P^{2}UIC: Plug-and-Play Physics-Aware Contrastive Mamba Framework for Underwater Image Captioning

Chunlei Wang, Wenquan Feng, Binghao Liu, Xianyu Zhao, Kejun Zhao, Qi Zhao, *Member, IEEE*

***Abstract*—Underwater Image Captioning (UIC) aims to automatically generate accurate descriptive textual for input underwater images, which faces unique challenges due to environmental distortions such as degraded image quality, ambiguous object categorization, and computational resource limitations. Existing Image Captioning (IC) models, primarily designed for natural scenes, often underperform in underwater environments because they fail to address these domain-specific issues. Recently, the Mamba framework with its linear reasoning capabilities and efficient performance has emerged as a popular alternative to Transformer framework. In this paper, we propose P^{2}UIC, a plug-and-play physics-aware contrastive Mamba framework for underwater image captioning consisting of Underwater Physics Environment-aware Enhancement (UPEE) and Contrastive Multi-sequence Mamba Decoder (CMMD). UPEE enables plug-and-play underwater perception enhancement for any IC model by inferring physical parameters and incorporating environment-specific keywords. Meanwhile, CMMD leverages semantic information from multi-sequence visual features to improve text generation via contrastive learning across multiple text embedding layers. To address the scarcity of evaluation benchmarks, we release two UIC benchmarks by extending existing underwater segmentation datasets and provide expert-annotated descriptions. Our P^{2}UIC effectively handles UIC tasks and achieves state-of-the-art (SOTA) results. Ablation studies and visualization experiments demonstrate the effectiveness of the proposed components. The code and models will be publicly available at https://github.com/cv516Buaa/ChunleiWang/P2UIC.**

***Index Terms*—Underwater image captioning models, Underwater image captioning datasets, Mamba framwork, Plug-and-play models.**

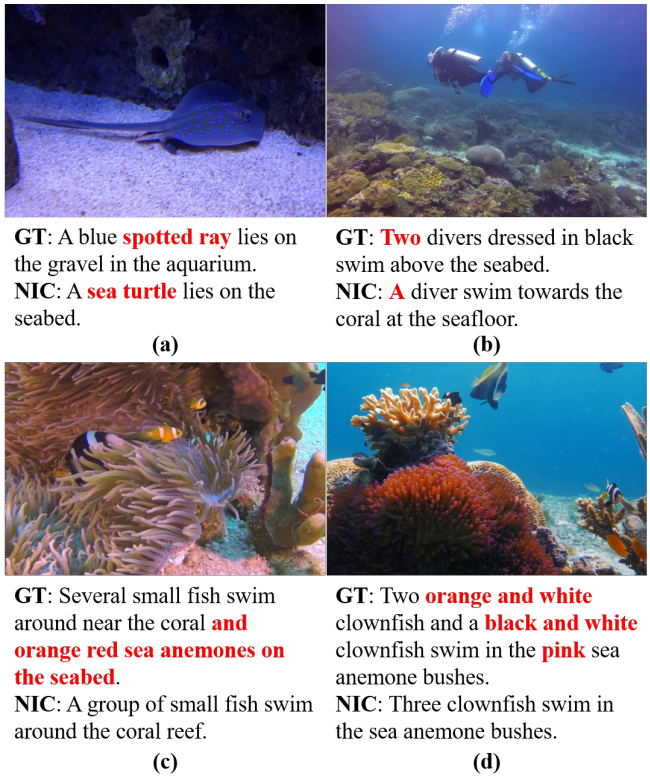

Fig. 1. Weaknesses illustration of the existing NIC approaches in underwater environments. (a) Wrong category. (b) False Negative. (c) Lack of detailed environment description. (d) Insufficient fine-grained information.

I. Introduction

UNDERWATER computer vision plays a crucial role in ocean energy exploitation and underwater life monitoring, with remarkable progress in fields such as underwater image enhancement [1], [2], underwater image restoration [3], [4], underwater image classification [5], and underwater image segmentation [6], [7]. Although existing underwater research has significantly improved our ability to observe underwater environments, underwater multimodal tasks that integrate visual perception with semantic understanding remain long-term research goals of underwater exploration [8]. Image captioning aims to describe visual content using natural language, serves as a cornerstone for multimodal computer vision tasks. By facilitating the understanding of visual-language interactions and bridging the semantic gap, IC holds significant potential in practical applications, including but not limited to robot navigation [9] and medical image analysis [10].

Natural Image Captioning (NIC) methods based on encoder-decoder architectures have achieved significant advancements. Previous studies typically employ Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) as image feature extractors and text generators, respectively [11]. Recently, the Transformer architecture has become the mainstream approach for cross-modal tasks, by leveraging self-attention mechanisms to enhance the extraction of global contextual information from images and improve image-text interaction, including image captioning and visual grounding. However, Transformer-based models demand substantial computational and storage resources. Due to the resource constraints in underwater envi-

This work was supported in part by the National Natural Science Foundation of China under Grants 62072021. (*Corresponding author:Qi Zhao*)

Chunlei Wang, Wenquan Feng, Binghao Liu, Qi Zhao are with the Department of Electronics and Information Engineering, Beihang University, Beijing 100191, China (e-mail:{wcl_buaa; buaafwq; liubinghao; zhaoqi}@buaa.edu.cn).

Xianyu Zhao is with Changchun University of Science and Technology, Changchun 130000, China and national key laboratory of air-based information of Luoyang Institute of Electro-Optical Equipment, Luoyang 471000, China (e-mail: zhao_xianyu@foxmail.com).

Kejun Zhao is with Luoyang Institute of Electro-Optical Equipment, AVIC, Luoyang 471000, China (e-mail: 58494573@qq.com).

ronments and domain-specific data disparities, natural scenes methods cannot be directly applied to UIC [12].

Nowadays, Spatial State Models (SSMs) also known as Mamba, have garnered significant attention due to their long-sequence modeling capabilities, linear computational complexity, and parallel computing efficiency. To better process sequential data, researchers have proposed various SSM variants, such as linear space state layer [13] and diagonal state space [14]. Mamba integrates time-varying parameters with selective SSM structures to achieve efficient training and inference. However, the Mamba application in visual tasks still faces challenges in unidirectional sequence modeling and lack of position awareness. Vim [15] addresses bidirectional SSM for visual context modeling and positional-aware encoding, enhancing robustness in dense prediction tasks. Vision Mamba [16] has demonstrated comparable performance to Transformer in both visual tasks and long text generation tasks, but has limited exploration in multimodal tasks. To leverage the strengths of SSM and Transformer for contextual features extraction with lower resource consumption, Jamba [17] interleaves Transformer and Mamba layers, balancing model capacity and parameter controllability. Pilault et al. [18] integrate SSM and Transformer block attention, optimizing runtime efficiency through parallel processing.

Existing IC models rely heavily on natural scene datasets, limiting their applicability to underwater environments. To advance the UIC task, By providing five natural language captions for each image, this paper manually annotate images from two existing underwater segmentation datasets, UWS [6] and SUIM [19], and construct two new underwater image captioning datasets: UWS-IC and SUIM-IC. Underwater lighting conditions is affected by scattering and refraction, result in images with low contrast, color cast, high object-background similarity, and blurred boundaries. Due to the difficulty in discerning subtle underwater target features, existing models often exhibit issues such as (a) misclassification, (b) false-negative predictions, (c) generated text descriptions frequently lack detailed environmental context and (d) in sufficient fine-grained target-specific information, as shown in Fig. 1. These challenges highlight that NIC models cannot be directly applied to underwater imagery.

To mitigate image domain discrepancy, we design a plug-and-play Underwater Physics Environment-aware Enhancement (UPEE) module. By modeling underwater physical parameters (e.g., light scattering, illumination attenuation) and incorporating environment-specific keywords (e.g., "turbidity," "blue wavelength"), UPEE enables domain-adaptive feature selection for any image captioning model, enhancing fine-grained feature extraction in degraded underwater visuals. Additionally, we develop a Contrastive Multi-sequence Mamba Decoder (CMMD), leverages semantic information from multi-scale visual features and improves the text generation capability through contrastive learning of multiple text embedding layers, thereby enhancing cross-modal alignment for underwater scene descriptions. The main contributions of this paper can be summarized as follows:

- We release two underwater image captioning benchmark datasets: UWS-IC and SUIM-IC. Then we design P^{2}UIC, a plug-and-play physics-aware contrastive Mamba framework for underwater image captioning.
- We propose Underwater Physics Environment-aware Enhancement (UPEE), which enables plug-and-play underwater perception enhancement for any NIC model.
- We design Contrastive Multi-sequence Mamba Decoder (CMMD), which exploits semantic information from visual features and improves text generation capability through contrastive learning multiple textual embedding.

II. Related Work

In this section, we will review the rateled works on Visual State Space Models, Natural Image Captioning and Underwater Image Captioning.

A. Visual State Space Models

Although widely used in vision tasks, Transformer architectures face challenges such as high computational complexity of attention and difficultly in processing long-sequence inputs, SSMs [20] have emerged as a promising alternative to Transformers. Gu et al. [21], [22] propose a Structured State Space Sequence (S4) that improves computational efficiency through the HiPPO matrix and designs a selective SSM [16] to address the linear time-invariant limitations of traditional SSMs. Through the integrated design of Mamba structure [23]–[26], this method achieves remarkable performance in natural language processing (NLP) tasks.

Recently, the application of Mamba in visual tasks has gradually become a research hotspot [27]–[29]. ViS4mer [30] reduces the spatiotemporal feature resolution and channel dimensions via 1-D multi-scale temporal S4 decoder, learning complex long-range spatiotemporal dependencies for video classification. S4ND [31] further extends 1-D S4 to multidimensional domains, enhancing the generalization capability of SSMs. DiffuSSM [32] replaces Transformer backbones with scalable state space models, enabling efficient generation of high-resolution images while preserving fine-grained representations. UMamba [28] introduces a hybrid CNN-SSM module for medical image segmentation, combining the local feature extraction ability of convolutional layers with the ability of SSM to capture long-range dependencies. Localmamba [33] partitions image into local windows to model fine-grained dependencies while maintaining global context awareness. VMamba [34] bridges the structural gap between 1-D sequential scanning and 2-D visual data through Visual State Space (VSS) block stacking, improving the performance of mamba in visual perception tasks. VisionMamba [15] represents multimodal features by embedding positional tags into image sequences and compressing visual features via bidirectional state-space modeling.

B. Natural Image Captioning

Previous works on NIC predominantly relied on RNN-LSTM based encoder-decoder architectures and improve the image caption generation ability through various attention methods [35]–[37]. You et al. [38] proposes a semantic attention model by selectively focusing on semantic concept

proposals and fusing them into the hidden states and outputs of a RNN. AoANet [39] generates a context information vector using the attention mechanism and further obtains the image-textual information through element-by-element multiplication. However, RNN and LSTM-based image captioning methods tend to ignore visual context information, resulting in the repetition of high-frequency phrases, which limits the expression of caption generation [40].

With the remarkable success of Transformer in NLP, Transformer-based image captioning models achieve superior performance over LSTM-based architectures via multi-head attention in encoding and decoding [41], [42]. M3ixup [43] implements data augmentation by mixing IC samples in terms of visual features, sentence embeddings, and loss functions, compelling models to prioritize visual-semantic alignment during captioning and mitigating dataset bias. CaDRel [44] employs inter-model interactive learning and a cascade diffusion mechanism guided by salient semantic evaluation to realize knowledge distillation for image captioning. MMGAT [45] uses multi-modal graph aggregation to improve the extraction of regional features and contextual information by representing images as three sub-graphs: context grid, region, and semantic text modality. I2OA [46] enhances IC capabilities through intra-head and inter-head orthogonal attention. Specifically, the intra-head attention enhances the multi-head attention learning of by introducing orthogonal constraints to each head, and the inter-head orthogonal attention mechanism reduces head redundancy and expands the diversity of representation subspaces by applying orthogonal constraints between heads. In addition, image captioning models leveraging Contrastive Language Image Pre-training (CLIP) to extract rich semantic information, enabling the generated captions to better align with human grammatical structures [47], [48].

