cost

As shown in Table 1, we compare the size and data collection methods of commonly used datasets with three types of annotations: image category labels, text descriptions, and bounding boxes. It can be observed that datasets with accurate bounding box labels, such as Cityscapes and COCO, are usually small in size due to the high cost of manual annotation. However, when using image category labels, the dataset size can be increased to 14 million (e.g., ImageNet). For datasets with a huge size, such as JFT-3B, WIT, and LAION-5B, manual annotation becomes impractical. Instead, semi-automatic annotation methods or web crawler algorithms are used to extensively collect noisy annotated data. Thanks to the rapid development of the Internet, a large number of image-text pairs can be found in websites, forums, and libraries, which are naturally annotated by citizens and can be easily obtained by crawler algorithms. Since collecting image-text pairs hardly requires human participation, their annotation cost is negligible. In this paper, the proposed method GenPromp is implemented based the Stable Diffusion model, which is pre-trained on LAION-5B. Accordingly, GenPromp hardly introduces additional annotation cost for a weakly supervised learning system.

*Corresponding author

Embedding	ImageNet-1K					
	Top-1 Loc	Top-5 Loc	GT-known Loc	M-Ins	Part	More
f_d	62.2	70.0	71.5	9.4	3.8	8.2
f_r	64.9	73.1	74.7	9.5	2.4	7.3
f_c	65.2	73.4	75.0	9.1	3.0	6.9

Table 2: **Localization error statistics.** The results are correspond to row 17-19 in Table 6. "M-Ins", "Part" and "More" denote the multi-instance error, localization part error and localization more error respectively.

B. Prompt Ensemble

To further improve the performance of GenPromp, we propose a prompt ensemble strategy. As shown in Fig. 1, during training, we random select a template from a template set. Then, we respectively fill the *meta* token (`goldfish`) and the *concept* token (⟨`goldfish`⟩) into the template to obtain the two input prompts, which are used to learn the representative embedding f_r. During inference, for each template in the template set, we combines it with the two tokens to form the input prompts. Then, all the prompts are encoded into prompt embeddings by the pre-trained CLIP model. After that, the discriminative embedding f_d is obtained by averaging the discriminative embeddings generated by different templates (*e.g.* $f_{d_1}, f_{d_2}, f_{d_3}, f_{d_4}$ in Fig. 1), the representative embedding f_r is obtained by averaging the representative embeddings generated by different templates (*e.g.* $f_{r_1}, f_{r_2}, f_{r_3}, f_{r_4}$ in Fig. 1). Finally, f_d and f_r are combined to f_c, which is fed into the network to generate attention maps. In experiments, we use a template set that consists of 7 templates:

"`a photo of a { }`"
"`a rendering of a { }`"
"`the photo of a { }`"
"`a photo of my { }`"
"`a photo of the { }`"
"`a photo of one { }`"
"`a rendition of a { }`"

