DViN: Dynamic Visual Routing Network for Weakly Supervised Referring Expression Comprehension

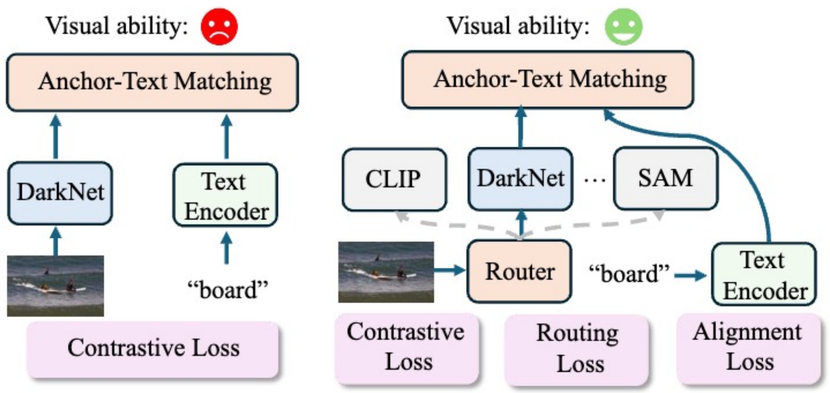
[1]Xiaofu Chen, [1]Yaxin Luo, [2]Gen Luo*, [3]Jiayi Ji, [4]Henghui Ding, [3]Yiyi Zhou

[1]MBZUAI, OpenGVLab, [2]Shanghai AI Laboratory, [3]Xiamen University, [4]Fudan University

CVPR Nashville JUNE 11-15, 2025

Paper, code, and data are available:
https://github.com/XxFChen/DViN

REC Needs More Views



Our contributions:

- Mixing visual features significantly enhances fine-grained recognition in weakly supervised REC.
- We propose DViN, which introduces a vision-conditioned router, diversity-aware routing, and routing-based feature alignment (RFA) for optimized visual representation.

Dynamic Visual Routing (DVR)

Dynamically select and combine visual features from different encoders to enhance fine-grained recognition while maintaining efficiency.

- Only two visual experts (e.g., CLIP, DINOv2) are activated per forward pass.
- The router generates routing weights based on the input visual features.

$$a^* = \arg\max_{f_{a_j} \in F_{v_j}} \varphi\left(\sum_{j \in S_e} f_{a_j} r_j, f_t\right)$$

Vision-Conditioned Routing:

$$R(F_{v_0}) = \text{softmax}\left(\sum_i F_{v_0}^i W_r + b_r\right)$$

DViN Framework

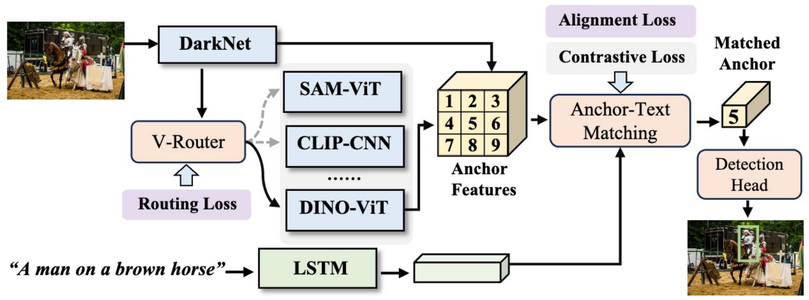


Figure 3. **Overview the proposed DViN framework.** In DViN, the vision-conditioned router will dynamically select the optimal visual encoder to refine the anchor features extracted by DarkNet. After that, anchor-text matching is conducted to locate the referent. DViN adopts the anchor-text contrastive loss for weakly supervised learning, which also includes an routing loss for efficient routing learning and an alignment loss for visual representation learning.

