Highlights

UMDM-USG: A Unified Multi-view Diffusion Model for Underwater Scene Generation via Cross-View Representation Alignment

Yifan Zhu, Chengjia Wang, Xinghui Dong

- Constructing the Underwater Text–Mask–Depth–Normal Dataset (U-TMDN), containing 53,403 images with textual descriptions, semantic segmentation masks, depth maps, and surface normal maps.
- Proposing a A Unified Multi-view Diffusion Model for Underwater Scene Generation via Cross-View Representation Alignment (UMDM-USG), jointly generating images and multiple key modalities conditioned on expressive, domain-specific textual descriptions.
- Designing the Conditional Generation Alignment Attention (CGA-Attn), which explicitly models bidirectional cross-view interactions, enabling accurate structural grounding and improved multi-view alignment.

UMDM-USG: A Unified Multi-view Diffusion Model for Underwater Scene Generation via Cross-View Representation Alignment

Yifan Zhu[a], Chengjia Wang[b], Xinghui Dong[a,*]

[a] *State Key Laboratory of Physical Oceanography and the Faculty of Information Science and Engineering, Ocean University of China, 238 Songling Road, Qingdao, 266100, Shandong, China*
[b] *School of Mathematical and Computer Sciences, Heriot-Watt University, Riccarton Mains Road, EH14 4AS, Edinburgh, United Kingdom*

Abstract

Underwater scene understanding is crucial for marine exploration, ecological monitoring, and robotic operations, yet the scarcity of large-scale, high-quality underwater datasets severely limits the performance of learning-based models. In this paper, we propose UMDM-USG, a Unified Multi-view Diffusion Model for Underwater Scene Generation based on cross-view representation alignment. By treating each modality, such as image, segmentation mask, depth map, and surface normal, as a distinct view of the same underwater scene, UMDM-USG jointly generates coherent multi-view outputs conditioned on expressive textual descriptions. To strengthen cross-view coherence, we introduce a Conditional Generation Alignment Attention (CGA-Attn) mechanism that explicitly enhances semantic and geometric alignment across views. In addition, we construct the U-TMDN dataset, consisting of 53,403 underwater images with comprehensive multi-view annotations. Extensive experiments demonstrate that UMDM-USG achieves superior or competitive performance in image quality and semantic consistency, and that the generated multi-view data consistently improve multiple downstream underwater vision tasks.

Keywords: Multi-view Learning; Cross-View Representation Alignment; Underwater

⋆Data set, source code and models are available at https://github.com/INDTLab/UMDM-USG.

*Corresponding author

Email addresses: zhuyifan@stu.ouc.edu.cn (Yifan Zhu), chengjia.wang@hw.ac.uk (Chengjia Wang), xinghui.dong@ouc.edu.cn (Xinghui Dong)

1. Introduction

Underwater scene understanding plays an important role in marine exploration [1, 2], robotic operations [3, 4] and underwater autonomous navigation [5, 6, 7, 8]. However, the acquisition of underwater datasets is typically costly and technically challenging due to the harsh and complex marine environment. Moreover, underwater imaging commonly suffers from severe light absorption, significant color distortion, strong scattering effects, and highly complex geometric structures. These factors further increase the difficulty of obtaining high-quality annotated data. As a result, data scarcity substantially constrains the training of underwater vision models, preventing them from fully learning the visual and geometric characteristics of complex underwater scenes. Since appearance, semantics, and geometry are tightly coupled in underwater environments, generating a single modality in isolation cannot capture these interdependencies. A principled solution must therefore jointly model multiple complementary views of the same scene within a unified framework [9, 10], a challenge that naturally connects to multi-view learning.

