# Neptune-X: Active X-to-Maritime Generation for Universal Maritime Object Detection

Yu Guo[1,3], Shengfeng He[2,*], Yuxu Lu[4], Haonan An[1], Yihang Tao[1], Huilin Zhu[5],
Jingxian Liu[3,*], Yuguang Fang[1]
[1]City University of Hong Kong, [2]Singapore Management University, [3]Wuhan University of Technology,
[4]The Hong Kong Polytechnic University, [5]Wuhan University of Science and Technology

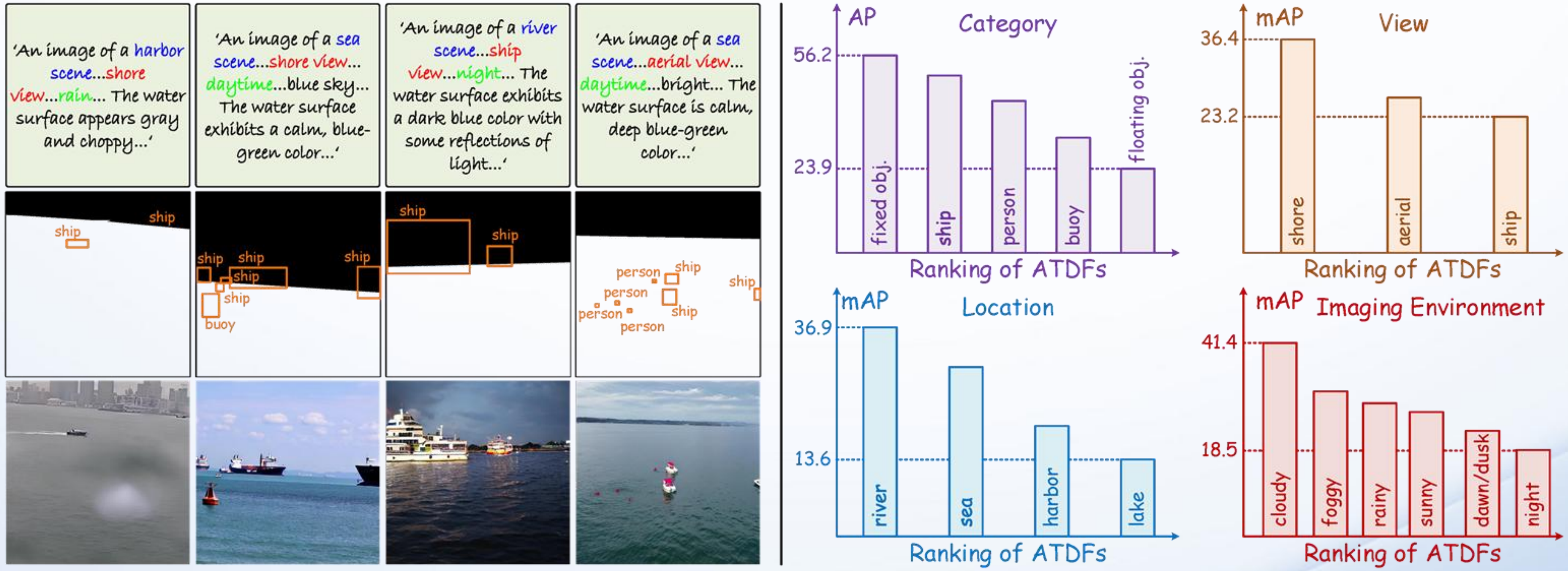
NEURAL INFORMATION PROCESSING SYSTEMS

# Content

# 01 Introduction

# 01 Introduction

- Maritime object detection is essential for maritime applications but relies on **scarce and costly annotated data**.
- Models generalize poorly due to **inherent imbalances in existing maritime datasets** across conditions and object categories.