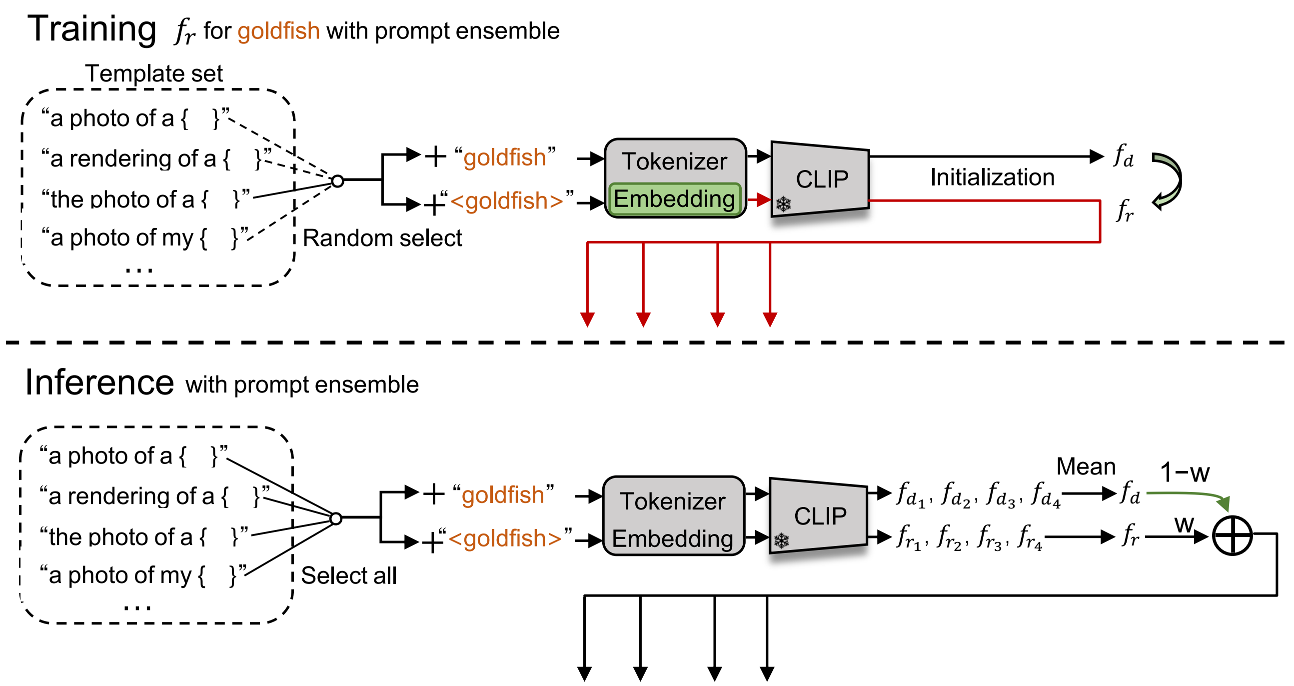

Figure 1: **Workflow of the proposed prompt ensemble strategy**. The image encoding and activate map generation procedure are omitted for clarity.

Noise ϵ	ImageNet-1K		
	Top-1 Loc	Top-5 Loc	GT-known Loc
	64.8	73.0	74.6
✓	**65.1**	**73.3**	**74.9**

Table 3: **Evaluation of noise levels in the inference time.** For the experiment that includes noise ϵ in the inference time, we conduct 10 experiments under different random seeds and average their results as the final result.

C. Additional Experimental Results

Complete Performance Comparison with SOTA Methods. Table 4 shows the complete performance comparison of the proposed GenPromp and the state-of-the-art (SOTA) models (extension of Table 1 in the main document). On CUB-200-2011 and ImageNet-1K dataset, GenPromp surpasses the SOTA methods by significant margins. Such strong results clearly demonstrate the superiority of the generative model over conventional discriminative models for weakly supervised object localization.

Localization Error Analysis. To further reveal the effect of the proposed prompt embeddings (*e.g.* f_d, f_r, f_c), following TS-CAM [8], we evaluate the localization errors of: multi-instance error (M-Ins), localization part error (Part), and localization more error (More). They are respectively defined as follows.

- M-Ins indicates that the predicted bounding box intersects with at least two ground-truth boxes, and IoG > 0.3.
- Part indicates that the predicted bounding box only cover the parts of object, and IoP > 0.5.
- More indicates that the predicted bounding box is larger than the ground truth bounding box by a large margin, and IoG > 0.7.

where IoG and IoP are defined as Intersection over Ground truth box and Intersection over Predict bounding box, respectively (similar to IoU (Intersection over Union)). Each metric calculates the percentage of images belonging to corresponding error in the validation/test set. Please refer to TS-CAM [8] for a detailed definition of the three metrics. Table 2 lists localization error statistics of M-Ins, Part, and More. Compare to the discriminative embedding f_d, the learned representative embedding f_r reduces both Part and More errors by 1.4% (3.8% $vs.$ 2.4%) and 0.9% (8.2% $vs.$ 7.3%) respectively, demonstrating that the representative embedding alleviates the partial object activation problem. By combining the representative embedding f_r with the discriminative embedding f_d, the More errors drop 0.4% (7.3% $vs.$ 6.9%) while the Part errors increase 0.6% (3.0% $vs.$ 2.4%) compared to f_r. This demonstrates that f_c can further depress the background noise while keeping relatively low Part errors.