In recent years, generative models have emerged as a promising avenue for data augmentation, offering a viable strategy to mitigate the scarcity of training data in underwater vision tasks. For example, Atlantis [11] (as illustrated in Fig. 1(a)) generates underwater images and depth data conditioned on ground-scene depth maps. Although this approach improves geometric controllability, it relies on ground-scene depth conditions that differ considerably from real underwater scenes in terms of physical imaging properties and structural distributions. In contrast, TIDE [12] (Fig. 1(b)) independently aligns visual and dense annotation modalities to improve structural modeling capabilities. However, this method requires multiple independent alignment modules for different modalities, resulting in a relatively complex overall architecture. Moreover, the lack of a unified and collaborative alignment mechanism across modalities limits its scalability in multimodal joint generation scenarios. In addition, existing methods generally adopt simplified textual conditions, failing to incorporate domain-

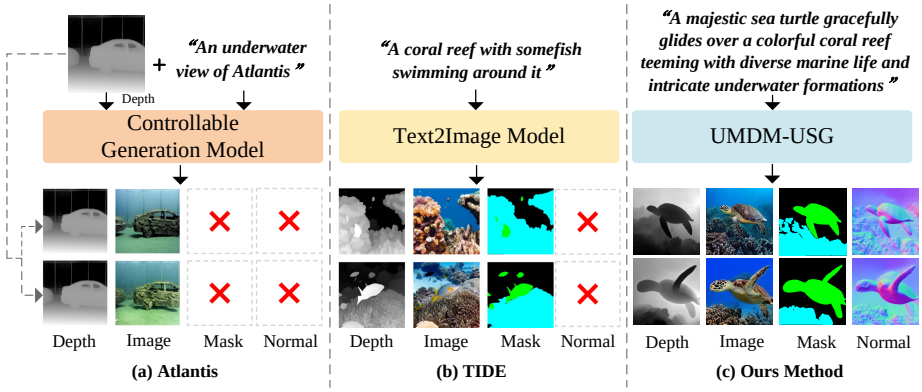

Figure 1: Illustration of different underwater scene generation methods. (a) Atlantis [11]: generation conditioned on text and depth maps as external control signals; (b) TIDE [12]: joint generation of images and dense annotations via alignment across multiple independent modules; (c) the proposed UMDM-USG: unified multi-view generation with improved structural consistency.

specific underwater terminology, which restricts their ability to provide fine-grained and physically meaningful guidance during generation.

To address these limitations, we propose a Unified Multi-view Diffusion Model for Underwater Scene Generation via Cross-View Representation Alignment, referred to as UMDM-USG. As shown in Fig. 1(c) and Fig. 2, UMDM-USG simultaneously generates underwater images and their corresponding key modalities, including semantic segmentation maps, depth maps, and surface normal maps, conditioned on expressive and domain-specific textual descriptions. By treating each modality as a distinct view of the same underwater scene, UMDM-USG introduces a Conditional Generation Alignment Attention (CGA-Attn) mechanism that explicitly strengthens cross-view structural and semantic alignment. Compared with existing approaches that require independent modality-specific modules, UMDM-USG avoids such complex modular designs and achieves more concise and coherent joint generation.

In addition, we construct a comprehensive underwater multi-view dataset, termed the Underwater Text–Mask–Depth–Normal Dataset, or U-TMDN for short. It contains 53,403 underwater images, each paired with high-quality text descriptions, semantic masks, depth maps, and surface normals. The dataset provides a solid foundation for training and evaluating unified multi-view generation models, and further supports the

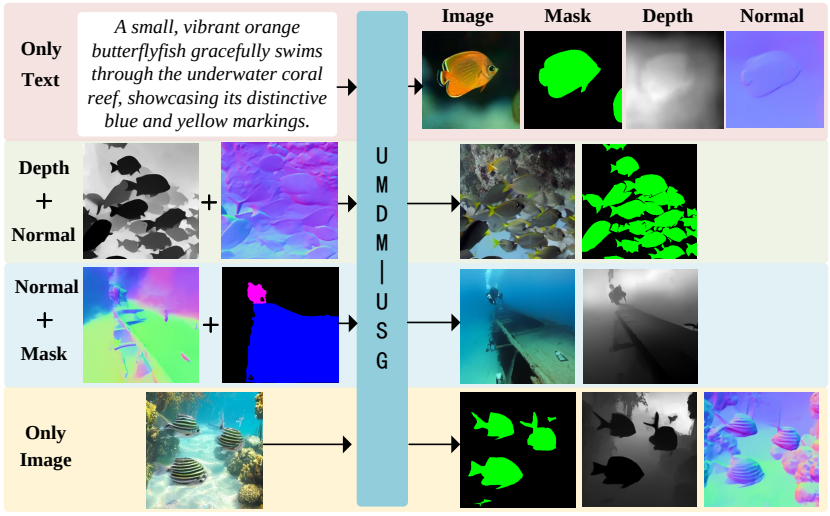

Figure 2: Visualization of the results generated by UMDM-USG. The model can generate corresponding multi-view outputs from textual descriptions, and can also perform joint inference to generate missing modalities given a combination of input modalities.

use of generated data to enhance downstream underwater vision tasks.

