Dual-Stream Attention with Multi-Modal Queries for Object Detection in Transportation Applications

Noreen Anwar[1], Guillaume-Alexandre Bilodeau[1], Wassim Bouachir[2]
[1]LITIV Lab, Polytechnique Montréal [2]Data Science Laboratory, TÉLUQ University

BMVC 2025

Abstract

We introduce DAMM, a detection transformer using multi-modal queries (appearance, polygonal positional, learnable) and a dual-stream attention decoder with adaptive per-layer fusion. A polygon-aware head improves localization of irregular and small objects, increasing robustness to occlusion and motion blur. DAMM sets a new state of the art on Cityscapes with **38.5 AP** (*vs. 35.5* for Grounding DINO (5)) and also surpasses prior methods on UAVDT/UA-DETRAC with **32.5/32.8 AP** (*vs. 27.0/27.3* Grounding DINO; *28.3/28.8* OV-DETR (7)).

Introduction

Traffic and aerial imagery pose challenges for object detection: occlusion, blur, irregular small objects, and limited spatial grounding. **DAMM** addresses these via a dual-stream attention decoder and multi-modal queries (appearance from Grounding DINO, positional from SAM, learnable).

Contributions

- **Multi-modal queries**: fuse appearance (Grounding DINO), positional priors (SAM), and learnable tokens.
- **Dual-stream cross-attention**: separate semantic and spatial decoder streams.
- **Adaptive per-layer fusion**: learnable gating merges outputs per layer.
- **Polygonal embeddings**: improve localization of irregular and small objects.

Methods

Overview: DAMM extends deformable-DETR (1, 8) with three query streams: appearance (Grounding DINO (5)), polygonal-positional (SAM (4)), and learnable random queries.
Pipeline (see Figure 1):

- Extract multi-scale encoder features.
- Build query banks $\{Q_{\text{app}}, Q_{\text{pos}}, Q_{\text{ran}}\}$.
- Dual-stream decoder (semantic & spatial).
- Adaptive fusion + classification + polygon-aware localization.

Training and Design Notes: Uses Hungarian matching (1), L1 + GIoU + polygon-alignment losses. Deformable attention ensures efficiency (8); Grounding DINO/SAM priors enhance robustness.

Evaluation

Datasets: Cityscapes (2), UAVDT (3), UA-DETRAC (6).
Baselines: Deformable DETR, DINO-Deform, Grounding DINO, OV-DETR (7).

Results: Cityscapes & VisDrone

Method	Cityscapes			VisDrone		
	AP	AP50	AP75	AP	AP50	AP75
Deformable DETR	22.5	41.7	22.5	27.2	44.1	28.5
DINO-Deform.	33.0	56.0	38.0	31.1	49.5	28.2
Grounding DINO	35.5	58.7	39.6	34.9	59.3	40.1
OV-DETR	31.5	54.3	36.2	32.0	50.0	30.0
DAMM (Ours)	**38.5**	**62.5**	**47.5**	**39.5**	**63.1**	**42.3**

Results: UAVDT & UA-DETRAC

Method	UAVDT			UA-DETRAC		
	AP	AP50	AP75	AP	AP50	AP75
Grounding DINO (5)	27.0	40.5	27.2	27.3	41.2	27.2
OV-DETR (7)	28.3	35.1	22.4	28.8	37.3	23.9