Method	Loc Back.	Cls Back.	CUB-200-2011			ImageNet-1K		
			Top-1 Loc	Top-5 Loc	GT-known Loc	Top-1 Loc	Top-5 Loc	GT-known Loc
CAM$_{CVPR'16}$ [30]	VGG16		41.1	50.7	55.1	42.8	54.9	59.0
ADL$_{CVPR'19}$ [4]	VGG16		52.4	-	75.4	44.9	-	-
DANet$_{ICCV'19}$ [26]	VGG16		52.5	62.0	67.7	-	-	-
SLT$_{CVPR'21}$ [9]	VGG16		67.8	-	87.6	51.2	62.4	67.2
FAM$_{ICCV'21}$ [13]	VGG16		69.3	-	89.3	52.0	-	71.7
TAFormer$_{TPAMI'22}$ [14]	VGG16		72.0	85.9	90.8	53.4	67.7	74.0
BAS$_{CVPR'22}$ [23]	VGG16		71.3	85.3	91.1	53.0	65.4	69.6
CAM$_{CVPR'16}$ [30]	MobileNetV1		48.1	59.2	63.3	43.4	54.4	59.0
HaS$_{ICCV'17}$ [20]	MobileNetV1		46.7	-	67.3	42.7	-	60.1
ADL$_{CVPR'19}$ [4]	MobileNetV1		47.7	-	-	43.0	-	-
FAM$_{ICCV'21}$ [13]	MobileNetV1		65.7	-	85.7	46.2	-	62.1
TAFormer$_{TPAMI'22}$ [14]	MobileNetV1		66.7	80.2	85.0	47.6	65.5	68.8
BAS$_{CVPR'22}$ [23]	MobileNetV1		69.8	86.0	92.4	53.0	66.6	72.0
CAM$_{CVPR'16}$ [30]	ResNet50		46.7	54.4	57.4	39.0	49.5	51.9
ADL$_{CVPR'19}$ [4]	ResNet50-SE		62.3	-	-	-	-	48.5
FAM$_{ICCV'21}$ [13]	ResNet50		73.7	-	85.7	54.5	-	64.6
SPOL$_{CVPR'21}$ [22]	ResNet50		80.1	93.4	96.5	59.1	67.2	69.0
TAFormer$_{TPAMI'22}$ [14]	ResNet50		75.0	87.8	91.2	57.5	69.9	75.5
DA$_{CVPR'22}$ [31]	ResNet50		66.7	-	81.8	55.8	-	70.3
BAS$_{CVPR'22}$ [23]	ResNet50		77.3	90.1	95.1	57.2	67.4	71.8
CAM$_{CVPR'16}$ [30]	InceptionV3		41.1	50.7	55.1	46.3	58.2	62.7
DANet$_{ICCV'19}$ [26]	InceptionV3		49.5	60.5	67.0	47.5	58.3	-
SLT$_{CVPR'21}$ [9]	InceptionV3		66.1	-	86.5	55.7	65.4	67.6
FAM$_{ICCV'21}$ [13]	InceptionV3		70.7	-	87.3	55.2	-	68.6
TAFormer$_{TPAMI'22}$ [14]	InceptionV3		73.3	84.1	88.7	56.0	66.5	69.8
BAS$_{CVPR'22}$ [23]	InceptionV3		73.3	86.3	92.2	58.5	69.0	71.9
CREAM$_{CVPR'22}$ [25]	InceptionV3		71.8	86.4	90.4	56.1	66.2	69.0
TS-CAM$_{ICCV'21}$ [8]	Deit-S		71.3	83.8	87.7	53.4	64.3	67.6
LCTR$_{AAAI'22}$ [3]	Deit-S		79.2	89.9	92.4	56.1	65.8	68.7
SCM$_{ECCV'22}$ [1]	Deit-S		76.4	91.6	96.6	56.1	66.4	68.8
DiPS$_{WACVW'23}$ [15]	Deit-S	TransFG [10]	88.2	-	-	-	-	-
PSOL$_{CVPR'20}$ [28]	DenseNet161	EfficientNet-B7	80.9	90.0	91.8	58.0	65.0	66.3
C^{2}AM$_{CVPR'22}$ [24]	DenseNet161	EfficientNet-B7	81.8	91.1	92.9	59.6	67.1	68.5
GenPromp (Ours)	Stable Diffusion	EfficientNet-B7	87.0	96.1	**98.0**	65.1	73.3	74.9
GenPromp† (Ours)	Stable Diffusion	EfficientNet-B7	87.0	96.1	**98.0**	**65.2**	**73.4**	**75.0**
GenPromp† (Ours)	Stable Diffusion	TransFG [10]	**89.3**	**96.5**	**98.0**	-	-	-

Table 4: **Performance comparison** of the proposed GenPromp approach with the state-of-the-art methods on the CUB-200-2011 test set and ImageNet-1K validation set. *Loc Back.* denotes the localization backbone, *Cls Back.* the backbone for classification, and † the prompt ensemble strategy, which ensembles the localization results from multiple prompts.

Effect of Noise ϵ. In Table 3, we evaluate the performance by setting the noise ϵ (in Eq. 1 and Eq. 2 of the main document) to 0 during inference. Without noise ϵ, the performance of GenPromp drops 0.3% in Top-1 Loc in average. Similar to the methods [4, 5, 12, 20, 27, 29] based on adversarial erasing, the input noise in GenPromp can also alleviate the part activation issue, which drives the network to mine the representative yet less discriminative object parts.

Additional Restults with respect to Model Size and Training Data. In Table 5, we re-implement TS-CAM with larger backbone (*e.g.*Deit-B, ViT-L, ViT-H) and more training data (*e.g.*LAION-2B). As the model size getting larger, the performance of TS-CAM becomes worse on CUB-200-2011 under GT-known Loc metric, Table 5(upper). As shown in Table 5(lower), by finetuning ViT-H-based TS-CAM for 3 epochs on ImageNet-1K, it achieves higher classification accuracy (74.7% *vs.* 74.3% on Top-1 Cls) while much lower localization accuracy (53.2% *vs.* 67.6% on GT-known Loc) compared to the Deit-S-based TS-CAM. By finetuning the model for more epochs (*e.g.*6 epochs), it achievea higher classification accuracy (77.4% *vs.* 74.7% on Top-1 Cls) but lower localization accuracy (52.2% *vs.* 53.2% under GT-known Loc metric), demonstrating that more epochs can not improve the localization performance of TS-CAM. We attribute this phenomenon to the inherent flaw of the discriminatively trained classification model, *i.e.*, local discriminative regions are capable of minimizing image classification loss but experience difficulty in accurate object localization. A larger backbone and more training data make this phenomenon even more serious.