The contributions of this research can be summarized as threefold.

- We construct the Underwater Text–Mask–Depth–Normal Dataset (U-TMDN). By integrating textual descriptions, semantic annotations, and geometric information, U-TMDN provides unified data support for both underwater multi-view generation and downstream vision research.
- We propose the Unified Multi-view Diffusion Model for Underwater Scene Generation via Cross-View Representation Alignment (UMDM-USG), featuring a Conditional Generation Alignment Attention (CGA-Attn) mechanism. This module explicitly aligns different views to ensure high structural consistency and generation quality.
- Extensive experiments on multiple representative downstream underwater vision tasks demonstrate that the multi-view data generated by UMDM-USG can effectively enhance the performance of downstream underwater tasks.

The rest of this paper is organized as follows. In Section 2, we review the related

literature. Our Underwater Text-Mask-Depth-Normal Dataset is introduced in Section 3. In Section 4, the proposed UMDM-USG is presented. The experimental setup, results, and downstream task evaluations are presented in Sections 5 and 6. Finally, we draw our conclusion in Section 7.

2. Related Work

2.1. Underwater Image Generation

Underwater image generation has gained significant attention due to its importance in marine perception, robotic exploration, and environmental monitoring. Early studies primarily focused on image enhancement and restoration to alleviate underwater degradations, such as color cast, light attenuation, and scattering effects [13, 6, 14].

Recently, research on underwater scene generation has gradually incorporated diffusion models and structural priors to constrain the synthesis process. Atlantis [11] introduces ControlNet into the Stable Diffusion framework and employs depth maps as external control signals to guide underwater image generation and depth estimation, thereby improving geometric consistency. However, this approach relies on plug-in control branches, and the interaction between depth information and the image generation process remains relatively indirect.

TIDE [12] proposes a framework for underwater image generation and dense annotation synthesis. It employs two fine-tuned, lightweight Transformer modules to separately align each modality, which improves structural expressiveness. Nevertheless, this design requires independent alignment components for different modalities, leading to a complex architecture that lacks a unified mechanism to jointly model cross-modal interactions.

In contrast, our UMDM-USG adopts a unified alignment mechanism to jointly model interactions among heterogeneous modalities, enabling the generation of more coherent underwater scenes.

2.2. Underwater Datasets

The development of underwater vision models is highly dependent on the availability of large-scale and high-quality datasets. Early underwater datasets mainly focus

on specific tasks, such as image enhancement [14], object detection [4], or semantic segmentation [15]. While these datasets provide valuable benchmarks, they are usually limited to a single modality and lack comprehensive semantic or geometric annotations.

Several recent datasets attempt to enrich underwater data with additional information. For example, Sea-thru [6] and SeathruNeRF [7] explore underwater image formation and geometry-aware modeling, but their scales remain relatively small and their annotations are task-specific. MarineInst [3] introduces instance-level visual descriptions for marine imagery, providing richer semantic annotations.

Atlantis [11] is one of the first efforts to incorporate depth supervision into underwater generative modeling, while TIDE [12] further introduces segmentation masks as additional structural conditions; however, both approaches are constrained by relatively limited dataset scale. As a result, existing underwater datasets still lack unified and large-scale multi-view annotations that jointly capture semantic, structural, and geometric properties of underwater scenes.

To address these limitations, we build a comprehensive underwater multi-view dataset containing 53,403 images, each paired with a textual description, segmentation mask, depth map, and surface normal, enabling joint multi-view learning and generation.

2.3. Multi-view Representation Learning

Multi-view learning aims to exploit the complementary information across different data representations or modalities, describing the same underlying objects [9]. Classical approaches learn shared representations by maximizing correlations between views. More recent methods used contrastive learning [16, 17] to align and fuse information from heterogeneous sources. In [18], multi-view knowledge fusion was further introduced to enhance representation learning by explicitly modeling complementary information across different views.

