VIP: Visual-guided Prompt Evolution for Efficient Dense Vision-Language Inference

Hao Zhu* [1,2] Shuo Jin* [3,4] Wenbin Liao[1,2] Jiayu Xiao[1,2] Yan Zhu [2] Siyue Yu [3] Feng Dai [1,2]

[1] Institute of Computing Technology, Chinese Academy of Sciences

[2] University of Chinese Academy of Sciences [3] XJTLU [4] University of Liverpool

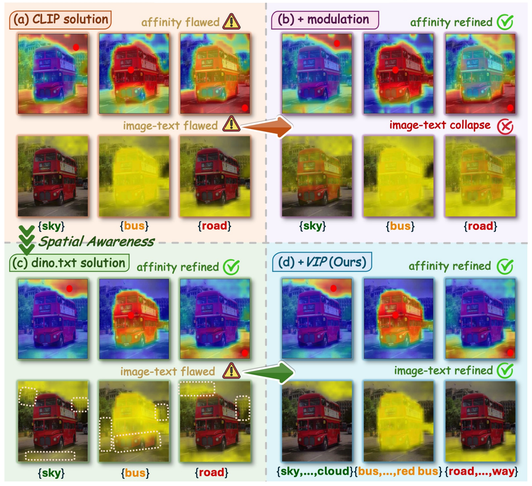
ICML International Conference On Machine Learning

Arxiv Paper

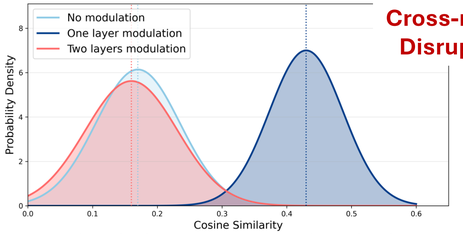
Github Page

Motivation

CLIP or DINO

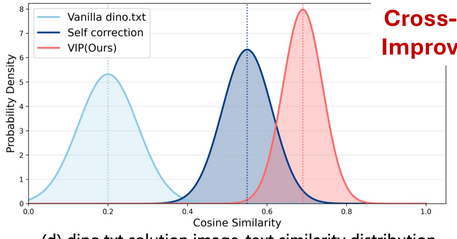


Cross-model gap



Cross-modal Disruption

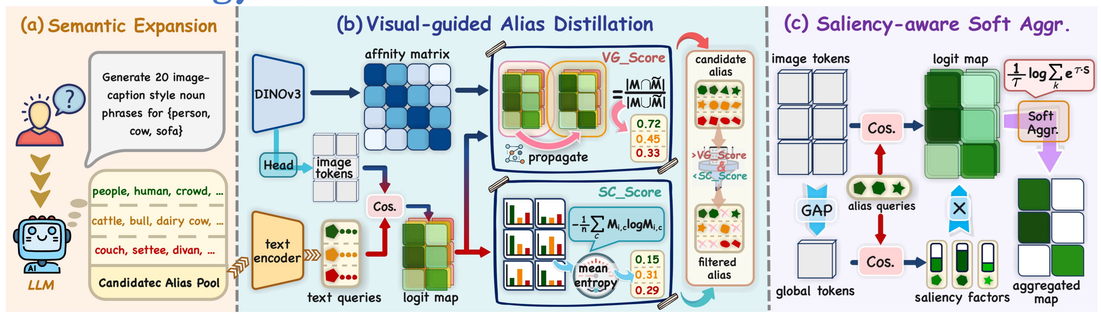
(c) CLIP-based solution image-text similarity distribution



Cross-modal Improvement

(d) dino.txt solution image-text similarity distribution

Methodology



Generation

We use LLMs to generate an alias pool to expand the semantic

How to filter the better one

The generated aliases are not always satisfactory. We hence use standard category names as anchors to filter them

Aggregation

These filtered aliases introduces multiple logits that need to be aggregated

Ablations

Methods	VOC	Object	City	ADE	Avg.	Δ
baseline	35.9	24.2	22.6	18.9	25.4	-
+ ***S-C*** §3.1	62.1	31.4	35.0	23.6	38.0	-
+ ***SE-LLM*** §3.2	61.9	28.6	36.7	23.0	37.6	-0.4
+ ***Alias Dis.*** §3.3	68.3	41.3	50.2	27.1	46.7	+8.7
+ ***Soft Aggr.*** §3.4	72.3	46.7	54.8	28.9	50.7	+12.7
+ ***Templates.*** §3.5	**73.2**	**47.3**	**55.7**	**29.1**	**51.3**	+13.3

Direct employing generated aliases can't work well without filtering out bad cases

The alias generation and filter can also work well on text templates

Detailed Text Descriptor

Methods	VOC21	Object	City	ADE	Avg.	Δ
baseline	35.9	24.2	22.6	18.9	25.4	-
+ ***S-C*** §3.1	62.1	31.4	35.0	23.6	38.0	-
+ *Category Descriptor*	58.2	27.6	32.8	21.0	34.9	-3.1
+ *Descriptor Distillation*	59.3	27.4	33.2	21.5	35.4	-2.6