C. Underwater Image Captioning

Underwater image classification and segmentation tasks have been extensively studied using multiple datasets [6], [19], [49]–[51]. Meanwhile, several image captioning studies have focused on natural and remote sensing scenes [52]–[54]. However, neither natural scene nor remote sensing image captioning models can be directly applied to underwater image captioning due to domain disparities and the lack of underwater caption datasets containing diverse marine organisms. Although existing open-vocabulary foundation models [55]–[57] can recognize out-of-distribution objects, they exhibit poor performance in underwater computer vision tasks due to significant domain gaps and target-level discrepancies.

Previous underwater image captioning studies predominantly rely on CNN-based architectures. Li et al. [58] employ Faster R-CNN to extract global and local features, then generate semantic descriptions of underwater targets through feature fusion and contextual information sorting. Kerai et al. [59] integrate object detection with image captioning via attention mechanisms, adopting an end-to-end framework for underwater captioning. However, these methods suffer from repetitive high-frequency words usage in generated captions, exacerbated by the limited size of available datasets. Recently, UICMSOFF [12] introduce an underwater image captioning dataset spanning diverse underwater scenes. The model employs physical-degradation pre-training and meta-learning strategies to fuse scene-target features, enabling robust handling of underwater captioning tasks.

In this paper, we will introduce two underwater image captioning benchmark datasets UWS-IC and SUIM-IC derived from existing underwater image segmentation datasets UWS [6] and SUIM [19], respectively. We further present a series of baseline models for IC tasks to establish performance benchmarks. Finally, we propose a novel network architecture specifically designed to tackle the unique challenges inherent in underwater image captioning, including degraded visual quality and domain-specific semantic gaps.

III. Methodology

The proposed P^{2}UIC is shown in Fig. 2, which aims to integrate the Mamba composed of selective SSM and the environment-aware transformer into UIC. This section first introduces the basic concept of Mamba, then introduces the architecture and overall process of P^{2}UIC, and finally introduces the plug-and-play UPEE and CMMD component modules.

A. Preliminaries

Both SSM-based models and Mamba are derived from the linear sequence model, which maps a 1-D sequence $x(t) \in \mathbb{R} \rightarrow y(t)$ through a hidden state $h(t) \in \mathbb{R}^N$. Then, we define the evolution parameter $\mathbf{A} \in \mathbb{R}^{N\times N}$ and the projection parameters $\mathbf{B} \in \mathbb{R}^{N\times 1}$ and $\mathbf{C} \in \mathbb{R}^{1\times N}$, the linear sequence model can be expressed as $h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t)$ and $y(t) = \mathbf{C}h(t)$. In order to be integrated into the deep learning model, S4 and Mamba need to discretize $\mathbf{A}$ and $\mathbf{B}$ through zero-order hold (ZOH), defined as follows:

$$\begin{aligned} \bar{\mathbf{A}} &= \exp(\Delta\mathbf{A}) \\ \bar{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B} \end{aligned} \tag{1}$$

After obtaining the discretized $\Delta\mathbf{A}$ and $\Delta\mathbf{A}$, the discretized SSM can be re-expressed as:

$$\begin{aligned} h_t &= \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t \\ y_t &= \mathbf{C}h_t \end{aligned} \tag{2}$$

Finally, the input sequence x of length L is subjected to global convolution with a structured convolution kernel of $\mathbf{K}$ to obtain the output:

$$\begin{aligned} y &= x * \bar{\mathbf{K}} \\ \bar{\mathbf{K}} &= \left(\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, \ldots, \mathbf{C}\bar{\mathbf{A}}^{L-1}\bar{\mathbf{B}}\right) \end{aligned} \tag{3}$$

B. Underwater Physics Environment-aware Enhancement

We use pre-trained CLIP with ViT-B to extract L image patch features $\mathbf{F_{CLIP}} = \{\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_L\} \in \mathbb{R}^{L\times d}$ as the input of the UPEE module, shown in Fig. 3.

Firstly, we convert image patch features $\mathbf{F_{CLIP}}$ into 2-D feature map $\mathbf{F}'$, which is input into the physical parameter simulation network together with the underwater keywords to obtain attenuation coefficient α and scattering coefficient β through the physical parameter simulation network.

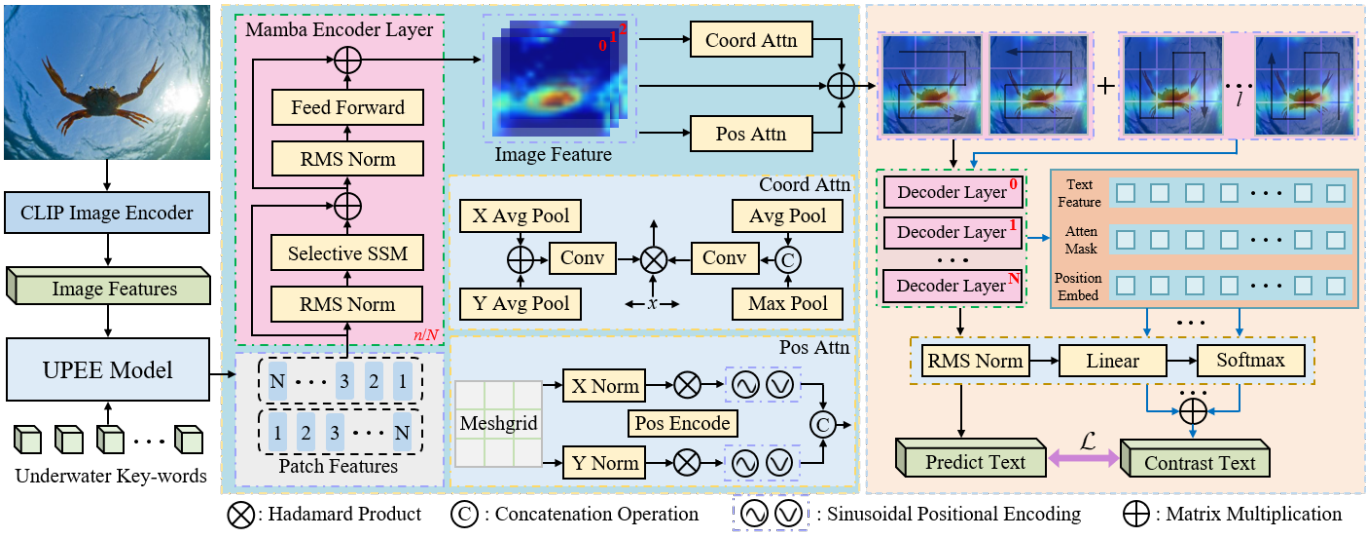

Fig. 2. Overview of the proposed P^{2}UIC framework. The patch features extracted by CLIP and environmental keywords are fed into UPEE to simulate the underwater environment and fine-tune the patch features. The encoder layer consists of N mamba blocks and fuse the image features from high to low. The fused features are fed into the CMMD, through different sequence arrangements to generate word embedding and fuse them to obtain the final output.

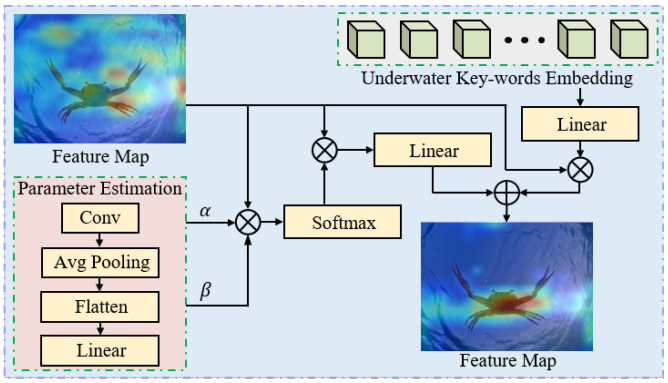

Fig. 3. The architecture of Underwater Physics Environment-aware Enhancement Model. α and β represent attenuation coefficient and scattering coefficient, respectively.

$$\alpha, \beta = \text{Linear}\left(\text{Flatten}\left(\text{Avgpool}\left(\text{Conv2D}\left(\mathbf{F}'\right)\right)\right)\right) \tag{4}$$

where $\text{Avgpool}(\cdot)$ means average pooling operation. We obtain $\mathbf{F_Q}$, $\mathbf{F_K}$ and $\mathbf{F_V}$ through the linear projection of $\mathbf{F}'$, and the image feature $\mathbf{F}''$ after the image feature $\mathbf{F}'$ is simulated by the physical parameter can be expressed as:

$$\mathbf{F}'' = \text{softmax}\left(\delta \cdot \beta \left(\frac{W_Q^T \mathbf{F}_Q \left(\frac{W_K^T \mathbf{F}_K}{\alpha}\right)^T}{\sqrt{d_k}}\right)\right) W_V^T \mathbf{F}_V \tag{5}$$

where δ is a fixed scaling parameter obtained by multi-head dimensions, W_Q, W_K and W_V are the linear projection weights for the *Query*, *Key* and *Value*, d_k is the projection channel dimension.

The training set noun embedding extracted by CLIP is used as the environmental keyword $\mathbf{L_K}$, and the keyword-guided image features are obtained through multi-head attention:

$$\mathbf{F}''' = \eta \cdot \text{softmax}\left(\frac{W_Q^T \mathbf{F}_Q \left(W_K^T \mathbf{L}_K\right)^T}{\sqrt{d_k}}\right) W_V^T \mathbf{F}_V + (1-\eta) \cdot \mathbf{F_{CLIP}} \tag{6}$$

where η represents the degree of integration of environmental key-words. Ablation studies show that we achieve the best results when $\eta = 0.5$.

The adaptive fusion score of the image feature $\mathbf{F}''$ after physical parameter simulation and the image feature $\mathbf{F}'''$ guided by the keyword is:

$$s = \text{Sigmoid}\left(\text{Linear}\left(\text{concat}\left(\mathbf{F}'', \mathbf{F}'''\right)\right)\right) \tag{7}$$

The output feature of the UPEE module is:

$$\mathbf{F} = s \cdot \mathbf{F}'' + (1-s)\, \mathbf{F}''' \tag{8}$$

Algorithm 1 Processing flow of UPEE.