In the context of scene understanding, different modalities—such as RGB images, depth maps, segmentation masks, and surface normals—can be naturally viewed as distinct views of the same physical scene, each encoding complementary appearance, semantic, or geometric properties. While multi-view learning has been extensively

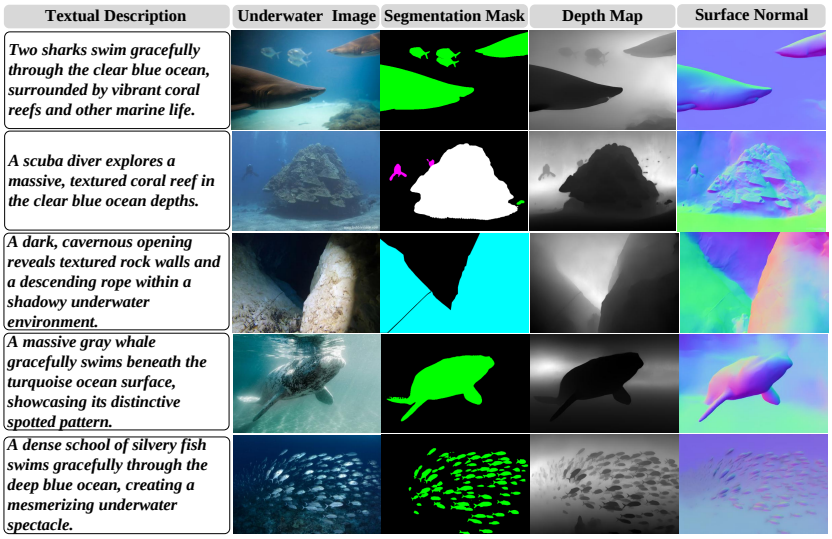

Figure 3: Examples from the proposed U-TMDN dataset, including underwater images paired with textual descriptions, segmentation masks, depth maps, and surface normals.

studied for discriminative tasks such as classification and clustering, its integration with generative frameworks remains relatively underexplored. Our work bridges this gap by formulating multimodal underwater scene generation as a multi-view learning problem and introducing explicit cross-view alignment within a unified diffusion architecture.

3. Underwater Text–Mask–Depth–Normal Dataset

This study focuses on multi-view underwater scene generation, which requires the joint modeling of semantic, structural, and geometric information. The existing Atlantis [11] dataset has notable deficiencies in both scale and modal diversity. To address this limitation, we construct a more comprehensive underwater multi-view dataset containing 53,403 images, each paired with a textual description, segmentation mask, depth map, and surface normal (see Fig. 3 for examples). We refer to this dataset as Underwater Text–Mask–Depth–Normal Dataset, or U-TMDN for short. In the following, we describe the construction pipeline of each modality.

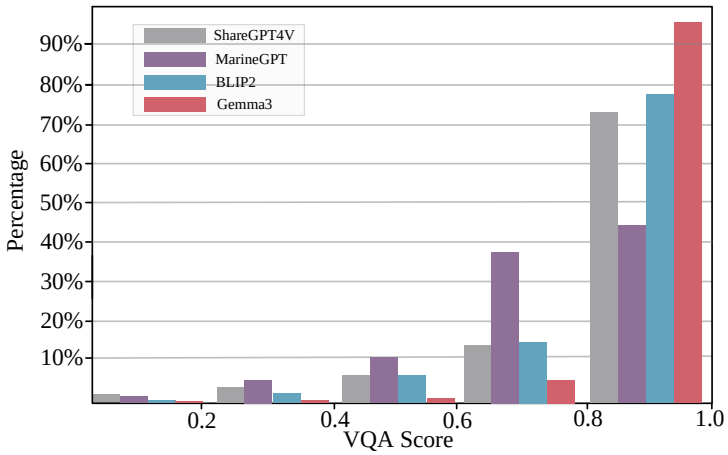

Figure 4: Comparison of VQA Scores [23] for captions generated by three baseline models and our fused caption.