(a) X-to-Maritime Generation Results

(b) Correlations between ATDFs and Detection Accuracy on Test Set

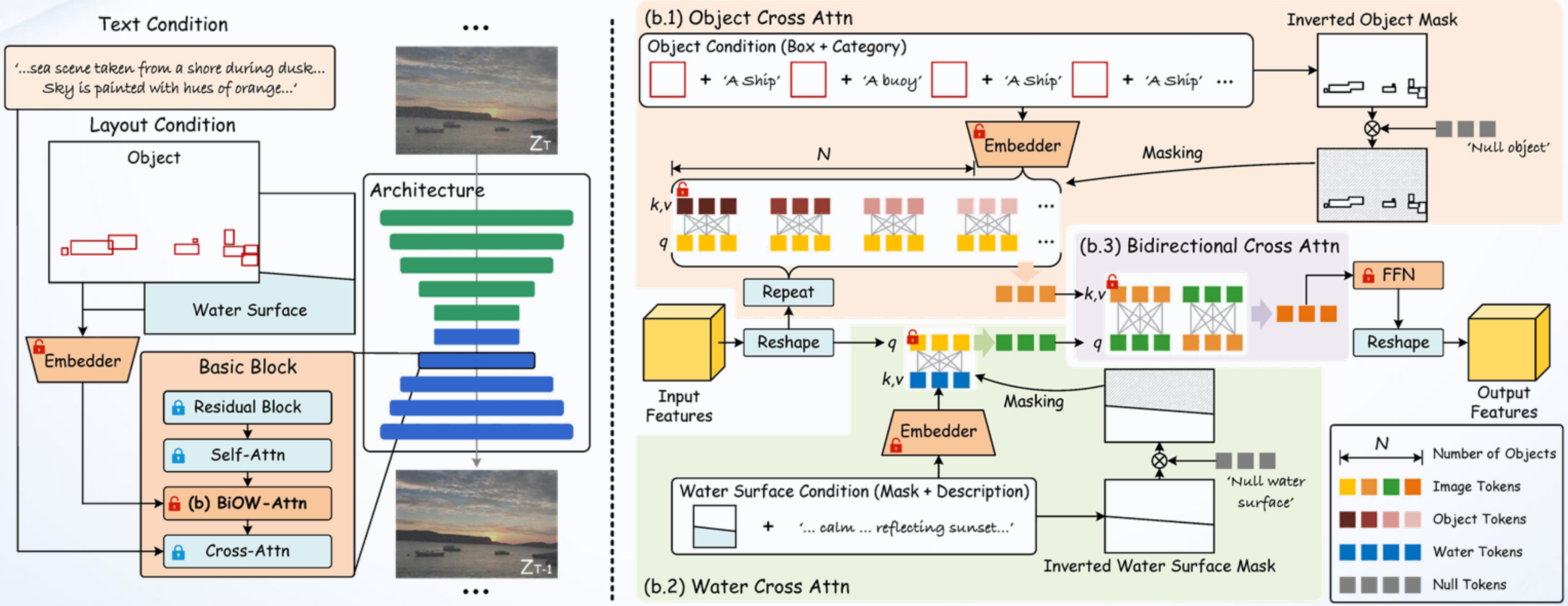
# 02 Neptune-X: Active X-to-Maritime Generation

# 02 Neptune-X

## X-to-Maritime Generation

## Objective Function →

$$\mathcal{L} = \mathbb{E}_{z,t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\epsilon - g_\theta(z_t, t, \mathcal{C}, \underbrace{\{\mathcal{C}_o^i, \mathcal{M}_o^i\}_{i=1}^{O}}_{\text{object conditions}}, \underbrace{\{\mathcal{C}_w, \mathcal{M}_w\}}_{\text{water surface condition}})\|_2^2\right],$$




(a) Data Generation Pipeline

(b) Bidirectional Object-Water Attention

**Bidirectional Object-Water Attention** (BiOW-Attn) module, which explicitly models the interactions between objects and their aquatic surroundings to generate physically plausible and realistic maritime scenes.