Input: $\mathbf{F}_{\text{CLIP}}, \mathbf{L_K}$
Output: UPEE features $\mathbf{F}$
$\alpha, \beta \leftarrow$ Physical Parameter Simulation $\leftarrow \mathbf{F}'$
$\mathbf{F_Q}, \mathbf{F_K}, \mathbf{F_V} \leftarrow$ Linear Projection $\leftarrow \mathbf{F}'$
$\mathbf{F}'' \leftarrow$ Attention $\leftarrow \alpha, \beta, \mathbf{F_Q}, \mathbf{F_K}, \mathbf{F_V}$
$\mathbf{F}''' \leftarrow$ Attention $\leftarrow \mathbf{F_Q}, \mathbf{L_K}, \mathbf{F_V}, \mathbf{F}$
$\mathbf{F} \leftarrow s \leftarrow \mathbf{F}'', \mathbf{F}'''$

C. Mamba Encoder

The Mamba encoder consists of N identical encoder layers, each of which contains a SSM, feedforward network (FFN), and RMS normalization (RMSNorm). The input of each encoder layer is the output of the previous encoder layer, so the output $\mathbf{V}_n \in \mathbb{R}^{L \times d}$ of the nth encoder layer is:

$$\mathbf{U}_n = \text{RMSNorm}\left(\text{S6}\left(\mathbf{F}_n\right)\right) + \mathbf{F}_n \tag{9}$$

$$\mathbf{V}_n = \text{RMSNorm}\left(\text{FFN}\left(\mathbf{U}_n\right)\right) + \mathbf{U}_n \tag{10}$$

where $\text{RMSNorm}\left(\cdot\right)$ means root mean square normalization, we refer to Vim to convert the input image features into image sequences $\mathbf{F}_a$ and $\mathbf{F}_b$ in both positive and negative

directions, obtain the output of the mamba encoder according to the above formula, and fuse them to get the final output $\mathbf{V} = \{\mathbf{V}_1, \ldots, \mathbf{V}_N\}$. Then, we use dilated convolutions with different dilation rates to obtain multiscale features $\mathbf{V}^0 = \{\mathbf{V}_0^0, \ldots, \mathbf{V}_N^0\}$, $\mathbf{V}^1 = \{\mathbf{V}_0^1, \ldots, \mathbf{V}_N^1\}$ and $\mathbf{V}^2 = \{\mathbf{V}_0^2, \ldots, \mathbf{V}_N^2\}$. Due to the problems of incomplete description and insufficient details in the existing IC results, we add CoordAttention to the low-level feature $\mathbf{V^0}$. At the same time, in order to reduce the target category error, we add PositionEmbeddingSine to the high-level feature $\mathbf{V^2}$. Then we obtain the output of the entire mamba encoder:

$$\mathbf{V} = \mu_0 \mathrm{CoordAttn}\left(\mathbf{V}^0\right) + \mu_1 \mathbf{V}^1 + \mu_2 \mathrm{PosAttn}\left(\mathbf{V}^2\right) \quad (11)$$

where μ_0, μ_1 and μ_2 are learnable adaptive fusion parameters, $\mathrm{CoordAttn}(\cdot)$ and $\mathrm{PosAttn}(\cdot)$ mean coordinate attention and PositionEmbeddingSine attention in Fig. 2, respectively.

D. Contrastive Multi-sequence Mamba Decoder

The proposed CMMD consists of a mamba decoder and a text sequence generator, where the mamba decoder is the same as the encoder and consists of N stacked mamba blocks. The output of the nth mamba block can be expressed as:

$$\mathbf{Y}_n = \mathrm{RMSNorm}\left(\mathrm{S6}\left(\mathbf{V}_n\right)\right) + \mathbf{V}_n \quad (12)$$

$$\mathbf{T}_n = \mathrm{RMSNorm}\left(\mathrm{FFN}\left(\mathbf{Y}_n\right)\right) + \mathbf{Y}_n \quad (13)$$

We scan the encoder output image feature in the same way as the encoder and obtain the output results $\mathbf{T}^a$ and $\mathbf{T}^b$ of the standard decoder. At the same time, we scan the encoder output image feature in l different ways from the encoder to obtain different image sequences, which are output by the mamba decoder to obtain $\{\mathbf{T}_1, \ldots, \mathbf{T}_l\}$ and enter the text sequence generator. We random choose multiple encoder layers of frozen BERT, RoBERTa, GPTv3 and XLNet to obtain text sequence generator, one of the encoder layers can be expressed as:

$$\mathbf{T'}_l = \mathrm{Linear}\left(\mathrm{Norm}\left(\mathbf{T}_l + \mathrm{MHAttn}\left(\mathbf{T}_l\right)\right)\right) \quad (14)$$

$$+ \mathrm{FFN}\left(\mathrm{Norm}\left(\mathbf{T}_l + \mathrm{MHAttn}\left(\mathbf{T}_l\right)\right)\right) \quad (15)$$

where $\mathrm{MHAttn}(\cdot)$ represents multi head attention. After that, $\mathbf{T}_n$ and $\mathbf{T'}_l$ pass through the fully connected layer and the softmax layer in turn to obtain the word probabilities of the predicted text $\mathbf{T}_{pred}$ and all the comparison texts, and the comparison texts are fused to obtain $\mathbf{T}_{contrast}$. Finally, we obtain the similarity through contrastive loss:

$$\mathcal{L}_{sim} = 1 - \frac{\mathbf{T}_{\mathrm{pred}} \cdot \mathbf{T}_{contrast}}{\|\mathbf{T}_{\mathrm{pred}}\| \, \|\mathbf{T}_{contrast}\|} \quad (16)$$

E. Loss function

Cross Entropy Loss: According to the standard image captioning loss function, we use the word-level cross-entropy loss(XE):

$$\mathcal{L}_{\mathrm{XE}} = -\sum_{t=1}^{T} \log\left(p\left(y_t^* | y_{1:t-1}^*\right)\right) \quad (17)$$

where $y_{1:t-1}^* = \left[y_1^*, \ldots, y_{t-1}^*\right]$ represents the GroundTruth (GT) sequence from the beginning to the $t-1$ time step, and T is the maximum time step.

CIDEr-D Score: During the training process, a beam search strategy is used to select the top k sentences with the highest probability and calculate their CIDEr-D scores as rewards:

$$\nabla \mathcal{L}_{\mathrm{RL}} = -\frac{1}{k} \sum_{i=1}^{k} \left(\left(r\left(w^i\right) - b\right) \nabla \log p\left(w^i\right)\right) \quad (18)$$

where k is the number of samples; w^i represents the ith sentence sampled in the beam; $r(\cdot)$ represents the reward function; $b = 1/k \sum_{i=1}^{k} r\left(w^j\right)$ is the baseline reward.

Therefore, the overall loss function of the proposed P^{2}UIC can be expressed as:

$$\mathcal{L} = \lambda_1 \nabla \mathcal{L}_{\mathrm{RL}} + \lambda_2 \mathcal{L}_{\mathrm{XE}} + \lambda_3 \mathcal{L}_{sim} \quad (19)$$

where λ_1, λ_2 and λ_3 are empirically adjusted, here we set them as 0.5, 0.4 and 0.1 by default for all the following experiments.

Algorithm 2 Processing flow of P^{2}UIC.

Input: Image $\mathbf{I}$, Keywords $\mathbf{L_K}$
Output: Predict Text $\mathbf{T_n}$
$\mathbf{F_{CLIP}} \leftarrow$ CLIP Image Backbone $\leftarrow \mathbf{I}$
$\mathbf{F} \leftarrow$ UPEE $\leftarrow \mathbf{F_{CLIP}}$
/* SSM Block Process* /
for n in N **do**
  $\mathbf{F}'_n \leftarrow \mathrm{Linear}\left(\mathrm{Norm}\left(\mathbf{F}_n\right)\right) + \mathbf{F}_n$
  $\mathbf{X}, \mathbf{Z} \leftarrow \mathrm{Linear}^{\mathrm{x}}\left(\mathbf{F}'_n\right), \mathrm{Linear}^{\mathrm{z}}\left(\mathbf{F}'_n\right)$
  for o in {forward, backward} **do**
    $\mathbf{X}'_o \leftarrow \mathrm{SiLU}\left(\mathrm{Conv1d_o}\left(\mathbf{X}\right)\right)$
    $\mathbf{B_o}, \mathbf{C_o} \leftarrow \mathrm{Linear}_o^{\mathrm{B}}\left(\mathbf{X}'_o\right), \mathrm{Linear}_o^{\mathrm{C}}\left(\mathbf{X}'_o\right)$
    $\Delta_{\mathrm{o}} \leftarrow \log\left(1 + \exp\left(\mathrm{Linear}_o^{\Delta}\left(\mathrm{x'}_o\right) + \mathrm{Parameter}_o^{\Delta}\right)\right)$
    $\bar{\mathbf{A}}_{\mathbf{o}}, \bar{\mathbf{B}}_{\mathbf{o}} \leftarrow \Delta_{\mathrm{o}} \otimes \mathrm{Parameter}_o^{\mathbf{\Delta}}, \Delta_{\mathrm{o}} \otimes \mathbf{B}_o$
    $\mathbf{H_{-1}} \leftarrow \mathbf{0}$
    $\mathbf{H_o} \leftarrow \mathrm{pscan}\left(\bar{\mathbf{A}}_{\mathbf{o}} \otimes \mathbf{H_{-1}} + \bar{\mathbf{B}}_{\mathbf{o}} \otimes \mathbf{X'_O}\right)$
    $\mathbf{H_x} \leftarrow \mathrm{Linear}\left(\mathrm{Att}\left(\mathrm{Linear}\left(\mathbf{H}_o\right)\right)\right)$
    $\mathbf{Y_o} \leftarrow \mathbf{H_x} \otimes \mathbf{C_o}$
  end for
  $\mathbf{F}_{n+1} \leftarrow \mathrm{Linear}\left(\mathbf{Z} \odot \mathbf{Y}_m + \mathbf{Z} \odot \mathbf{Y}_c\right)$
end for
$\mathbf{V} \leftarrow \mathrm{CoordAttn}\left(\mathbf{V^0}\right), \mathbf{V^1}, \mathrm{PosAttn}\left(\mathbf{V^2}\right) \leftarrow \mathrm{DC}\left(\mathbf{V}\right)$
$\mathbf{T_{pred}}, \{\mathbf{T_1}, \ldots, \mathbf{T_k}\} \leftarrow$ SSM Process $\leftarrow \mathrm{Scan}\left(\mathbf{V}\right)$
$\mathbf{T_{contrast}} \leftarrow$ Text Sequence Generator $\leftarrow \{\mathbf{T_1}, \ldots, \mathbf{T_k}\}$
Predict Text $\mathbf{T_n} \leftarrow \mathbf{T_{pred}}, \mathbf{T_{contrast}}$

IV. Experiments

A. Datasets

This paper evaluates the proposed method on three UIC datasets: UICM-SOFF [12], SUIM-IC and UWS-IC, where UICM-SOFF is a public dataset. We derive SUIM-IC and UWS-IC from existing underwater segmentation datasets, providing two UIC datasets with more semantically rich, expert-annotated descriptions, shown in Fig. 4.

UICM-SOFF: The UICM-SOFF dataset contains 3,176 images spanning 10 categories: fish, coral reefs, divers, aquatic

TABLE I
EXPERMENT RESULTS ON UICM-SOFF DATASET. BN, M, R, C AND S DENOTE BLEU-N, METEOR, ROUGE$_L$, CIDER-D AND SPICE, RESPECTIVELY. ALL RESULTS ARE REPORTED AS PERCENTAGE (%) AND THE BEST RESULTS ARE PRESENTED IN BOLD.