3.1. Image Source

The images in U-TMDN are integrated from multiple high-quality public sources, including CaveSeg [19], EUVP [13], FLSea [5], HICRD [20], LUIQD [21], MarineInst [3], Roboflow [1], RUIE [8], RUOD [4], Sea-thru (D1 and D2) [6], SeathruNeRF [7], SUIM [15], UIEB [14], UIIS [1], and USIS [2]. To ensure diversity and validity, we calculated the Structural Similarity Index (SSIM) [22] for all candidate images and removed duplicates with excessive overlap. This process yielded a total of 53,403 underwater images.

3.2. Text Description Generation

For each image, we generate three candidate captions using BLIP2 [24], ShareGPT4V [25], and MarineGPT [26]. These initial descriptions are then refined by the Gemma3 [27]. Specifically, Gemma3 [27] is prompted to merge, optimize, and deduplicate the candidates to produce a single, semantically accurate caption that precisely describes the underwater scene.

Figure 4 reports the Visual-Question-Answering (VQA) Score [23] for individual

[1] https://universe.roboflow.com/datathon-ce18l/underwater-imagery

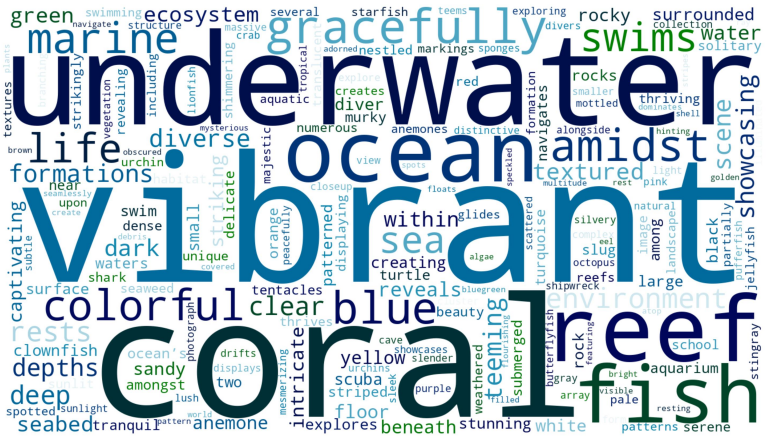

Figure 5: Word cloud visualization of all refined textual descriptions in U-TMDN.

and fused captions. The fused version consistently outperforms all single models, validating our refinement strategy. Furthermore, a word cloud of all refined descriptions (see Fig. 5) highlights the global textual distribution, confirming that U-TMDN provides rich, semantically coherent annotations for multi-view training.

3.3. Segmentation Mask Construction

To construct high-quality segmentation masks for U-TMDN, we integrate annotations from three complementary sources, as shown in Fig. 6(a). In total, three different strategies are used, as detailed below.

1) For datasets with ground-truth masks [19, 15, 1, 2], we directly adopt their pixel-level annotations. 2) For datasets containing only instance masks (e.g., MarineInst [3]), we employ a CLIP-based similarity strategy [28] for semantic labeling. Specifically, each cropped instance region is compared against the text embeddings of the seven SUIM [15] categories, with the highest similarity score determining the final label. 3) For unlabeled raw images, we fuse predictions from three underwater-specific models, including WaterMask [1], WaterSAM [2], and SPDA-SAM [29]. Given that general segmentation models often struggle with underwater light attenuation and scattering, we employ these specialized models and follow a fusion strategy to generate high-quality pseudo masks, substantially reducing the need for manual annotation.