Various LLM for Alias Generation

LLMs	VOC21	Object	City	ADE	Avg.
Gemini-2.5 (Comanici et al., 2025)	73.0	47.2	55.4	28.9	51.1
DeepSeek-R1 (Guo et al., 2025)	72.8	46.5	55.1	28.8	50.8
GPT-5	73.2	47.3	55.7	29.1	51.3

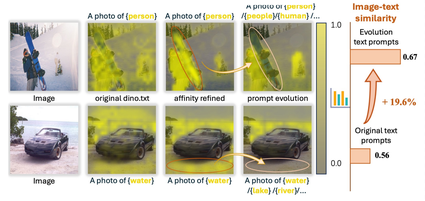
Experiments

Methods	Extra Backbone	With Background VOC21	PC60	Object	Without Background VOC20	PC59	Stuff	City	ADE	Avg.↑ (mIoU)	Time.↓ (ms)	Mem.↓ (MB)
Dense image-text inference with multi vision models												
LaVG [ECCV24]	*DINO*	61.8	31.5	33.3	81.9	34.6	22.8	25.0	14.8	38.2	140	1750
ProxyCLIP [ECCV24]	*DINO*	61.3	35.3	37.5	80.3	39.1	26.5	38.1	20.2	42.3	105	1600
LPOSS [CVPR25]	*DINO*	62.4	35.4	34.3	79.3	38.6	26.5	37.9	22.3	42.1	-	-
CASS [CVPR25]	*DINO*	65.8	36.7	37.8	87.8	40.2	26.7	39.4	20.4	44.4	440	2890
FSA [ICCV25]	*DINO*	63.7	36.1	38.0	82.3	39.9	27.0	38.8	20.5	43.3	110	1650
CLIPer [ICCV25]	*Stable Diffusion*	66.5	38.3	40.0	86.0	42.4	28.6	38.7	22.0	45.3	180	3390
Trident [ICCV25]	*DINO+SAM*	67.1	38.6	41.1	84.5	42.2	28.3	42.9	21.9	45.8	120	3710
CorrCLIP [ICCV25]	*DINO+SAM2*	74.8	44.2	43.7	88.8	48.8	31.6	49.4	26.9	51.0	1440	3200
Dense image-text inference with single vision model												
CLIP [ICML21]	✗	16.2	7.7	5.5	41.8	9.2	4.4	5.5	2.1	11.6	-	-
MaskCLIP [ECCV22]	✗	38.8	23.6	20.6	74.9	26.4	16.4	12.6	9.8	27.9	-	-
SCLIP [ECCV24]	✗	59.1	30.4	30.5	80.4	34.2	22.4	32.2	16.1	38.2	60	1580
ClearCLIP [ECCV24]	✗	51.8	32.6	33.0	80.9	35.9	23.9	30.0	16.7	38.1	45	650
CdamCLIP [ICLR25]	✗	58.7	30.6	35.2	-	-	24.8	23.7	17.2	-	100	920
ResCLIP [CVPR25]	✗	60.9	33.4	34.7	85.9	36.6	24.6	35.6	18.0	41.2	85	2550
dino.txt [CVPR25]	✗	35.9	22.9	24.2	84.5	26.1	19.2	22.6	18.9	31.8	90	1500
FreeCP [ICCV25]	✗	64.5	35.7	36.9	81.5	39.3	26.1	34.4	18.9	42.2	120	750
SFP [ICCV25]	✗	63.9	37.2	37.9	84.5	39.9	26.4	41.1	20.8	44.0	100	1600
SC-CLIP [IEEE TIP25]	✗	64.6	36.8	37.7	84.3	40.1	26.6	41.0	20.1	43.9	85	1600
VIP (Ours)	✗	**73.2**↑8.6	**41.2**↑4.0	**47.4**↑9.5	**92.5**↑6.6	**46.5**↑6.4	**33.3**↑6.7	**55.7**↑14.6	**29.1**↑8.3	**52.4**↑8.4	100	1650

Remote Sensing

Methods	iSAID	Vaihing.	Potsdam	VDD	Avg.
CLIP [ICML21]	7.5	10.3	14.5	14.2	11.6
ProxyCLIP [ECCV24]	20.7	27.8	44.1	44.3	34.2
dino.txt [CVPR25]	19.3	18.9	24.3	32.3	23.7
Trident [ICCV25]	20.0	27.7	44.4	45.7	34.5
CorrCLIP [ICCV25]	16.9	24.7	42.6	37.7	30.5
SC-CLIP [IEEE TIP25]	18.4	29.6	43.4	41.0	33.1
SegEarth-OV [CVPR25]	21.7	29.1	47.1	45.3	35.8
VIP (Ours)	**26.1**↑4.4	**47.0**↑17.4	**49.8**↑2.7	**54.3**↑8.6	**44.3**↑8.5

Text Activations



Qualitative Results