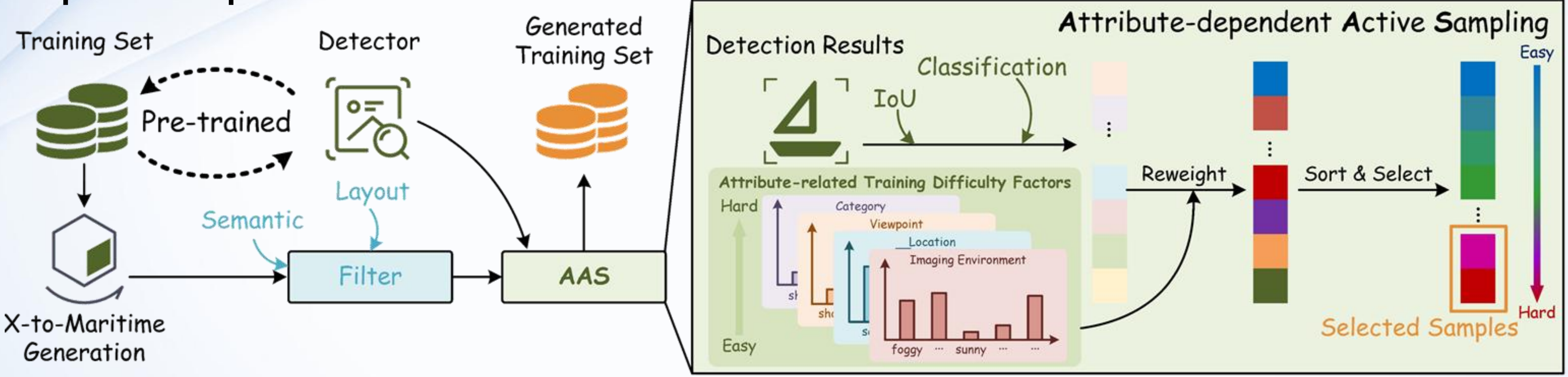
## Conditional Injection

$$\text{Cross-Att}(Q, K_k, V_k) = \text{Softmax}\left(\frac{Q \cdot K_k^\top}{\lambda}\right) V_k, \quad k \in \{\mathcal{C}_o^i, \mathcal{C}_w\},$$

$$\mathbf{F}_o = \left(\sum_{i=1}^{O} f_o^i\right) \odot \mathbf{M} + \mathbf{null}_{\text{obj}} \odot (1 - \mathbf{M}), \quad \text{where } \mathbf{M} = \bigcup_{i=1}^{O} \mathcal{M}_o^i.$$

# 02 Neptune-X

## Neptune-X Pipeline




### Attribute-correlated Training Difficulty Factors

$$d_s^j = \frac{1}{N_s^j}\sum_{n=1}^{N_s^j}(1-\mathrm{Acc}_n).$$

**Exponential Moving Average**

$$d_s^j \leftarrow m_s^{j-1} d_s^{j-1} + (1-m_s^{j-1}) d_s^j,$$

### Comprehensive Consideration

Category Viewpoint Location Imaging Environment

**Training Difficulty**

$$d = \delta \prod_{\alpha\in A} d_\alpha \cdot \frac{1}{N}\sum_{n=1}^{N} d_{\mathrm{cls}}^n \cdot (1-\mathrm{Acc}_n),$$