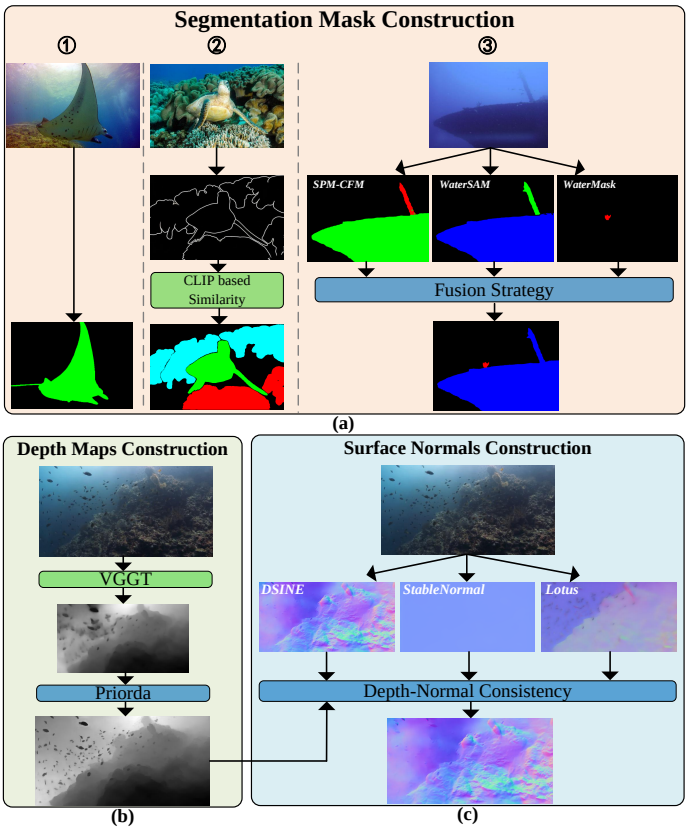

Figure 6: Overview of the construction pipeline for underwater segmentation masks, depth maps, and surface normals.

In summary, the final mask set comprises 28,349 manually annotated masks and 25,054 masks generated using our multi-model fusion strategy. This design ensures U-TMDN covers both expert-validated semantic knowledge and the broad scenario coverage of cross-model predictions, avoiding overfitting to the distribution of individual segmentation models and providing a more comprehensive and authentic semantic foundation for multi-modal feature alignment and fusion.

3.4. Depth Maps Construction

Following the coarse-to-fine pipeline of Prior Depth Anything (Priorda) [30], we estimate depth maps in two stages. First, the Visual Geometry Grounded Transformer

(VGGT) [31] generates an initial relative depth map and a spatial confidence map. To mitigate noise and blurred boundaries caused by underwater scattering, we then employ Priorda [30] to refine these results. By using the confidence map as a guide, this process preserves reliable depth values while correcting ambiguous regions, yielding cleaner and more consistent geometry for multimodal modeling. The overall pipeline is illustrated in Fig. 6(b).

3.5. Surface Normals Construction

To generate reliable surface normal maps, we adopt a selection-based fusion strategy involving three advanced estimation models, including StableNormal [32], DSINE [33], and Lotus [34]. To ensure geometric fidelity, we evaluate each candidate prediction using a depth-normal consistency metric. The prediction with the highest consistency score is selected as the final normal map. The overall pipeline is illustrated in Fig. 6(c).

4. Proposed Method

The proposed Unified Multi-view Diffusion Model for Underwater Scene Generation via Cross-View Representation Alignment (UMDM-USG) is built upon the Sana framework [35]. Conceptually, UMDM-USG treats each target modality as a distinct view of the same underwater scene and learns to align these views in a shared latent space during the diffusion process. Considering the strong structural correlations and significant scale discrepancies among multi-view data in underwater scenes, we extend Sana by designing a Multi-view Alignment Module and introducing a representation alignment regularization [36]. Unlike diffusion models that focus on a single modality or rely on auxiliary plug-in controls [11, 12], UMDM-USG preserves the inherent stability and uniformity of the diffusion process while enhancing its capability to represent and align complex structural information across views.

As illustrated in Fig. 7, the overall architecture of UMDM-USG consists of four key components, including a deep compression autoencoder [35], a pre-trained text encoder, a Role Assignment Module (RAM), and our proposed Multi-view Alignment