Method	Loc Back.	Cls Back.	Params.	CUB-200-2011				
				Top-1 Loc	Top-5 Loc	GT-known Loc	Top-1 Cls	Top-5 Cls
TS-CAM [8]	Deit-S (ImageNet-1K)		22.4M	71.3	83.8	87.7	80.3	94.8
TS-CAM [8]	Deit-B (ImageNet-1K)		87.2M	75.8	84.1	86.6	86.8	96.7
TS-CAM [8]	ViT-L (LAION-2B [18], ImageNet-1K, CUB(60epochs ft))		304M	63.4	76.0	80.1	77.3	93.8
TS-CAM [8]	ViT-H (LAION-2B [18], ImageNet-1K, CUB(60epochs ft))		633M	10.7	20.2	32.9	29.1	56.8
GenPromp†	Stable Diffusion	EfficientNet-B7	1017M + 66M	**87.0**	**96.1**	**98.0**	**88.7**	**97.9**

Method	Loc Back.	Cls Back.	Params.	ImageNet-1K				
				Top-1 Loc	Top-5 Loc	GT-known Loc	Top-1 Cls	Top-5 Cls
TS-CAM [8]	Deit-S (ImageNet-1K)		22.4M	53.4	64.3	67.6	74.3	92.1
TS-CAM [8]	ViT-H (LAION-2B [18], ImageNet-1K(3epochs ft))		633M	41.9	50.7	53.2	74.7	92.8
TS-CAM [8]	ViT-H (LAION-2B [18], ImageNet-1K(6epochs ft))		633M	42.1	49.9	52.2	77.4	93.7
GenPromp†	Stable Diffusion	EfficientNet-B7	1017M + 66M	**65.2**	**73.4**	**75.0**	**85.1**	**97.2**

Table 5: Performance comparison with respect to model size and training data. With a larger backbone and pre-training dataset, the discriminatively trained method TS-CAM does not achieve higher performance.

	Multi-resolution	Multi-timesteps	Prompt Ensemble	Prompt Embedding	Finetune	ImageNet-1K		
						Top-1 Loc	Top-5 Loc	GT-known Loc
1				f_d		58.5	66.0	67.4
2	✓			f_d		58.6	66.1	67.5
3		✓		f_d		58.6	66.0	67.5
4	✓	✓		f_d		61.2	69.0	70.4
5	✓	✓		f_r (w/o init)		44.6	50.2	51.3
6	✓	✓		f_r		64.0	72.1	73.7
7	✓	✓		f_c (w/o init)		56.2	63.2	64.5
8	✓	✓		f_c		64.5	72.7	74.2
9	✓	✓	✓	f_d		61.5	69.2	70.7
10	✓	✓	✓	f_r		64.2	72.3	73.8
11	✓	✓	✓	f_c		64.6	72.8	74.3
12	✓	✓		f_d	✓	62.0	69.8	71.4
13	✓	✓		f_r	✓	64.9	73.1	74.6
14	✓	✓		f_c	✓	65.1	73.3	74.9
15	✓		✓	f_c	✓	65.0	73.2	74.8
16		✓	✓	f_c	✓	62.3	70.3	71.8
17	✓	✓	✓	f_d	✓	62.2	70.0	71.5
18	✓	✓	✓	f_r	✓	64.9	73.1	74.7
19	✓	✓	✓	f_c	✓	**65.2**	**73.4**	**75.0**

Table 6: **Ablation of main components of GenPromp**. For experiments that do not have a "✓" in Multi-resolution or Multi-timesteps, we use a single resolution (16× 16) or a single timestep ($t = 100$) for model inference.

Detailed Ablation Study. Table 6 provides a detailed ablation of the performance contribution of each component and their combinations, with respect to Multi-resolution, Multi-timesteps, Prompt ensemble, Prompt embedding and Finetuning.

D. Additional Visualization Results

In Fig. 2, we visualize the localization results of GenPromp and compared them with the discriminatively trained model (*e.g.*CAM [30]). The object Localization maps of CAM (column b) suffer from partial object activation. Localization maps of GenPromp (column d) with sole representative embeddings (f_r) covers more object extent but introducing background noise. Those of GenPromp (column e) with combined embeddings (f_c) not only activate full object extent but also depress background noise for precise object localization.

We also provide additional visualization results of Fig. 2, Fig. 5 and Fig. 6 in the main document. The results are shown in Fig. 3, Fig. 4 and Fig. 5 respectively.