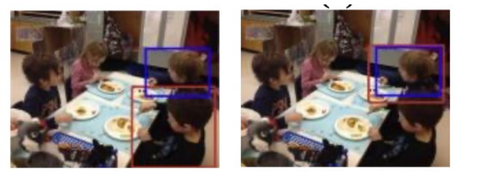
Key Results:

Method	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val-g	ReferIt test	Inference speed
GT Proposals:									
VC [36]CVPR'18	-	33.29	30.13	-	34.60	31.58	30.26	-	-
ARN [24]ICCV'19	38.05	36.43	36.47	34.53	36.40	36.12	39.62	-	-
KPRN [25]MM'19	36.34	35.28	37.72	37.16	36.06	39.29	38.37	33.87	-
DTWREG [47]TPAMI'21	39.21	41.14	37.72	39.18	40.01	38.08	43.24	-	-
EARN [27]TPAMI'22	38.08	38.25	38.59	37.54	37.58	37.92	45.33	36.86	-
Det Proposals:									
VC[36]CVPR'18	-	32.68	27.22	-	34.68	28.10	29.65	14.50	-
KAC Net [3]CVPR'18	-	-	-	-	-	-	-	15.83	-
MATN [54]CVPR'18	-	-	-	-	-	-	-	13.61	-
ARN [24]ICCV'19	32.17	35.25	30.28	32.78	34.35	32.13	33.09	26.19	5.7fps
IGN [53]NeurIPS'20	34.78	37.64	32.59	34.29	36.91	33.56	34.92	-	-
DTWREG [47]TAPAMI'21	38.35	39.51	37.01	38.91	39.91	37.09	42.54	-	5.9fps
RelR [28]CVPR'21	-	-	-	-	-	-	-	37.68	-
NCE+Dist [49]CVPR'21	-	-	-	-	-	-	-	38.39	-
RefCLIP [10]CVPR'23	60.36	58.58	57.13	40.39	40.45	38.86	47.87	39.58	31.3fps
APL [32]ECCV'24	64.51	61.91	63.57	42.70	42.84	39.80	50.22	**41.80**	26.7fps
DViN (ours)	**67.67**	**70.90**	**59.39**	**52.54**	**57.52**	**45.31**	**55.04**	40.63	29.8fps
Pesudo Labels:									
RefCLIP_SimREC [10]	62.57	62.70	61.22	39.13	40.81	36.59	45.68	42.33	54.8fps
RefCLIP_Transvg [10]	64.08	63.67	63.93	39.32	39.54	36.29	45.70	42.64	19.3fps
DViN_SimREC (ours)	**67.29**	**73.09**	60.65	**51.54**	**59.06**	**39.59**	**51.73**	**45.43**	38.4fps
DViN_Transvg (ours)	64.99	68.87	**64.48**	50.72	57.36	38.64	50.47	44.26	19.3fps

Ablation studies:

Method	RefCOCO test A	RefCOCO test B	RefCOCO+ test A	RefCOCO+ test B	Inference Speed
RefCLIP	58.75	57.01	39.99	39.33	31.3 fps
+ V-Router	68.68	57.03	56.84	43.50	29.8 fps
+ R-loss	69.11	59.01	56.96	43.85	29.8 fps
+ T-loss	70.30	59.01	57.19	44.24	29.8 fps
+ V-loss	70.90	59.39	57.52	45.31	29.8 fps

Visualization Results:

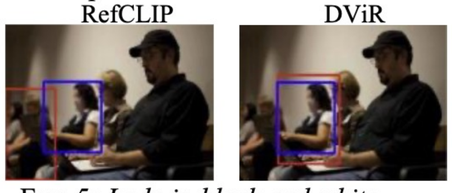


Exp-1: *Kid on right in back blondish hair*
RefCLIP DViR

Exp-2: *Glass closer to italian*

Exp-4: *Man inforeground in dark shirt*

RefCLIP DViR



Exp-5: *Lady in black and white slightly blurry*

Exp-6: *Yellow and blue vehicle close to the camera*