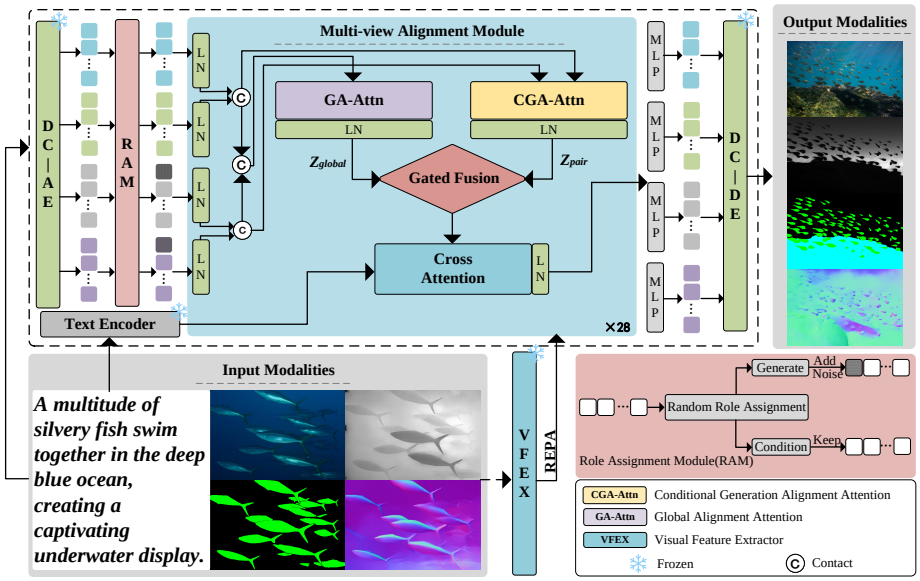

Figure 7: Overall architecture of the Unified Multi-view Diffusion Model for Underwater Scene Generation via Cross-View Representation Alignment (UMDM-USG).

Modules. In the following subsections, we present a detailed description of the core design principles and implementation details of the proposed model.

4.1. Deep Compression Autoencoder

To reduce the computational cost and improve generation efficiency, we employ a deep compressed autoencoder [35]. This module maps high-dimensional images into a compact latent space, achieving efficient feature compression while preserving critical structural and semantic information. Compared with conventional autoencoders, this approach reduces the computational burden and memory consumption during both the training and inference processes.

4.2. Pre-trained Text Encoder

To provide expressive and semantically rich language guidance for underwater scene generation, we adopt the pre-trained Gemma3 [27] as the text encoder in UMDM-USG. Given an input underwater textual description, the text encoder first tokenizes the

sequence and encodes it into a set of contextualized text embeddings. The resulting text representation is denoted as

$$\mathbf{z}^{(t)} \in \mathbb{R}^{L_t \times C}, \tag{1}$$

where L_t denotes the length of the token sequence, and C is the embedding dimension which is aligned with the latent feature dimension used in the diffusion backbone. During the training process, the parameters of the text encoder are frozen.

4.3. Role Assignment Module

To enable unified modeling and flexible generation across multiple modalities, we introduce and extend the Role Assignment Module (RAM) [37]. Unlike static frameworks, the RAM avoids predefining fixed input–output relationships. Instead, it employs a random role assignment strategy, training the model to generate arbitrary target modalities conditioned on any subset of the remaining ones.

As shown in Fig. 7, data from various modalities are first mapped into a unified latent space via the deep compressed autoencoder [35] and represented as sequences. Let the latent representation of the m-th modality be denoted as

$$\mathbf{z}^{(m)} \in \mathbb{R}^{L_m \times C}, \tag{2}$$

where L_m denotes the sequence length of modality m, and C is the latent feature dimension. During the training process, the RAM randomly partitions the modality set $\mathbf{z}^{(m)}$ at each iteration into a generation modality set M_g and a condition modality set M_c, satisfying

$$M_g \cup M_c = M, \quad M_g \cap M_c = \emptyset, \tag{3}$$

where M denotes the set of all available modalities.

The intersection between M_g and M_c is constrained to be an empty set because the two modality sets serve different purposes during diffusion learning. For modalities assigned to the generation modality set M_g, which serve as the prediction targets of the diffusion model, noise is injected following the standard diffusion formulation at timestep t:

$$\tilde{\mathbf{z}}_t^{(m)} = \sqrt{\alpha_t}\,\mathbf{z}^{(m)} + \sqrt{1-\alpha_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{4}$$

where α_t is the predefined noise scheduling coefficient. These modalities serve as the prediction targets of the diffusion model. In contrast, latent representations belonging to the condition modality set M_c are directly sent to the diffusion network as conditioning signals, offering semantic or structural priors to guide the generation process and therefore remain unchanged, which can be expressed as:

$$\tilde{\mathbf{z}}_t^{(m)} = \mathbf{z}^{(m)}, \quad m \in M_c. \tag{5}$$

The RAM exposes the model to diverse modality combinations during the training process. Consequently, at inference time, UMDM-USG inherently supports arbitrary modality-to-modality generation conditioned on any available subset. This unified mechanism eliminates the need for modality-specific pathways or specialized alignment modules, simplifying system architecture while enhancing the generalization of the model across complex multi-view tasks.