References

[1] Haotian Bai, Ruimao Zhang, Jiong Wang, and Xiang Wan. Weakly supervised object localization via transformer with implicit spatial calibration. In *ECCV*, pages 612–628, 2022. 3

[2] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE CVPR*, pages 3558–3568, 2021. 1

[3] Zhiwei Chen, Changan Wang, Yabiao Wang, Guannan Jiang, Yunhang Shen, Ying Tai, Chengjie Wang, Wei Zhang, and Liujuan Cao. LCTR: on awakening the local continuity of transformer for weakly supervised object localization. In *AAAI*, pages 410–418, 2022. 3

[4] Junsuk Choe, Seungho Lee, and Hyunjung Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. *IEEE TPAMI*, pages 4256–4271, 2021. 3

[5] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *IEEE CVPR*, pages 2219–2228, 2019. 3

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE CVPR*, pages 3213–3223, 2016. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[8] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. TS-CAM: token semantic coupled attention map for weakly supervised object localization. In *ICCV*, pages 2866–2875, 2021. 2, 3, 4

[9] Guangyu Guo, Junwei Han, Fang Wan, and Dingwen Zhang. Strengthen learning tolerance for weakly supervised object localization. In *IEEE CVPR*, pages 7403–7412, 2021. 3

[10] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *AAAI*, pages 852–860, 2022. 3

[11] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV*, pages 740–755, 2014. 1

[12] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *IEEE CVPR*, pages 8763–8772, 2020. 3

[13] Meng Meng, Tianzhu Zhang, Qi Tian, Yongdong Zhang, and Feng Wu. Foreground activation maps for weakly supervised object localization. In *IEEE ICCV*, pages 3365–3375, 2021. 3

[14] Meng Meng, Tianzhu Zhang, Zhe Zhang, Yongdong Zhang, and Feng Wu. Task-aware weakly supervised object localization with transformer. *IEEE TPAMI*, pages 1–13, 2022. 3

[15] Shakeeb Murtaza, Soufiane Belharbi, Marco Pedersoli, Aydin Sarraf, and Eric Granger. Discriminative sampling of proposals in self-supervised transformers for weakly supervised object localization. In *IEEE WACVW*, pages 1–11, 2023. 3

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *ICML*, pages 8748–8763, 2021. 1

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, pages 211–252, 2015. 1

[18] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. *CoRR*, abs/2210.08402, 2022. 1, 4

[19] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021. 1

[20] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *IEEE ICCV*, pages 3544–3553, 2017. 3

[21] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[22] Jun Wei, Qin Wang, Zhen Li, Sheng Wang, S. Kevin Zhou, and Shuguang Cui. Shallow feature matters for weakly supervised object localization. In *IEEE CVPR*, pages 5993–6001, 2021. 3

[23] Pingyu Wu, Wei Zhai, and Yang Cao. Background activation suppression for weakly supervised object localization. In *IEEE CVPR*, pages 14228–14237, 2022. 3

[24] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. C^2 AM: contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In *IEEE CVPR*, pages 979–988, 2022. 3

[25] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Rui-Wei Zhao, Tao Zhang, Xuequan Lu, and Shang Gao. CREAM: weakly supervised object localization via class re-activation mapping. In *IEEE CVPR*, pages 9427–9436, 2022. 3

[26] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *IEEE ICCV*, pages 6588–6597, 2019. 3

[27] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE ICCV*, pages 6022–6031, 2019. 3

[28] Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. Rethinking the route towards weakly supervised object localization. In *IEEE CVPR*, pages 13457–13466, 2020. 3

[29] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S. Huang. Adversarial complementary learning for weakly supervised object localization. In *IEEE CVPR*, pages 1325–1334, 2018. 3

[30] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE CVPR*, pages 2921–2929, 2016. 3, 4

[31] Lei Zhu, Qi She, Qian Chen, Yunfei You, Boyu Wang, and Yanye Lu. Weakly supervised object localization as domain adaption. In *IEEE CVPR*, pages 14617–14626, 2022. 3