# 03 Experimental Evaluation

# 03 Experimental Evaluation

## Maritime Generation Dataset

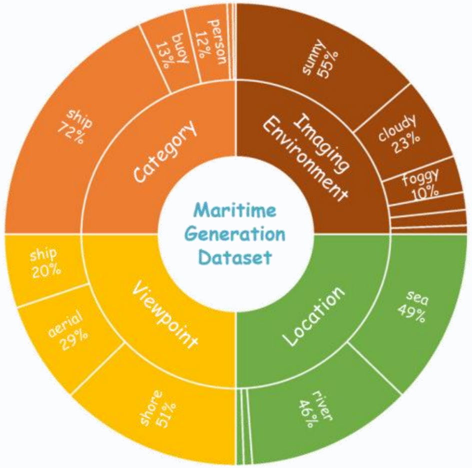


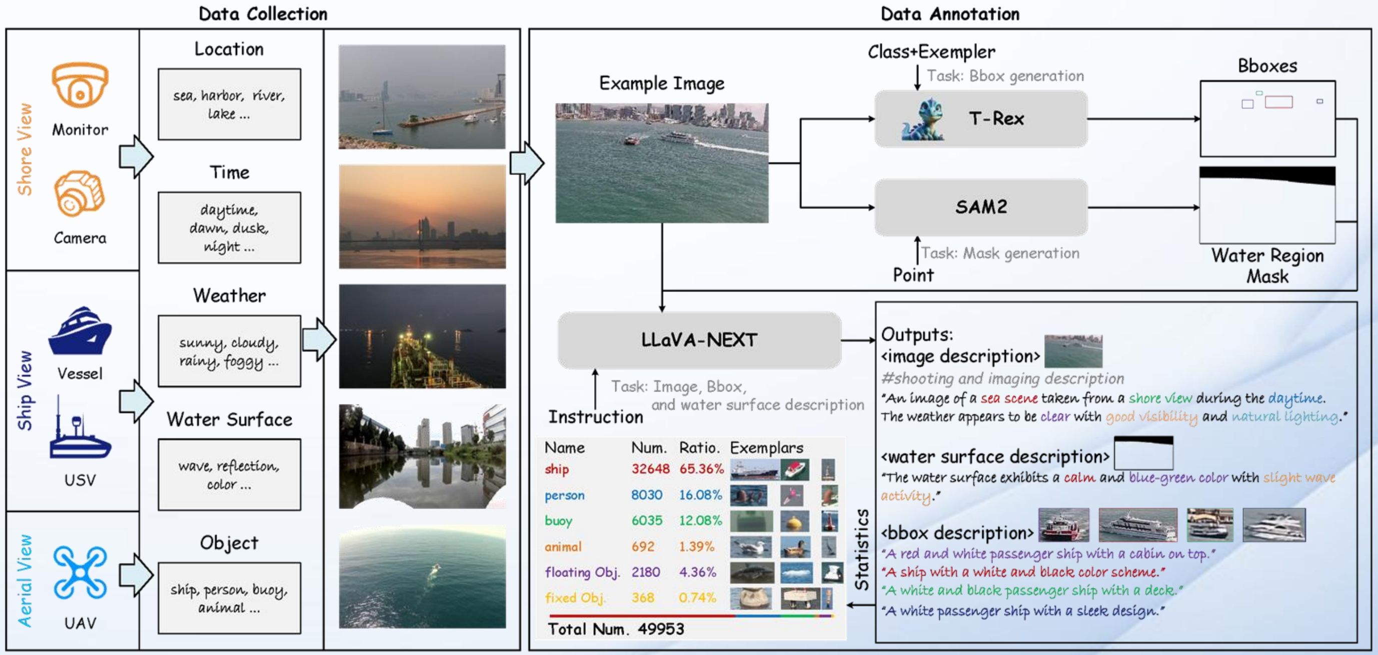

Figure 7: The percentages of various dimensions and attributes in our MGD dataset.

Table 7: Sample numbers and percentages of various dimensions and attributes.

| Dimensions | Attributes | Number | Proportion |
|---|---|---|---|
| Category | ship | 29313 | 72.44% |
| | buoy | 5326 | 13.16% |
| | person | 4843 | 11.97% |
| | floating obj. | 618 | 1.53% |
| | fixed obj. | 366 | 0.90% |
| View | shore | 6042 | 50.77% |
| | ship | 2459 | 20.66% |
| | aerial | 3399 | 28.56% |
| Location | sea | 5829 | 48.98% |
| | river | 5531 | 46.48% |
| | harbor | 282 | 2.37% |
| | lake | 258 | 2.17% |
| Imaging Environment | sunny | 6491 | 54.55% |
| | cloudy | 2794 | 23.48% |
| | foggy | 1225 | 10.29% |
| | rainy | 515 | 4.33% |
| | dawn/dusk | 583 | 4.90% |
| | night | 292 | 2.45% |

Table 1: Data source of MGD.

| Source | Imaging Viewpoint | Num. |
|---|---|---|
| MaSTr1325 [3] | ship view | 800 |
| USVInland [6] | ship view | 1000 |
| MIT Sea Grant [9] | ship view | 100 |
| SMD [24] | shore and ship view | 400 |
| Seaships [33] | shore view | 1500 |
| Seagull [29] | aerial view | 2996 |
| Fvessel [12] | shore view | 1500 |
| LaRS [52] | shore, ship, and aerial view | 1973 |
| Others | shore, ship, and aerial view | 1631 |
| MGD | shore, ship, and aerial view | 11900 |