Framework Overview

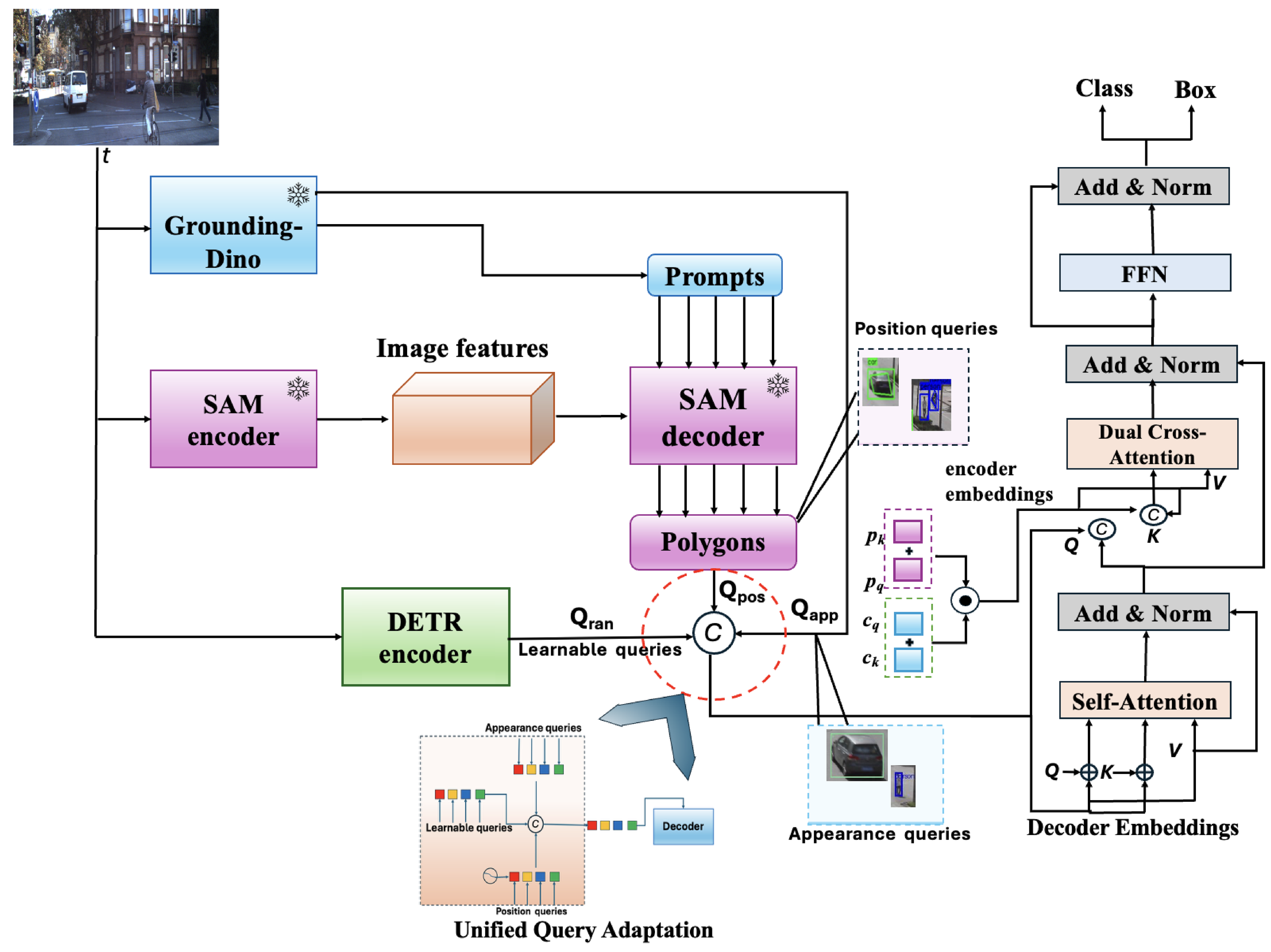

Figure 1: DAMM architecture: multi-modal queries feed a dual-stream decoder with adaptive fusion and polygon-aware heads.

Ablation: Query Types (Cityscapes, Table 6)

Setting	G-DINO (app.)	SAM (pos.)	RQ (learn.)	AP
RQ only			✓	34.1
G-DINO + RQ	✓		✓	36.8
SAM + RQ		✓	✓	37.2
All (G-DINO + SAM + RQ)	✓	✓	✓	**38.5**

Conclusion

Design: Multi-modal queries + dual-stream fusion of appearance/position.
Effect: Polygon-aware heads improve localization for small/irregular and occluded objects.
Result: Consistent gains over strong baselines on Cityscapes, UAVDT, UA-DETRAC, VisDrone.
Code: `https://github.com/DET-LIP/DAMM`

References

(1) Nicolas Carion et al. "End-to-End Object Detection with Transformers". In: *ECCV*. 2020.

(2) Marius Cordts, Mohamed Omran, et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *CVPR*. 2016.

(3) Dawei Du et al. "UAVDT: A Large-Scale Benchmark for UAV Tracking". In: *ECCV Workshops*. 2018.

(4) Alexander Kirillov, Eric Mintun, et al. "Segment Anything". In: *arXiv:2304.02643* (2023).

(5) Shilong Liu et al. "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection". In: *CVPR Workshops*. arXiv:2303.05499. 2023.

(6) Longyin Wen et al. "UA-DETRAC: A New Benchmark and Protocol for Multi-Object Detection and Tracking". In: *Computer Vision and Image Understanding* (2020).

(7) Yuhang Zang et al. "OV-DETR: Open-Vocabulary DETR". In: *arXiv:2209.08788* (2023).

(8) Xizhou Zhu et al. "Deformable DETR: Deformable Transformers for End-to-End Object Detection". In: *ICLR*. 2021.

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

NSERC CRSNG