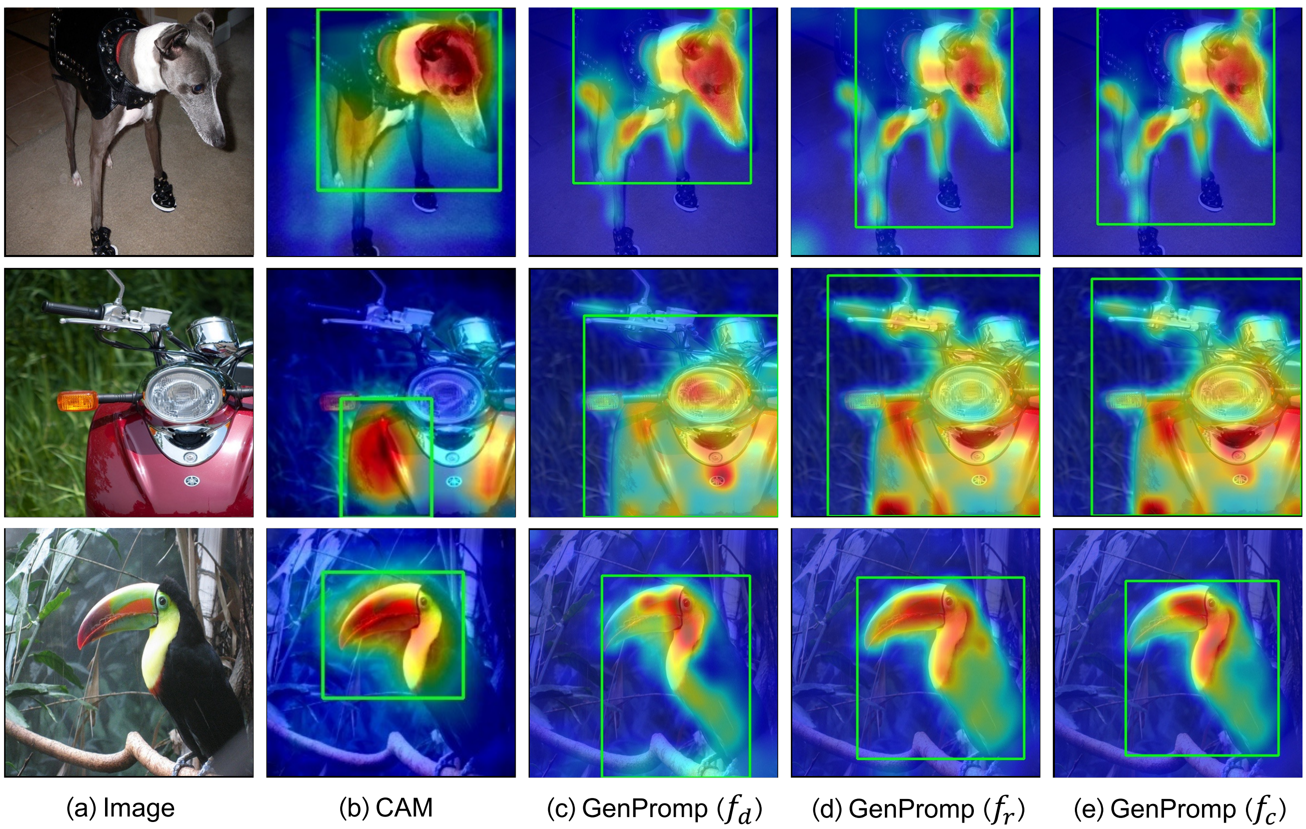

Figure 2: **Comparison of activation maps between CAM and GenPromp**.