# 03 Experimental Evaluation

## Data Generation Results

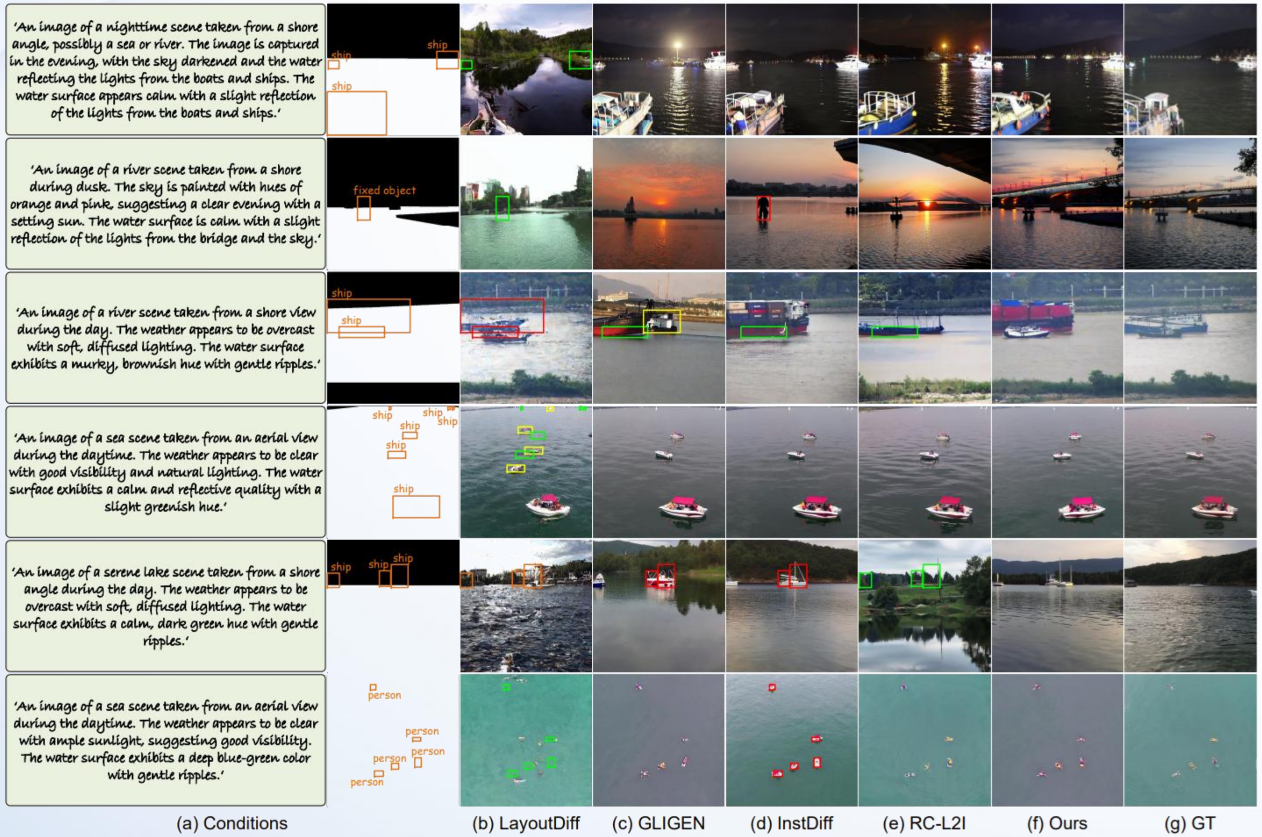



Figure 4: Comparison of image generation on MGD. The red, green, and yellow bounding boxes indicate low-quality/incorrect generation, missed generation, and unexpected generation, respectively.