Methods	B1	B2	B3	B4	M	R	C	S	S_m^*
mRNN [60]	42.03±1.69	28.90±1.76	17.40±2.08	10.53±2.38	10.96±0.29	29.27±0.93	28.47±5.51	6.34±0.92	19.81±1.49
Soft-Attention [61]	47.66±1.57	32.84±1.52	21.46±2.47	16.29±1.44	16.93±0.18	37.15±0.67	31.82±4.38	8.25±0.88	22.55±1.62
ORT [62]	48.15±1.06	33.72±1.43	23.07±1.26	17.12±1.02	18.23±0.09	37.92±0.85	33.18±7.59	9.71±0.43	26.61±3.15
AoANet [39]	52.74±1.11	36.24±0.99	26.71±1.77	19.39±1.29	19.21±0.31	40.55±0.52	37.92±3.27	11.61±0.38	29.27±0.93
World Sentence [63]	54.36±1.47	36.51±1.79	26.91±0.89	19.67±1.50	19.25±0.15	41.96±0.40	37.20±5.08	11.66±0.66	29.52±1.57
GVFG+LSGA [64]	55.22±1.60	36.78±1.36	26.94±1.20	19.45±1.08	19.08±0.11	42.31±0.62	36.27±4.92	11.59±0.53	29.28±1.43
RASG [65]	61.77±1.28	43.26±1.92	32.08±1.52	25.53±0.79	23.62±0.13	48.91±0.44	75.26±4.11	15.17±0.75	43.33±1.09
PKG-Transformer [52]	67.66±1.03	52.84±1.88	41.46±1.56	32.29±1.48	28.93±0.20	56.15±0.81	91.82±7.26	21.25±0.61	52.30±3.09
MG-Transformer [53]	71.38±0.42	58.34±0.52	46.98±0.50	37.18±0.35	28.68±0.12	57.12±0.35	112.01±6.58	21.82±0.16	58.75±0.98
CaDReL [44]	68.63±1.25	54.89±1.74	43.75±1.86	34.44±1.51	30.12±0.13	56.35±0.55	98.57±10.10	22.46±0.67	54.87±2.87
UICM-SOFF [12]	70.88±0.87	57.62±0.93	46.83±0.33	37.61±1.24	29.29±0.14	58.74±0.51	119.35±3.4	22.87±0.41	61.25±1.42
Mamba-IC [54]	71.26±1.84	57.80±2.23	46.69±2.27	37.282±2.15	30.11±0.07	58.478±1.22	112.36±14.88	**23.80±0.50**	59.56±4.58
RSIC-Gmamba [54]	71.51±1.43	58.04±1.39	46.73±1.87	37.24±1.67	29.64±0.10	58.87±0.65	116.72±10.08	23.91±0.65	58.95±3.13
Ours	**72.78±0.58**	**59.20±0.57**	**48.24±0.52**	**38.50±0.46**	**30.83±0.13**	**59.85±0.44**	**118.53±3.78**	23.77±0.03	**61.93±1.75**

TABLE II
EXPERMENT RESULTS ON UWS-IC DATASET. BN, M, R, C AND S DENOTE BLEU-N, METEOR, ROUGE$_L$, CIDER-D AND SPICE, RESPECTIVELY. ALL RESULTS ARE REPORTED AS PERCENTAGE (%) AND THE BEST RESULTS ARE PRESENTED IN BOLD.

Method	B1	B2	B3	B4	M	R	C	S	S_m^*
mRNN [60]	51.80±1.27	39.05±1.07	25.11±0.93	16.09±1.15	9.82±0.99	35.77±0.59	16.15±6.39	6.09±0.66	19.46±1.91
Soft-Attention [61]	56.67±1.42	43.29±0.77	29.97±0.88	19.16±2.06	14.71±0.53	39.50±0.61	18.06±7.08	6.99±0.54	22.86±2.13
ORT [62]	57.21±1.22	43.81±1.12	30.77±2.15	19.84±1.61	15.58±1.10	40.17±0.81	18.93±5.91	7.28±0.39	23.63±1.86
AoANet [39]	61.77±0.88	47.08±1.35	33.19±1.32	22.67±1.52	18.12±0.36	43.95±0.44	22.17±6.08	7.92±0.47	26.73±2.28
World Sentence [63]	64.63±1.10	50.41±2.08	36.61±1.42	25.99±0.93	20.24±0.75	47.23±0.18	25.88±4.76	8.55±0.31	29.84±1.74
GVFG+LSGA [64]	66.79±0.78	51.47±1.30	37.07±1.08	26.47±1.47	21.20±1.24	46.85±0.37	23.97±5.12	8.63±0.60	29.62±0.86
RASG [65]	67.22±0.91	50.87±0.88	36.03±1.21	25.31±1.38	20.65±0.86	44.47±0.52	18.25±5.69	8.21±0.55	27.17±1.81
PKG-Transformer [52]	68.08±1.36	51.74±0.50	37.32±0.53	26.93±0.72	21.96±0.89	45.12±0.75	29.84±5.88	8.67±0.57	30.96±1.96
MG-Transformer [53]	72.07±0.51	65.28±0.52	57.74±0.73	**51.52±0.95**	24.78±0.62	**57.40±0.19**	156.60±7.50	18.54±0.11	72.58±0.93
CaDReL [44]	68.61±0.99	62.65±1.62	54.91±1.15	47.49±1.76	25.71±0.34	57.00±0.22	141.93±6.39	18.84±0.71	68.03±1.69
UICMSOFF [12]	69.69±0.63	64.44±0.72	56.82±1.54	49.46±1.34	25.72±0.87	55.95±0.21	147.95±7.39	18.41±0.38	69.77±2.02
Mamba-IC [54]	70.14±0.96	62.93±1.63	55.01±2.05	47.59±2.33	26.22±1.03	57.37±0.39	142.77±5.91	18.88±0.44	68.49±2.15
RSIC-Gmamba [54]	70.42±1.13	64.02±0.98	55.93±1.37	48.20±1.23	25.77±0.59	56.49±0.28	148.95±7.33	18.30±0.58	68.44±2.40
Ours	**72.88±0.93**	**66.05±0.58**	**58.30±0.81**	50.74±0.96	**26.68±0.78**	56.86±0.16	**159.09±6.47**	**20.35±0.32**	**73.34±1.92**

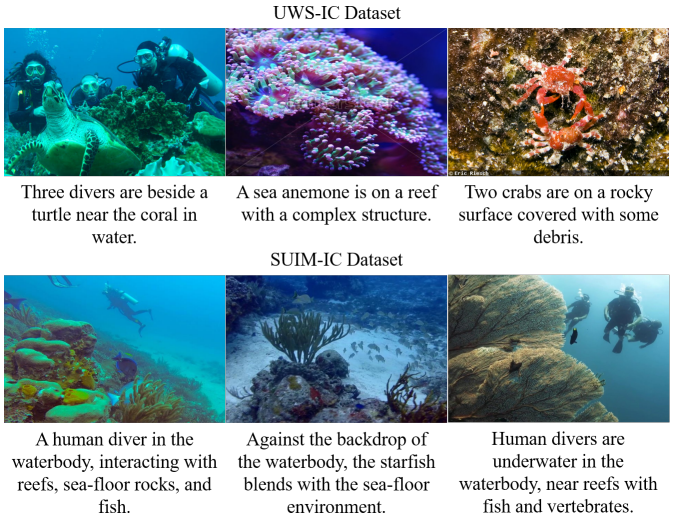

Fig. 4. Examples of UWS-IC and SUIM-IC datasets. It is worth mentioning that only one of the five descriptions corresponding to each image is shown.

plants, turtles, man-made objects, whales, other organisms, submersibles, and natural environments. Each image is paired with five expert-annotated descriptions. Image resolutions range from 119×300 to 8432×4743 pixels and all images are uniformly resized to 224×224 pixels for model input. The average word count per description is 10.07.

SUIM-IC: The SUIM-IC dataset comprises 1,635 images across 8 categories: background, fish, reefs, aquatic plants, wrecks/ruins, sea-floor, human divers and robots. Each image is annotated with five semantically rich, expert-curated descriptions, with an average word count of 12.26 words per caption. The dataset features diverse spatial resolutions (e.g., 1906×1080, 1280×720, 640×480), reflecting real-world underwater imaging scenarios.

UWS-IC: The UWS-IC dataset consists 576 images spanning 29 diverse categories: crab, crocodile, dolphin, frog, nettles, octopus, otter, penguin, polar bear, sea anemone, sea urchin, seahorse, seal, shark, shrimp, star fish, stingray, squid, turtle,

TABLE III
EXPERMENT RESULTS ON SUIM-IC DATASET. BN, M, R, C AND S DENOTE BLEU-N, METEOR, ROUGE$_L$, CIDER-D AND SPICE, RESPECTIVELY. ALL RESULTS ARE REPORTED AS PERCENTAGE (%) AND THE BEST RESULTS ARE PRESENTED IN BOLD.

Method	B1	B2	B3	B4	M	R	C	S	S_m^*
mRNN [60]	55.82±1.03	43.29±1.76	33.76±2.13	26.13±1.72	13.29±2.37	39.22±0.82	35.75±9.37	11.72±1.42	28.60±1.78
Soft-Attention [61]	60.17±1.15	48.56±2.03	38.72±2.17	30.28±1.45	15.91±2.21	42.83±0.96	45.31±10.49	13.08±1.77	33.58±1.82
ORT [62]	61.82±0.93	49.23±1.41	39.86±2.04	31.75±1.81	17.28±1.77	45.01±0.74	47.82±9.52	13.59±1.74	35.47±1.65
AoANet [39]	65.08±0.92	53.14±1.15	42.96±1.76	34.07±1.58	19.85±2.04	47.32±0.86	50.08±9.43	14.02±1.52	37.83±1.33
World Sentence [63]	68.16±0.64	56.99±1.37	45.34±1.57	36.61±1.65	22.51±1.96	50.13±0.66	52.60±10.07	14.52±1.46	40.46±1.41
GVFG+LSGA [64]	70.36±0.71	58.50±1.48	47.78±0.83	38.22±1.39	25.14±1.65	52.93±0.77	57.37±9.89	15.40±1.09	43.42±1.39
RASG [65]	71.17±0.65	59.23±1.19	48.81±1.92	39.41±1.27	25.36±1.83	54.14±0.79	57.48±10.15	15.69±1.38	44.10±1.36
PKG-Transformer [52]	72.87±0.98	60.68±1.57	49.92±0.51	40.37±1.52	24.64±1.98	54.74±0.44	54.82±5.17	16.68±1.24	43.64±0.78
MG-Transformer [53]	75.25±0.55	65.62±1.87	**57.03±0.31**	**49.12±1.68**	26.17±2.46	63.21±0.71	94.06±11.31	21.22±1.36	58.14±1.15
CaDReL [44]	74.99±1.31	64.53±1.02	54.02±0.78	45.61±1.59	23.80±1.81	62.39±0.91	92.89±10.77	20.39±1.68	56.17±1.52
UICMSOFF [12]	75.23±0.88	65.03±1.35	55.03±1.84	45.31±1.44	25.39±2.03	64.81±0.59	91.59±10.53	20.08±1.52	56.78±1.60
Mamba-IC [54]	75.88±1.67	64.44±1.12	56.17±1.58	45.19±1.35	24.02±2.58	65.44±0.77	77.37±10.04	19.67±1.93	53.01±1.69
RSIC-Gmamba [54]	76.19±0.91	64.88±1.23	56.93±1.74	46.06±1.69	**28.57±2.14**	64.02±0.94	93.69±10.88	20.62±1.13	58.09±1.85
Ours	**77.46±0.63**	**66.50±1.50**	56.21±1.67	46.60±1.60	27.75±1.75	**65.72±0.60**	**97.89±8.70**	**22.52±1.04**	**59.49±1.61**

TABLE IV
PLUG-AND-PLAY EXPERIMENT RESULTS ON UICM-SOFF DATASET. BN, M, R, C AND S DENOTE BLEU-N, METEOR, ROUGE$_L$, CIDER-D AND SPICE, RESPECTIVELY. ALL RESULTS ARE REPORTED AS PERCENTAGE (%) AND THE BEST RESULTS ARE PRESENTED IN BOLD. $*$ INDICATES THE MODEL WITH PLUG-AND-PLAY UPEE AND CMMD.