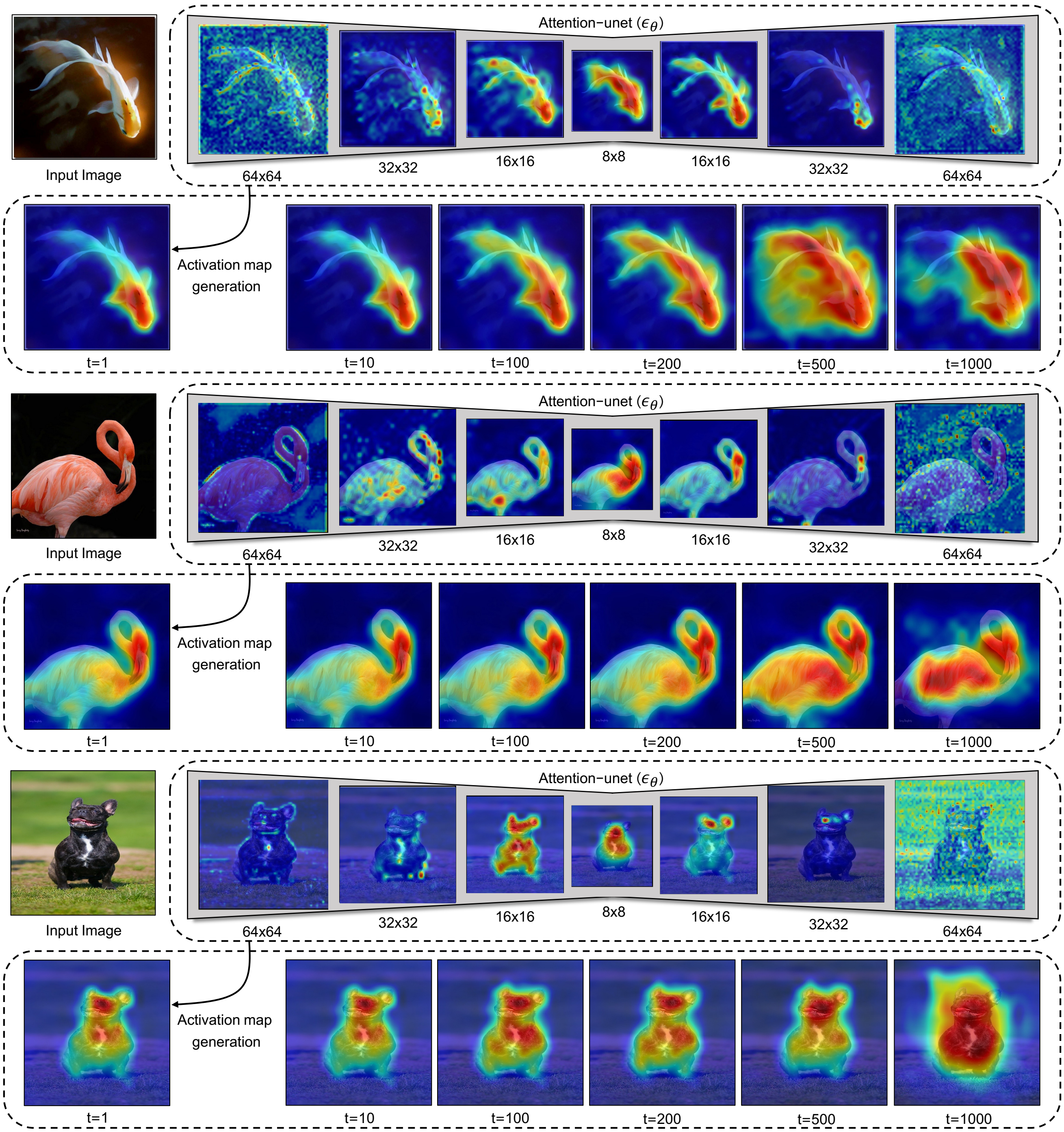

Figure 3: **Visualization of cross attention maps.** Attention maps with respect to multiple resolutions and multiple noise levels (timesteps t) are aggregated to obtain the final localization map. The characteristics of these attention maps can be concluded as follows: (1) Attention maps with higher resolution can provide more detailed localization clues but introduce more noise. (2) Attention maps of different layers can focus on different parts of the target object. (3) Smaller t provides a less noisy background but tends to partial object activation. (4) Larger t activates the target object more completely but introduces more background noise.