Table 2: FID, CAS, and YOLO Score comparisons of different methods on image generation. The best and second-best results are highlighted in **bold** and underlined.

| Methods | Conditions | Venue & Year | FID ↓ | CAS ↑ | YOLO Score ↑ mAP/mAP$_{50}$/mAP$_{75}$ |
|---|---|---|---|---|---|
| SD1.5 [30] | Text | CVPR2022 | 27.65 | – | – |
| LayoutDiff [50] | Box | CVPR2023 | 18.17 | 63.77 | 0.83/2.68/0.29 |
| GLIGEN [18] | Text + Box | CVPR2023 | 20.02 | 77.06 | 12.74/30.36/8.99 |
| InstDiff [38] | Text + Box + Mask | CVPR2024 | 19.43 | 76.65 | 12.46/29.73/9.07 |
| RC-L2I [5] | Text + Box + Mask | NeurIPS2024 | 25.63 | 74.84 | 8.75/22.99/5.48 |
| Ours | Text + Box + Mask | | **18.05** | **79.34** | **17.08/39.14/13.52** |

Table 5: Ablation study of different generation configurations.

| ObjCA | WatCA | BiCA Obj2WatCA | BiCA Wat2ObjCA | FID ↓ | CAS ↑ | YOLO Score ↑ mAP/mAP$_{50}$/mAP$_{75}$ |
|---|---|---|---|---|---|---|
| ✓ | | | | 21.44 | 76.23 | 10.69/26.01/6.99 |
| ✓ | ✓ | | | 19.57 | 78.15 | 13.37/29.60/10.78 |
| ✓ | ✓ | ✓ | | 18.35 | 78.00 | 12.52/27.58/10.06 |
| ✓ | ✓ | | ✓ | 18.37 | 78.68 | 15.60/36.13/12.09 |
| ✓ | ✓ | ✓ | ✓ | **18.05** | **79.34** | **17.08/39.14/13.52** |

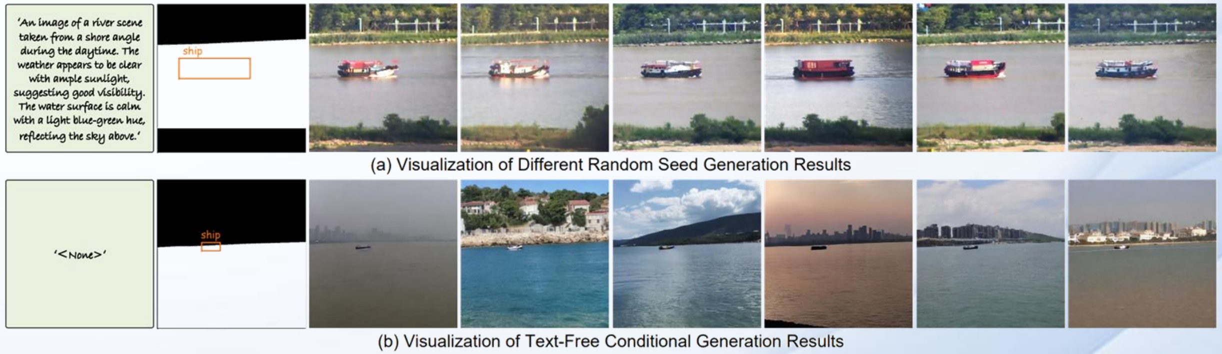



Figure 11: Image generation cases using (a) different random seeds and (b) only removing text conditions. The main reason for the scene similarity in (a) is that the text specifies background and hydrological conditions, while the unspecified objects exhibit diversity.

## Data Augmentation Results

Table 3: mAP and $mAP_{50}$ comparison with/without generated data.

| Model | mAP ↑ | $mAP_{50}$ ↑ |
|---|---|---|
| YOLOv10 [37] | 39.99 | 61.13 |
| +*Gen Data* | **43.62 (+9.08%)** | **65.50 (+7.15%)** |
| YOLOv11 [16] | 41.29 | 62.51 |
| +*Gen Data* | **44.43 (+7.60%)** | **66.15 (+5.82%)** |
| YOLOv12 [35] | 39.06 | 60.53 |
| +*Gen Data* | **42.91 (+9.86%)** | **63.85 (+5.48%)** |