Method	B1	B2	B3	B4	M	R	C	S	S_m^*
PKG-Transformer	67.66±1.03	52.84±1.88	**41.46±1.56**	32.29±1.48	28.93±0.20	**56.15±0.81**	91.82±7.26	21.25±0.61	52.30±3.09
PKG-Transformer*	**68.43±0.87**	**53.69±1.56**	40.75±1.39	**33.04±1.56**	**29.62±0.23**	55.35±0.76	**92.57±6.93**	**21.46±0.68**	**52.65±2.97**
MG-Transformer	71.38±0.42	**58.34±0.52**	46.98±0.50	37.18±0.35	28.68±0.12	57.12±0.35	**112.01±6.58**	21.82±0.16	58.75±0.98
MG-Transformer*	**71.97±0.55**	58.19±0.49	**47.54±0.87**	**38.06±0.41**	**29.11±0.10**	**57.77±0.41**	110.91±6.75	**22.64±0.18**	**58.96±1.06**
UICMSOFF	70.88±0.87	57.62±0.93	46.83±0.33	37.61±1.24	29.29±0.14	58.74±0.51	119.35±3.40	22.87±0.41	61.25±1.42
UICMSOFF*	**71.62±0.77**	**58.43±1.08**	**47.52±0.67**	**38.12±1.30**	**30.01±0.13**	**59.10±0.49**	**119.82±3.57**	**23.25±0.37**	**61.76±1.48**
RSIC-Gmamba	71.51±1.43	58.04±1.39	**46.73±1.87**	37.24±1.67	29.64±0.10	**58.87±0.65**	**116.72±10.08**	23.91±0.65	58.95±3.13
RSIC-Gmamba*	**72.19±1.31**	**58.67±1.53**	46.49±1.76	**38.03±1.44**	**30.15±0.11**	57.67±0.64	115.08±9.77	**24.23±0.41**	**60.23±2.86**

whale, nudibranch, coral, rock, water, sand, plant, human, reef and others. Each image is accompanied by five richer expert descriptions with an average description length of 17.13 words.

B. Evaluation Metrics

To assess UIC model performance, we follow previous works [12], [53], [54], [66] and employ six standard image captioning metrics. The definitions of these metrics are as follows: 1) BLEU [67] evaluates similarity by calculating the phrase overlap ratio between candidate sentences and reference sentences. 2) METEOR [68] aims to calculate the harmonic mean of precision and recall. 3) ROUGE$_L$ [69] evaluates similarity by measuring the longest common subsequence between candidate sentences and reference sentences. 4) CIDEr-D [70] weights the importance of different words by calculating the Term Frequency Inverse Document Frequency (TF-IDF) and calculates cosine similarity. 5) SPICE [69] converts captions into a series of semantic tuples containing objects, attributes, and relations, and measures the degree of match between the generated annotations and the reference annotations by F-score. 6) S_m^* is the mean of BLEU-4, METEOR, ROUGE$_L$ and CIDEr-D.

C. Experimental Setting

Dataset splitting: To ensure fair comparison with existing methods, we randomly partition each dataset into training (70%), validation (15%) and testing (15%) subsets, following the spilt ratio in UICM-SOFF dataset [12]. Inspired by prior evaluation protocols [53], [54], we conduct five independent experiments on each dataset using distinct random splits. Results are averaged across these splits to mitigate randomness and provide statistically robust performance estimates.

Implementation Details: We use the pre-trained CLIP [52] image encoder to extract patch features of the image and match its dimensions with the embeddings of the text decoder [71]. The Mamba visual encoder generates multi-scale visual features through dilated convolutions with dilation rates of 2, 4, and 8. The text features are obtained using word embeddings and position encoding with self-attention, and the maximum length of the predicted sentence is 20. We perform experiments using NVIDIA GeForce RTX 3090ti GPU, an Intel(R) Core(TM) i9-12900KF CPU, a batch size of 50, and run training for 30 epochs using the AdamW optimizer with a learning rate of 2×10^{-4} and a weight decay of 0.1.

Compared Methods: We compare the proposed method

TABLE V
PLUG-AND-PLAY EXPERIMENT RESULTS ON UWS-IC DATASET. BN, M, R, C AND S DENOTE BLEU-N, METEOR, ROUGE$_L$, CIDER-D AND SPICE, RESPECTIVELY. ALL RESULTS ARE REPORTED AS PERCENTAGE (%) AND THE BEST RESULTS ARE PRESENTED IN BOLD. $*$ INDICATES THE MODEL WITH PLUG-AND-PLAY UPEE AND CMMD.

Method	B1	B2	B3	B4	M	R	C	S	S_m^*
PKG-Transformer	68.08±1.36	51.74±0.50	37.32±0.53	26.93±0.72	21.96±0.89	45.12±0.75	29.84±5.88	8.67±0.57	30.96±1.96
PKG-Transformer*	**69.13±1.48**	**54.39±1.28**	**39.55±1.54**	**30.49±1.07**	**23.08±1.10**	**47.28±1.59**	**34.46±7.39**	**11.01±1.29**	**33.83±2.12**
MG-Transformer	**72.07±0.51**	65.28±0.52	57.74±0.73	51.52±0.95	24.78±0.62	**57.40±0.19**	**166.60±7.50**	18.54±0.11	75.08±0.93
MG-Transformer*	71.31±0.77	**65.76±0.74**	**58.12±0.79**	**52.25±1.03**	**26.06±0.84**	56.92±0.17	165.33±7.84	**19.70±0.24**	**75.14±0.98**
UICMSOFF	69.69±0.63	64.44±0.72	56.82±1.54	**49.46±1.34**	25.72±0.87	55.95±0.21	147.95±7.39	18.41±0.38	69.77±2.02
UICMSOFF*	**70.44±0.42**	**65.08±0.81**	**57.94±1.34**	48.75±1.35	**26.41±0.93**	**56.12±0.18**	**149.33±7.09**	**19.50±0.45**	**70.15±2.22**
RSIC-Gmamba	70.42±1.13	64.02±0.98	55.93±1.37	48.20±1.23	**25.77±0.59**	56.49±0.28	148.95±7.33	18.30±0.58	68.44±2.40
RSIC-Gmamba*	**72.48±1.00**	**65.73±0.82**	**57.49±1.22**	**50.06±1.03**	25.64±0.71	**56.53±0.31**	**151.35±6.88**	**20.01±0.48**	**70.90±2.16**

TABLE VI
PLUG-AND-PLAY EXPERIMENT RESULTS ON SUIM-IC DATASET. BN, M, R, C AND S DENOTE BLEU-N, METEOR, ROUGE$_L$, CIDER-D AND SPICE, RESPECTIVELY. ALL RESULTS ARE REPORTED AS PERCENTAGE (%) AND THE BEST RESULTS ARE PRESENTED IN BOLD. $*$ INDICATES THE MODEL WITH PLUG-AND-PLAY UPEE AND CMMD.

Method	B1	B2	B3	B4	M	R	C	S	S_m^*
PKG-Transformer	**72.87±0.98**	**60.68±1.57**	**49.92±0.51**	40.37±1.52	24.64±1.98	54.74±0.44	54.82±5.17	16.68±1.24	43.64±0.78
PKG-Transformer*	72.22±1.24	60.01±1.61	49.56±0.77	**40.75±1.34**	**25.91±1.66**	**56.43±0.52**	**61.29±6.73**	**17.02±1.11**	**46.10±0.88**
MG-Transformer	75.25±0.55	65.62±1.87	57.03±0.31	49.12±1.68	26.17±2.46	63.21±0.71	**94.06±11.31**	21.22±1.36	58.14±1.15
MG-Transformer*	**76.03±0.61**	**66.23±1.64**	**57.91±0.69**	**49.32±1.57**	**26.71±2.19**	**63.79±0.67**	93.08±10.85	**21.86±1.29**	**58.23±1.26**
UICMSOFF	75.23±0.88	**65.03±1.35**	55.03±1.84	45.31±1.44	25.39±2.03	64.81±0.59	91.59±10.53	20.08±1.52	56.78±1.60
UICMSOFF*	**76.15±0.79**	64.66±1.48	**55.86±1.90**	**45.67±1.51**	**26.09±1.79**	**65.08±0.71**	**92.39±10.22**	**20.09±1.37**	**57.31±1.53**
RSIC-Gmamba	76.19±0.91	64.88±1.23	**56.93±1.74**	46.06±1.69	**28.57±2.14**	64.02±0.94	93.69±10.88	20.62±1.13	58.09±1.85
RSIC-Gmamba*	**76.98±0.73**	**66.61±1.42**	56.37±1.88	**46.50±1.77**	27.94±1.99	**65.27±0.77**	**94.80±10.09**	**22.07±1.28**	**58.63±1.79**

TABLE VII
ABLATION STUDIES ON UICM-SOFF DATASET. BN, M, R, C AND S DENOTE BLEU-N, METEOR, ROUGE$_L$, CIDER-D AND SPICE, RESPECTIVELY. ALL RESULTS ARE REPORTED AS PERCENTAGE (%) AND THE BEST RESULTS ARE PRESENTED IN BOLD.

UPEE	CMMD	B1	B2	B3	B4	M	R	C	S	S_m^*
-	-	71.26±1.84	57.80±2.23	46.69±2.27	37.28±2.15	30.11±0.07	58.48±1.22	112.36±14.88	**23.80±0.50**	59.56±4.58
✓	-	72.18±1.42	58.63±1.77	47.74±1.62	38.09±1.49	30.67±0.10	59.55±1.07	116.18±8..71	23.74±0.41	61.12±3.72
-	✓	71.94±1.07	58.42±1.29	47.35±1.33	37.88±1.37	30.39±0.09	59.02±0.89	115.88±6.14	23.59±0.32	60.79±3.28
✓	✓	**72.78±0.58**	**59.20±0.57**	**48.24±0.52**	**38.50±0.46**	**30.83±0.13**	**59.85±0.44**	**118.53±3.78**	23.77±0.03	**61.93±1.75**

against 13 IC methods. 1) CNN-RNN framework: mRNN [60] uses different CNNs as encoders to process image features and uses different RNNs as decoders to generate text. Soft-attention [61] establish connections between image regions and words based on soft attention. 2) Transformer framework: ORT [62] improves the spatial relationship between objects in region proposals by geometric attention. AoANet [39] acquires textual information by extending traditional attention. World Sentence [63] extracts as many words as possible from images through a word extractor to assist sentence generation. GVFG+LSGA [64] combines global and local visual features to provide salient visual features and proposes language states to provide text features. RASG [65] enhances the representation of the current word state through recurrent attention and semantic gating and focuses on effective information for understanding the image. PKG-Transformer [52] guides the fusion of the relevance and difference of scene-level and object-level features through prior knowledge. MG-Transformer [53] extracts multi-scale image features through dilated convolution and combines image with text using semantic correlation module. CaDRel [44] fuses visual information from different scales through cascaded bridging diffusion to capture rich contextual information. UICM-SOFF [12] UICM-SOFF deeply explores UIC from three perspectives: challenges, models, and datasets, and proposes an UIC model based on scene-target feature fusion. 3) Mamba framework: Mamba-IC [54] introduces the mamba architecture to the image captioning task. RSIC-Gmamba [54] integrates heuristic genetic algorithm and Transformer into Mamba framework to fully capture contextual information.

D. Comparisons with Existing SOTA Methods

We summarize the performance comparisons between our proposed method and existing SOTA methods on the UICM-

TABLE VIII
ABLATION STUDIES ON UWS-IC DATASET. BN, M, R, C AND S DENOTE BLEU-N, METEOR, ROUGE$_L$, CIDER-D AND SPICE, RESPECTIVELY. ALL RESULTS ARE REPORTED AS PERCENTAGE (%) AND THE BEST RESULTS ARE PRESENTED IN BOLD.