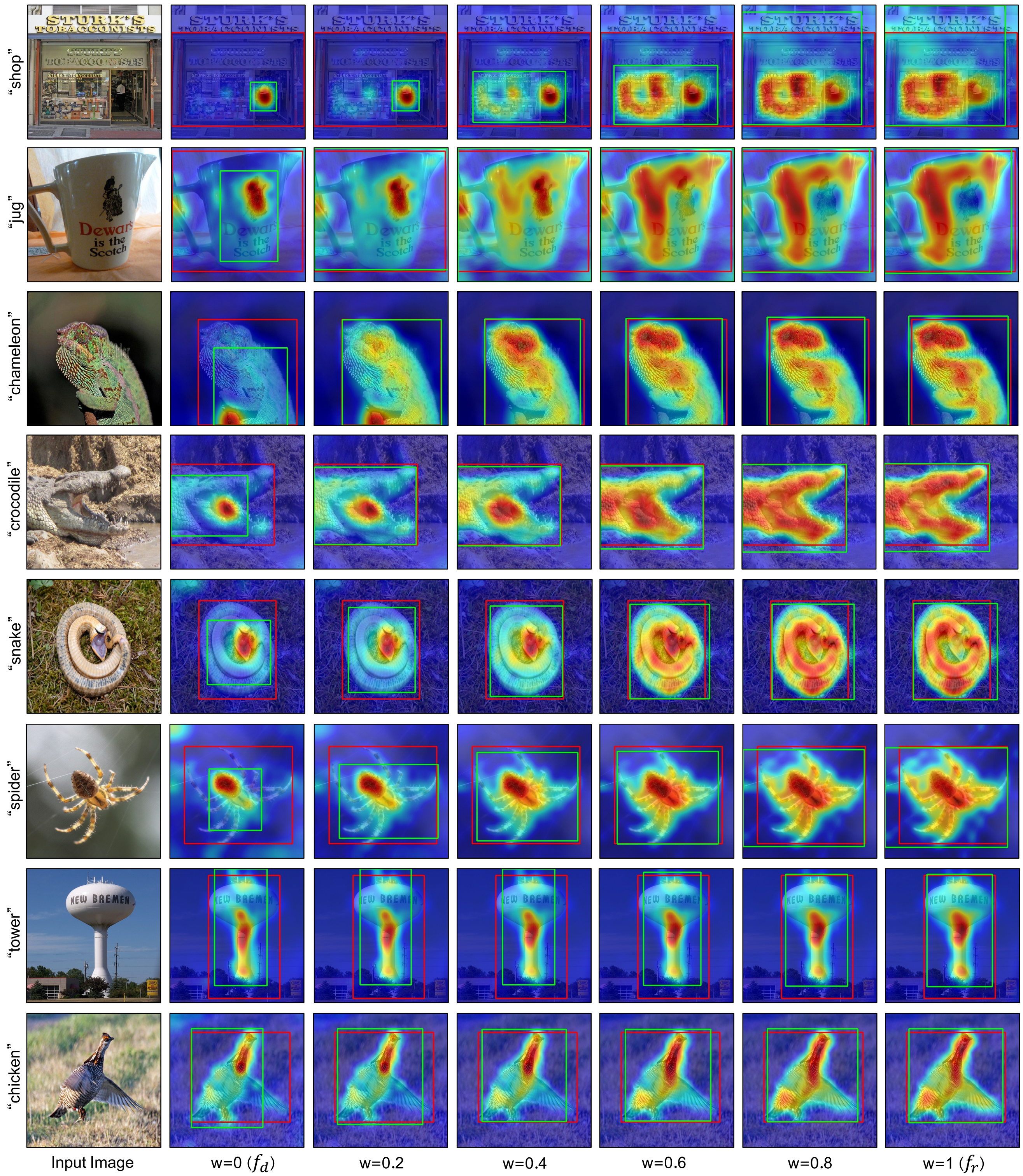

Figure 4: **Activation maps and localization results using discriminative and representative embeddings.** A proper combination of discriminative embeddings f_d with representative embedding f_r as the prompt produces precise activation maps and good WSOL results (green boxes).

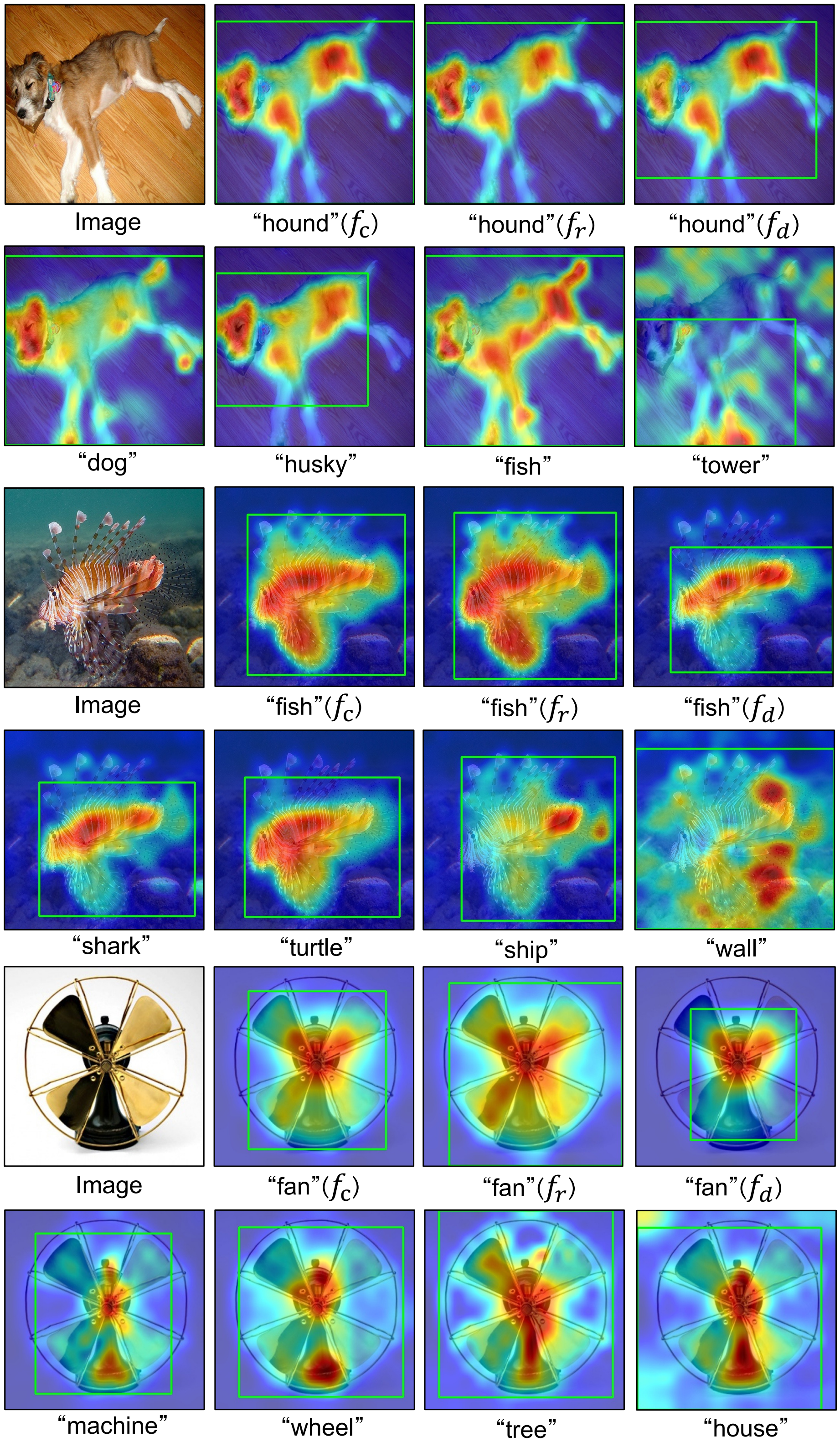

Figure 5: **Object localization results of GenPromp using different prompt words.**