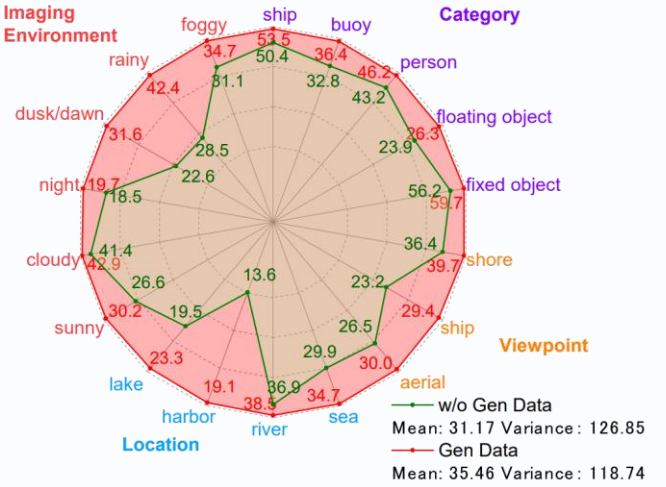



Figure 5: YOLOv11 accuracy improvement visualization across various attributes.

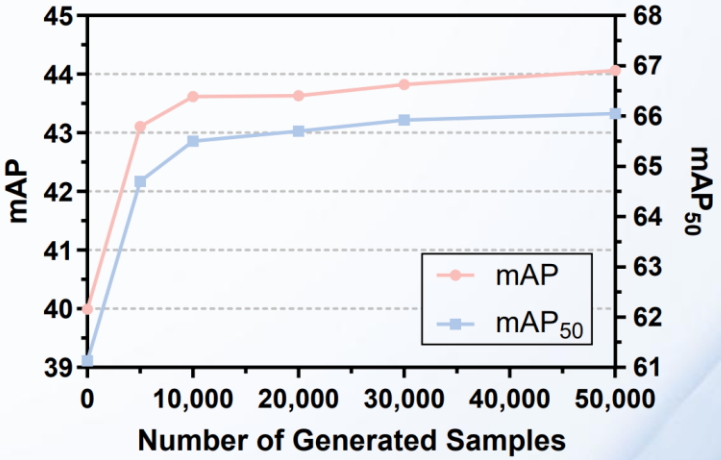



Figure 6: Correlation between detection accuracy and the number of generated samples used.

Table 4: mAP and $mAP_{50}$ comparison with/without generated data. † denotes fine-tuned on our dataset.

| Model | mAP ↑ | $mAP_{50}$ ↑ |
|---|---|---|
| Grounding DINO | 8.42 | 12.60 |
| Grounding DINO† | 65.03 | 86.12 |
| +*Gen Data* | **68.04 (+4.63%)** | **89.86 (+4.34%)** |

Table 6: Ablation study of different sampling strategies.

| Methods | Number | mAP ↑ | $mAP_{50}$ ↑ |
|---|---|---|---|
| N/O | 0 | 39.99 | 61.13 |
| Random | 5,000 | 41.48 | 63.19 |
| | 10,000 | 43.31 | 64.95 |
| AAS | 5,000 | 43.11 | 64.70 |
| | 10,000 | **43.62** | **65.50** |

NEURAL INFORMATION PROCESSING SYSTEMS

# 04 Conclusion

# 04 Conclusion

NEURAL INFORMATION PROCESSING SYSTEMS

## Contribution of our Work

- **Generative Framework:** We introduce X-to-Maritime, a novel framework incorporating a Bidirectional Object-Water Attention module to generate realistic maritime scenes under multi-condition inputs.
- **Sampling Strategy:** We propose an Attribute-dependent Active Sampling approach that dynamically estimates training difficulty across semantic dimensions to select high-value synthetic samples.
- **Benchmark Dataset:** We construct the Maritime Generation Dataset (MGD), the first dedicated benchmark for generative maritime learning, featuring comprehensive annotations and diverse scenarios

## Code (Github)

# Neptune-X: Active X-to-Maritime Generation for Universal Maritime Object Detection

Yu Guo[1,3], Shengfeng He[2,*], Yuxu Lu[4], Haonan An[1], Yihang Tao[1], Huilin Zhu[5], Jingxian Liu[3,*], Yuguang Fang[1]

[1]City University of Hong Kong, [2]Singapore Management University, [3]Wuhan University of Technology, [4]The Hong Kong Polytechnic University, [5]Wuhan University of Science and Technology