UPEE	CMMD	B1	B2	B3	B4	M	R	C	S	S_m^*
-	-	70.14±0.96	62.93±1.63	55.01±2.05	47.59±2.33	26.22±1.03	**57.37±0.39**	142.77±5.91	18.88±0.44	68.49±2.15
✓	-	72.08±0.98	65.17±1.21	57.21±1.54	49.78±1.58	26.17±0.88	56.58±0.28	153.43±6.77	20.03±0.39	71.49±2.08
-	✓	71.59±0.82	64.73±1.07	56.66±1.18	48.89±1.25	**26.89±0.64**	57.12±0.10	151.01±5.99	19.69±0.30	70.98±1.68
✓	✓	**72.88±0.93**	**66.05±0.58**	**58.30±0.81**	**50.74±0.96**	26.68±0.78	56.86±0.16	**159.09±6.47**	**20.35±0.32**	**73.34±1.92**

TABLE IX
ABLATION STUDIES ON SUIM-IC DATASET. BN, M, R, C AND S DENOTE BLEU-N, METEOR, ROUGE$_L$, CIDER-D AND SPICE, RESPECTIVELY. ALL RESULTS ARE REPORTED AS PERCENTAGE (%) AND THE BEST RESULTS ARE PRESENTED IN BOLD.

UPEE	CMMD	B1	B2	B3	B4	M	R	C	S	S_m^*
-	-	75.88±1.67	64.44±1.12	56.17±1.58	45.19±1.35	24.02±2.58	65.44±0.77	77.37±10.04	19.67±1.93	53.01±1.69
✓	-	77.09±0.93	66.02±1.38	**56.37±1.61**	46.28±1.54	26.88±2.13	**65.87±0.75**	94.18±9.26	22.29±1.36	58.30±1.57
-	✓	76.91±0.55	**65.87±1.59**	56.00±1.74	45.96±1.57	25.49±2.22	65.53±0.62	87.37±8.83	21.44±1.11	56.09±1.67
✓	✓	**77.46±0.63**	66.50±1.50	56.21±1.67	**46.60±1.60**	**27.75±1.75**	65.72±0.60	**97.89±8.70**	**22.52±1.04**	**59.49±1.61**

TABLE X
ABLATION STUDIES OF l IN CMMD. BN, M, R, C AND S DENOTE BLEU-N, METEOR, ROUGE$_L$, CIDER-D AND SPICE, RESPECTIVELY. ALL RESULTS ARE REPORTED AS PERCENTAGE (%) AND THE BEST RESULTS ARE PRESENTED IN BOLD.

l	B1	B2	B3	B4	M	R	C	S	S_m^*
0	71.92±0.61	58.56±0.63	47.71±0.57	38.02±0.55	30.27±0.15	59.16±0.56	116.37±3.49	22.86±0.06	60.96±1.79
1	72.51±0.49	59.02±0.53	47.97±0.47	38.34±0.41	30.62±0.12	59.43±0.34	117.69±3.91	23.65±0.03	61.52±1.67
2	**72.78±0.58**	**59.20±0.57**	**48.24±0.52**	**38.50±0.46**	**30.83±0.13**	**59.85±0.44**	**118.53±3.78**	**23.77±0.03**	**61.93±1.75**
3	71.35±0.60	58.17±0.74	47.18±0.66	37.62±0.68	29.88±0.17	58.92±0.65	115.15±4.18	22.67±0.05	60.39±1.93

TABLE XI
ABLATION STUDIES OF δ IN EQUATION 6. BN, M, R, C AND S DENOTE BLEU-N, METEOR, ROUGE$_L$, CIDER-D AND SPICE, RESPECTIVELY. ALL RESULTS ARE REPORTED AS PERCENTAGE (%) AND THE BEST RESULTS ARE PRESENTED IN BOLD.

δ	B1	B2	B3	B4	M	R	C	S	S_m^*
0.3	71.61±0.64	58.16±0.69	47.33±0.59	37.51±0.55	29.76±0.15	58.96±0.53	113.40±4.52	23.29±0.04	59.91±2.03
0.4	72.35±0.52	58.96±0.56	48.01±0.57	38.39±0.52	30.51±0.16	59.49±0.49	117.18±3.59	23.63±0.03	61.39±1.85
0.5	**72.78±0.58**	**59.20±0.57**	**48.24±0.52**	**38.50±0.46**	**30.83±0.13**	**59.85±0.44**	**118.53±3.78**	**23.77±0.03**	**61.93±1.75**
0.6	72.15±0.55	58.03±0.51	47.12±0.52	38.05±0.48	30.12±0.15	59.31±0.41	116.57±4.02	23.18±0.05	61.01±1.77

TABLE XII
COMPARION OF THE PROPOSED P^{2}UIC WITH OTHER METHODS IN PARAMETERS AND FLOPS IN UICM-SOFF TEST SET.

Method	Parameters(M)	FLOPs(G)
AoANet [39]	87.37	-
RASG [65]	53.44	-
PKG-Transformer [52]	31.94	1.58
MG-Transformer [53]	38.56	1.60
CaDReL [44]	68.90	-
Mamba-IC [54]	23.78	0.80
RSIC-Gmamba [54]	39.21	1.50
Ours	40.84	1.58

SOFF, UWS-IC, and SUIM-IC datasets in Tab I, Tab II and Tab III, respectively. All results are reported as percentages (%). As can be seen from these tables, compared with the existing methods, the proposed method has achieved SOTA results on three datasets. We will analyze the experiments on each dataset in detail as follows.

Tab I presents performance results on the UICM-SOFF dataset. The proposed method improves by about 1.3 % over the best existing method in BLEU-N, METEOR, ROUGE$_L$, CIDEr-D and S_m^*, and is slightly lower than Mamba-IC in SPICE. Our training and testing data follow the division of UICM-SOFF, that is, 70%, 15%, and 15% for training, validation, and testing, respectively. It is worth mentioning that the UICM-SOFF paper uses additional pre-training data for the backbone. In order to ensure the fairness of the experiment, we choose a unified backbone network for all methods and do not add additional pre-training data.

Tab II presents the results on proposed UWS-IC dataset. P^{2}UIC improves by about 0.8% on BLEU-1, BLEU-2, BLEU-

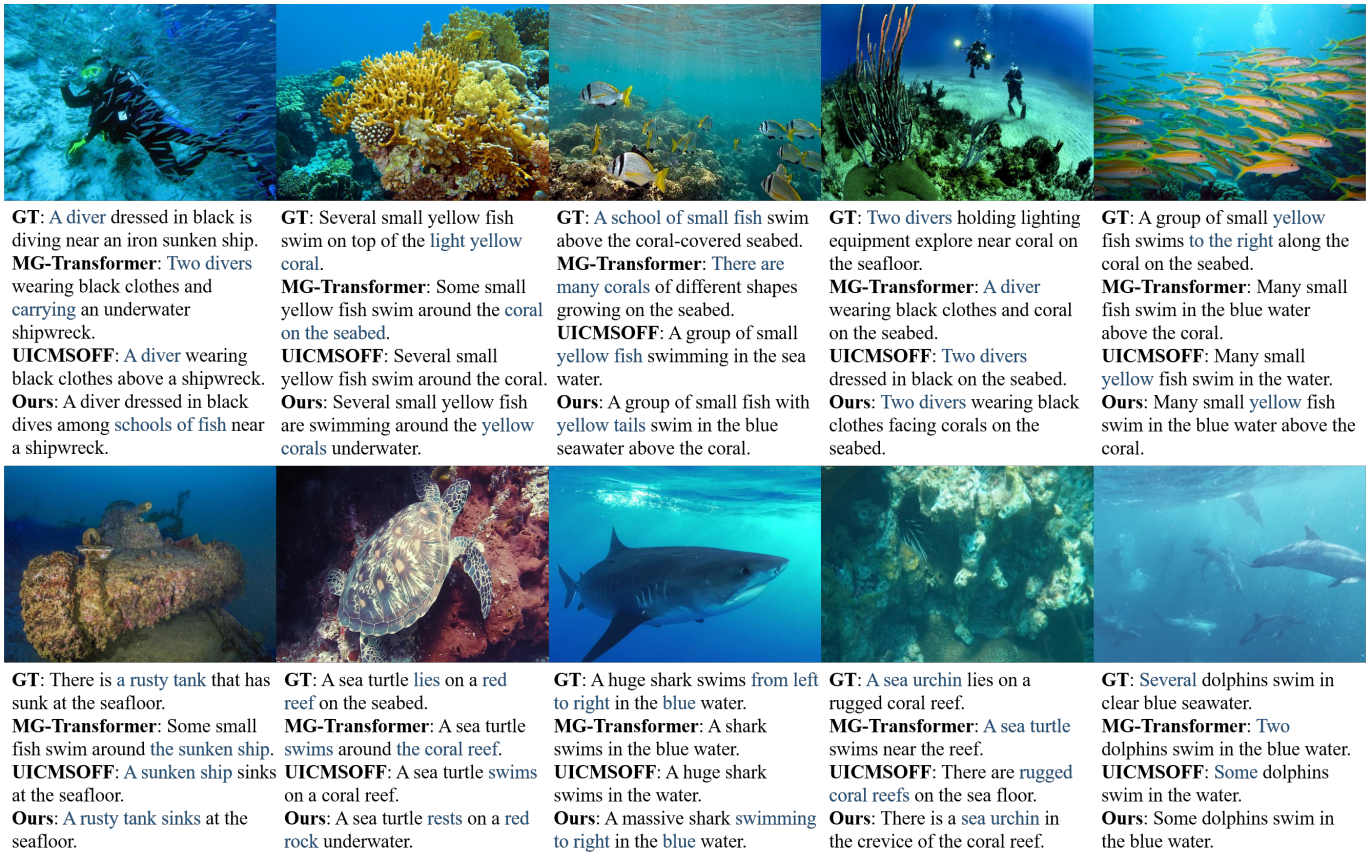

Fig. 5. Examples of captions generated by MG-Transformer, UICMSOFF and our P^{2}UIC on UICM-SOFF dataset. GT means ground truth and blue represents different word descriptions.

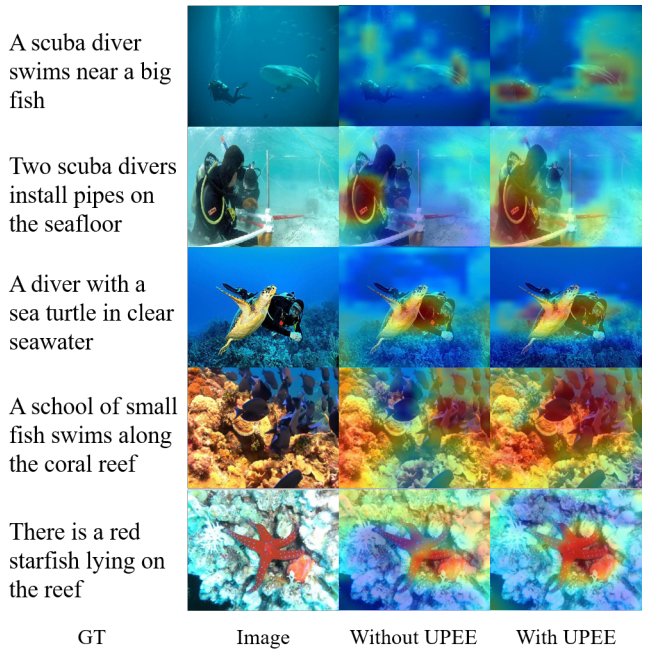

Fig. 6. Visualization of heatmaps generated by P^{2}UIC with/without UPEE module, where red means higher degree of attention and blue means lower degree of attention.

3, METEOR and S_m^*, 1.5% on SPICE, and 2.5% on CIDEr-D, while it is slightly lower than the MG-Transformer method in BLEU-4 and ROUGE$_L$. The underwater target categories of this dataset are much larger than those of other datasets, which can further illustrate the adaptability of the proposed method to underwater environment description.

Tab III presents the results on proposed SUIM-IC dataset. The proposed method improves by about 1.4% on BLEU-1 and S_m^*, 1.7% on BLEU-2 and ROUGE$_L$, 4% on CIDEr-D, and 2% on SPICE, while it is slightly lower than the existing methods in BLEU-3, BLEU-4 and METEOR. This dataset contains a large number of detailed description annotations of underwater target contexts, which can be used to verify the model's ability to generate diverse text.

E. Ablation Study

In this subsection, we conduct extensive ablation studies of the proposed model on three UIC datasets. Firstly, we verify the impact of the proposed plug-and-play UPEE and CMMD modules on the existing model and the results are shown in Tab IV, Tab V and Tab VI. Secondly, we analyze the effectiveness of the proposed components on proposed P^{2}UIC, as shown in Tab VII, Tab VIII and Tab IX. Finally, we experimentally verify the important parameters of the modules in Tab X and Tab XI.

1) Validity of plug-and-play modules UPEE and CMMD: We embed the UPEE module into the encoder of existing methods and replace the decoder with CMMD. Then we verify the generalization ability of these two modules in other IC

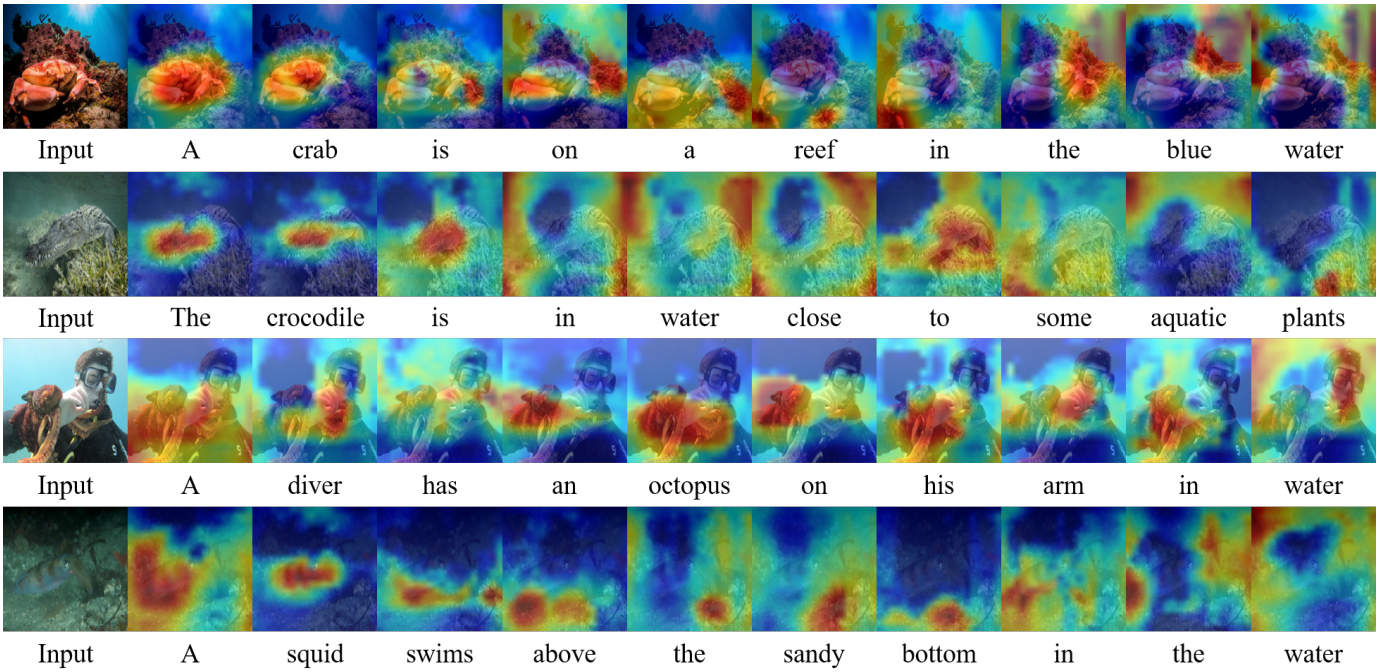

Fig. 7. Visualization of attented image regions along the caption generation processes for P^{2}UIC on UWS-IC dataset, where red means higher degree of attention and blue means lower degree of attention.

models on three UIC datasets, where $*$ indicates the model with plug-and-play UPEE and CMMD. It is worth mentioning that we do not choose CNN-RNN based method, but choose the latest transformer and Mamba methods, which is because UPEE is based on the transformer structure, while CMMD is a Mamba based decoder. Tab IV, Tab V and Tab VI represent the results on the UICM-SOFF, UWS-IC and SUIM-IC datasets, respectively. It can be seen that after adding these two plug-and-play modules to existing methods, the performance of the model has been improved in most indicators.

2) Effectiveness of UPEE and CMMD: Tab VII , Tab VIII and Tab IX represent the ablation studies of proposed UPEE and CMMD on UICM-SOFF, UWS-IC and SUIM-IC datasets, respectively. Overall, UPEE and CMMD can improve the performance of text generation and have improvements in most indicators, but there are some differences in each dataset. On UICM-SOFF dataset, SPICE is reduced by 0.03% and other indicators are the best when we add both UPEE and CMMD. On UWS-IC dataset, METEOR has the best results when we only add CMMD. When we add both UPEE and CMMD, ROUGE$_L$ is reduced by 0.51% and other indicators are the best results. On SUIM-IC dataset, BLEU-3 and ROUGE$_L$ have the best results when we only add UPEE. BLEU-2 has the best results when we only add CMMD. When we add both UPEE and CMMD, our method has the best results.

3) Important parameters: We conducted ablation studies on two important parameters in UPEE and CMMD on the UICM-SOFF dataset, where Tab X represents the number of additional scans in the decoder. It can be seen that the model achieves the best performance when the number of additional scans is 2. Tab XI represents the degree to which environmental keywords participate in guiding image features. When the weight of the environmental keyword score is 0.5, the performance is best.

F. Visualization Experiments

In this subsection, we verify the advantages of the proposed P^{2}UIC through visualization experiments, mainly including the results of generated text, the attention heatmaps of the encoder with or without UPEE, and the image region attention heatmaps during caption generation process, which are shown in Fig. 5, Fig. 6, Fig. 7, and Fig. 8, respectively.

To visually demonstrate the performance of the proposed P^{2}UIC in underwater image captioning, we compare the existing methods with the proposed P^{2}UIC on the UICM-SOFF dataset, as shown in Fig. 5. The results show that the proposed method generates more comprehensive and accurate underwater captions compared with MG-Transformer and UICM-SOFF. In the first sub-figure, MG-Transformer inaccurately inaccurately describes ”two divers” and mislabels the ”shipwreck” in the environment as a ”carrying” while UICM-SOFF accurately identifies the divers and shipwreck but fails to mention the school of fish in the background. The proposed P^{2}UIC can not only accurately describe the content contained in GT, but also more comprehensively describe the environment in the image. In the second sub-figure, MG-Transformer and UICM-SOFF lack attention to the details of the environment, only describing ”coral” but not describing the color of ”coral”, while the proposed method can accurately describe ”yellow coral”. These examples illustrate that the physics-aware feature enhancement and contrastive decoding framework of P^{2}UIC enable more holistic capture of underwater scene semantics, outperforming baselines in both object accuracy and descriptive richness.

To better demonstrate the role of proposed UPEE module in underwater environment modeling, we analyze the heatmaps

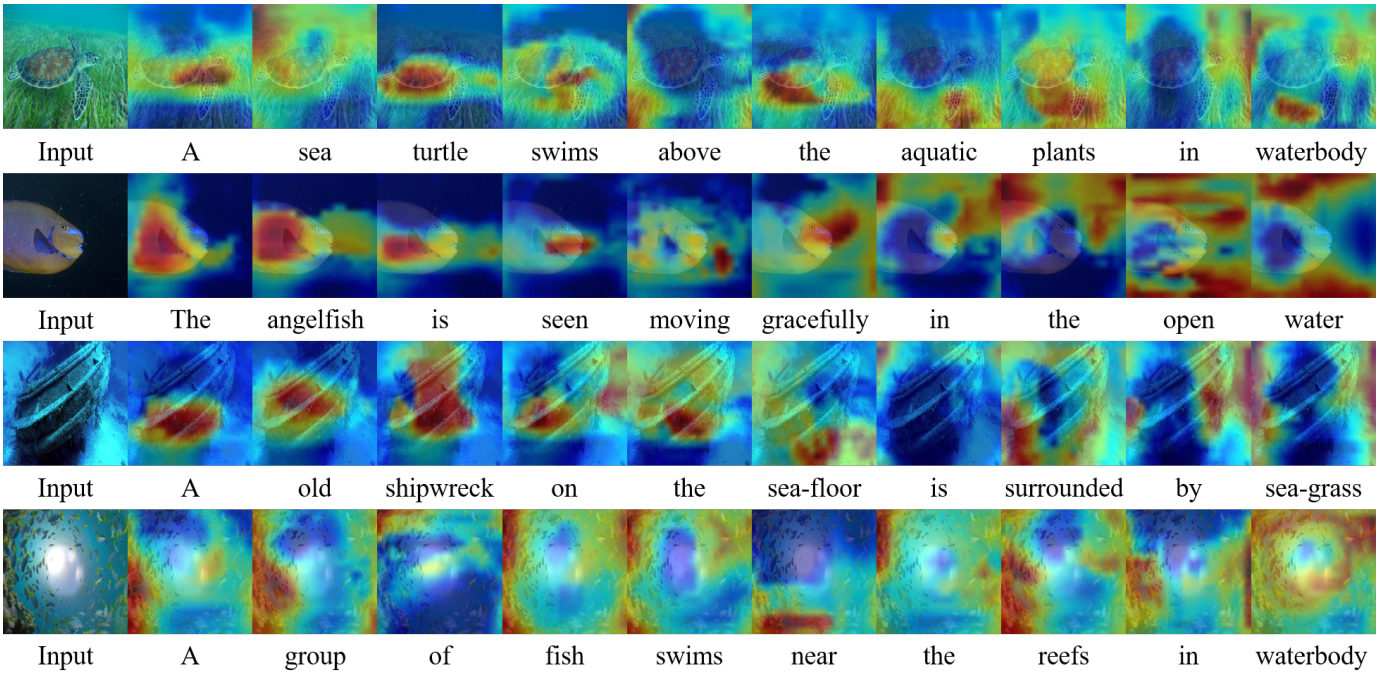

Fig. 8. Visualization of attented image regions along the caption generation processes for P^{2}UIC on SUIM-IC dataset, where red means higher degree of attention and blue means lower degree of attention.

of the P^{2}UIC encoding layers shown in Fig. 6. In the first sub-image, the model struggles to identify underwater targets due to color distortion and insufficient lighting inherent underwater image without UPEE. After UPEE simulation, the model accurately focuses on "diver", "big fish" and small fish in the background, showcasing improved target detection in visually degraded conditions. In the second sub-image, since the "diver" target is large, the model ignores environmental details like the "seafloor". By simulating physical parameters, UPEE can not only enable the model to balance focus on both prominent targets and subtle scene elements, but also enhance the model's dual capability to recognize underwater target.

In Fig. 7 and Fig. 8, we visualize the evolution of the image region attention heatmaps during caption generation process on the UWS-IC and SUIM-IC datasets, respectively, where red pixels denote higher attention weights. For the first row of images in Fig. 7, the caption "A crab is on a reef in the blue water" generated by P^{2}UIC not only accurately identifies the target "crab" but also emphasizes environmental elements "reef" and "blue water" as reflected by focused heatmap regions. In the second row in Fig. 7, the caption "The crocodile is in water close to some aquatic plants" accurately describes the location information of the target and the environment. For the third row in Fig. 8, the caption "A old shipwreck on the sea-floor is surrounded by sea-grass" accurately captures the interaction between the target and its environment. For the fourth row in Fig. 8, the caption "A group of fish swims near the reefs in waterbody" not only accurately locates the "group of fish" and "reefs", but also explains their relationship, providing a clear and detailed description of the underwater image. Overall, the attention heatmaps show that the underwater image captions generated by P^{2}UIC can not only effectively balance scene and target descriptions, but also produce linguistically fluent captions.

V. Conclusion

In this paper, we propose P^{2}UIC, an innovative Mamba-based framework for underwater image captioning. Firstly, we extend existing underwater image recognition datasets to create two expert-annotated UIC benchmarks: UWS-IC and SUIM-IC, which provide rich semantic descriptions of marine organisms and environments. Secondly, we propose two plug-and-play modules: UPEE and CMMD. UPEE enhances model sensitivity to underwater visual characteristics through physical degradation simulation and underwater key-words attention. CMMD improves cross-modal alignment by decoding different image sequences with NLP priors, generating linguistically precise captions that balance object localization and contextual detail. Extensive experiments on three UIC datasets donete that P^{2}UIC achieves SOTA performance across multiple metrics, outperforming Transformer and Mamba-based baselines in both quantitative scores and qualitative caption richness. Ablation studies demonstrate the effectiveness of the proposed components and visualization experiments validate that our method effectively captures fine-grained visual-semantic relationships and produce linguistically fluent captions. Our future research direction is to explore foundation model adaptation for underwater captioning and design a more elegant UIC pipeline to seamlessly generalize across diverse underwater scenarios.

References

[1] Z. Fu, W. Wang, Y. Huang, X. Ding, and K.-K. Ma, "Uncertainty inspired underwater image enhancement," in *ECCV*, 2022.

[2] Z. Wang, L. Shen, M. Xu, M. Yu, K. Wang, and Y. Lin, "Domain adaptation for underwater image enhancement," *IEEE Transactions on Image Processing*, 2023.

[3] S. Huang, K. Wang, H. Liu, J. Chen, and Y. Li, "Contrastive semi-supervised learning for underwater image restoration via reliable bank," in *CVPR*, 2023.

[4] Z. Fu, H. Lin, Y. Yang, S. Chai, L. Sun, Y. Huang, and X. Ding, "Unsupervised underwater image restoration: From a homology perspective," in *AAAI*, 2022.

[5] S. Mittal, S. Srivastava, and J. P. Jayanth, "A survey of deep learning techniques for underwater image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[6] I. Kabir, S. Shaurya, V. Maigur, N. Thakurdesai, M. Latnekar, M. Raunak, D. Crandall, and M. A. Reza, "Few-shot segmentation and semantic segmentation for underwater imagery," in *IROS*, 2023.

[7] S. Lian, H. Li, R. Cong, S. Li, W. Zhang, and S. Kwong, "Watermask: Instance segmentation for underwater imagery," in *ICCV*, 2023.

[8] J. Wang, Z. Wu, Y. Zhang, S. Kong, M. Tan, and J. Yu, "Integrated tracking control of an underwater bionic robot based on multimodal motions," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023.

[9] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *ICRA*, 2023.

[10] D. Singh, M. Kaur, J. M. Alanazi, A. A. AlZubi, and H.-N. Lee, "Efficient evolving deep ensemble medical image captioning network," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[11] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *ICCV*, 2017.

[12] H. Li, H. Wang, Y. Zhang, L. Li, and P. Ren, "Underwater image captioning: Challenges, models, and datasets," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025.

[13] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *NeurIPS*, 2021.

[14] A. Gupta, A. Gu, and J. Berant, "Diagonal state spaces are as effective as structured state spaces," *NeurIPS*, 2022.

[15] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and W. Xinggang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *ICML*, 2024.

[16] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[17] O. Lieber, B. Lenz, H. Bata, G. Cohen, J. Osin, I. Dalmedigos, E. Safahi, S. Meirom, Y. Belinkov, S. Shalev-Shwartz *et al.*, "Jamba: A hybrid transformer-mamba language model," *arXiv preprint arXiv:2403.19887*, 2024.

[18] J. Pilault, M. Fathi, O. Firat, C. Pal, P.-L. Bacon, and R. Goroshin, "Block-state transformers," *NeurIPS*, 2023.

[19] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar, "Semantic segmentation of underwater imagery: Dataset and benchmark," in *IROS*, 2020.

[20] R. E. Kalman, "A new approach to linear filtering and prediction problems," in *J. Basic Eng.*, 1960.

[21] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[22] A. Gu, I. Johnson, A. Timalsina, A. Rudra, and C. Ré, "How to train your hippo: State space models with generalized orthogonal basis projections," *arXiv preprint arXiv:2206.12037*, 2022.

[23] J. Wang, J. N. Yan, A. Gu, and A. M. Rush, "Pretraining without attention," *arXiv preprint arXiv:2212.10544*, 2022.

[24] H. Mehta, A. Gupta, A. Cutkosky, and B. Neyshabur, "Long range language modeling via gated state spaces," *arXiv preprint arXiv:2206.13947*, 2022.

[25] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré, "Hungry hungry hippos: Towards language modeling with state space models," *arXiv preprint arXiv:2212.14052*, 2022.

[26] Y. Bai and H. Cui, "A class sensitivity feature guided t-type generative model for noisy label classification," *Machine Learning*, 2024.

[27] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[28] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024.

[29] M. M. Islam, M. Hasan, K. S. Athrey, T. Braskich, and G. Bertasius, "Efficient movie scene detection using state-space transformers," in *CVPR*, 2023.

[30] M. M. Islam and G. Bertasius, "Long movie clip classification with state-space video models," in *ECCV*, 2022.

[31] E. Nguyen, K. Goel, A. Gu, G. Downs, P. Shah, T. Dao, S. Baccus, and C. Ré, "S4nd: Modeling images and videos as multidimensional signals with state spaces," *NeurIPS*, 2022.

[32] J. N. Yan, J. Gu, and A. M. Rush, "Diffusion models without attention," in *CVPR*, 2024.

[33] T. Huang, X. Pei, S. You, F. Wang, C. Qian, and C. Xu, "Localmamba: Visual state space model with windowed selective scan," *arXiv preprint arXiv:2403.09338*, 2024.

[34] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," *NeurIPS*, 2024.

[35] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[36] X. Zhu, L. Li, J. Liu, Z. Li, H. Peng, and X. Niu, "Image captioning with triple-attention and stack parallel lstm," *Neurocomputing*, 2018.

[37] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *ECCV*, 2018, pp. 684–699.

[38] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *CVPR*, 2016.

[39] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *ICCV*, 2019.

[40] N. Li, Z. Chen, and S. Liu, "Meta learning for image captioning," in *AAAI*, 2019.

[41] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," in *AAAI*, 2021.

[42] J. Zhang, Y. Xie, W. Ding, and Z. Wang, "Cross on cross attention: Deep fusion transformer for image captioning," *TCSVT*, 2023.

[43] Y. Li, J. Ji, X. Sun, Y. Zhou, Y. Luo, and R. Ji, "M3ixup: A multi-modal data augmentation approach for image captioning," *Pattern Recognition*, 2025.

[44] J. Zhang, K. Zhang, Y. Xie, and Z. Wang, "Deep reciprocal learning for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[45] L. Chen and K. Li, "Multi-modal graph aggregation transformer for image captioning," *Neural Networks*, 2025.

[46] X. Zhang, A. Jia, J. Ji, L. Qu, and Q. Ye, "Intra-and inter-head orthogonal attention for image captioning," *IEEE Transactions on Image Processing*, 2025.

[47] Y. Li, Y. Pan, T. Yao, and T. Mei, "Comprehending and ordering semantics for image captioning," in *CVPR*, 2022.

[48] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[49] C. Zhao, W. Cai, C. Dong, and C. Hu, "Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration," in *CVPR*, 2024.

[50] H. Wang, W. Zhang, L. Bai, and P. Ren, "Metalantis: A comprehensive underwater image enhancement framework," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[51] H. Wang, W. Zhang, and P. Ren, "Self-organized underwater image enhancement," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024.

[52] L. Meng, J. Wang, Y. Yang, and L. Xiao, "Prior knowledge-guided transformer for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[53] L. Meng, J. Wang, R. Meng, Y. Yang, and L. Xiao, "A multiscale grouping transformer with clip latents for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[54] L. Meng, J. Wang, Y. Huang, and L. Xiao, "Rsic-gmamba: A state space model with genetic operations for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[55] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *ECCV*, 2024.

[56] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *ICCV*, 2023.

[57] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo, "Promptcap: Prompt-guided image captioning for vqa with gpt-3," in *ICCV*, 2023.

[58] L. Li, Y. Wei, and P. Ren, "Underwater image captioning based on feature fusion," in *Proceedings of the 2024 7th International Conference on Image and Graphics Processing*, 2024.

[59] S. Kerai and G. Khekare, "Contextual embedding generation of underwater images using deep learning techniques," *IAES Int J Artif Intell (IJ-AI)*, 2024.

[60] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *CITS*, 2016.
[61] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, 2017.
[62] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *NeurIPS*, 2019.
[63] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word–sentence framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.
[64] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
[65] Y. Li, X. Zhang, J. Gu, C. Li, X. Wang, X. Tang, and L. Jiao, "Recurrent attention and semantic gate for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
[66] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.
[67] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002.
[68] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.
[69] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *ECCV*, 2016.
[70] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015.
[71] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

Chunlei Wang Chunlei Wang received the B.E degree in communication engineering from USTB, Beijing, China, in 2019, and the M.E. degree in information and communication engineering, USTB, Beijing, China, in 2022. He is currently pursuing the Ph.D. degree with the School of Electronic and Information Engineering, Beihang University, Beijing. His interests include image captioning, visual grounding and open vocabulary learning.

Wenquan Feng received Ph.D in communication and information system from Beihang University, Beijing, China. He is a professor and works in Beihang University. He is also the president of Qingdao Research Institute of Beihang University. Recently, He has been working on using artificial intelligence techniques to handle image data from satellites. His research interests include pattern recognition, computer vision and multi-modality image processing.

Binghao Liu received B.E. degree in electronics and information engineering from Beihang University, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Electronicand Information Engineering, Beihang University, Beijing. His research interests include few-shot learning, remote sensing and image captioning.

Xianyu Zhao received the B.E degree in measurement and control technology and Instruments from China University of Geosciences, Wuhan, China, in 2015, and the Ph.D. degree in instrument science and technology from Tianjin University, Tianjin, China, in 2020. He is currently studing photophysics in Changchun University of Science and Technology. His research interests include image captioning, object detection and multimedia analysis.

Kejun Zhao received the B.E degree in Optoelectronics Technology from Zhengzhou University, Zhengzhou, China, in 2001, and the M.E. degree in optical engineering from Tianjin University, Tianjin, China, in 2018. He is currently studing photophysics in Luoyang Institute of Electro-Optical Equipment, AVIC. His research interests include object detection, underwater image recognition and natural language processing.

Qi Zhao received Ph.D in communication and information system from Beihang University, Beijing, China. She is a professor and works in Beihang University. She was in the Department of Electrical and Computer Engineering at the University of Pittsburgh as a visiting scholar from 2014 to 2015. Since 2016, she has been working on wearable device based first-view image processing and deep learning based image recognition. Her current research interests include underwater image segmentation, multimedia analysis and object detection.