4.4. Multi-view Alignment Module

As one of the core components of the proposed UMDM-USG, the primary objective of the Multi-view Alignment Module is to efficiently capture structural correlations and semantic complementarity across different views of the same scene within a unified diffusion modeling framework. This module is composed of four key components, including the Global Alignment Attention (GA-Attn), Conditional Generation Alignment Attention (CGA-Attn), Gated Fusion, and Cross Attention mechanisms. Specifically, GA-Attn and CGA-Attn extract macro-scale global correlations and micro-scale cross-view structural priors, respectively. These complementary features are subsequently integrated via the Gated Fusion mechanism before being injected with textual semantics through the Cross Attention mechanism.

4.4.1. Global Alignment Attention

To effectively model long-range cross-view dependencies under multi-view conditions while controlling computational complexity, the Global Alignment Attention (GA-Attn) mechanism is built upon the linear self-attention mechanism adopted by

Sana [35]. Specifically, let the layer-normalized latent representation of the m-th modality be

$$\hat{\mathbf{z}}^{(m)} \in \mathbb{R}^{N\times C}, \tag{6}$$

where N denotes the number of tokens for a single modality and C represents the feature dimension. By concatenating all modalities along the token dimension, we obtain a global sequence representation:

$$\hat{\mathbf{Z}} = \mathrm{Concat}\left(\hat{\mathbf{z}}^{(1)}, \hat{\mathbf{z}}^{(2)}, \ldots, \hat{\mathbf{z}}^{(L)}\right) \in \mathbb{R}^{(LN)\times C}, \tag{7}$$

which is then processed by linear self-attention [35] to model global dependencies:

$$\mathbf{Z}_{\mathrm{global}} = \mathrm{LA}(\hat{\mathbf{Z}}), \tag{8}$$

where $\mathrm{LA}(\cdot)$ denotes the linear attention operation.

Compared with the standard Transformer, the time complexity is reduced from $O((LN)^2C)$ to $O(LN \cdot C^2)$, and the memory complexity is reduced from $O((LN)^2)$ to $O(LN \cdot C)$, where L denotes the number of modalities, N is the number of tokens per modality, and C represents the feature dimension. This optimization yields substantial gains in computational efficiency, particularly when scaling to multiple modalities and high-resolution latent representations.

Despite its efficiency, relying solely on global alignment attention across concatenated sequences is insufficient for multi-view underwater generation. Such an approach primarily captures coarse statistical correlations and tends to be dominated by large-scale regions, limiting its capacity to model directed dependencies between conditioning and generation sets. Moreover, fine-grained structures, small objects, and localized cross-view correspondences are difficult to preserve under purely global interactions. Consequently, rather than relying on diluted global context, generation views require explicitly targeted structural guidance from conditioning views.

4.4.2. Conditional Generation Alignment Attention

To address these limitations, we design the Conditional Generation Alignment Attention (CGA-Attn) mechanism. By explicitly constructing bidirectional cross-view attention between generation views and conditioning views, this design enables the

model to preserve global consistency while selectively strengthening the mutual synergy between the evolving content and its structural priors. This targeted interaction ensures that the generated views are precisely anchored to the conditioning information, enhancing the quality of cross-view alignment.

Specifically, according to the output of the RAM, the modality set is partitioned into the generation modality set M_g and the conditioning modality set M_c. For any generation modality $m \in M_g$, its latent representation is used as the query, while the conditioning modalities serve as keys and values to construct conditional generation alignment attention:

$$\mathbf{Z}_{g\leftarrow c} = \mathrm{Attn}\left(\hat{\mathbf{Z}}_g, \hat{\mathbf{Z}}_c\right), \tag{9}$$

where $\hat{\mathbf{Z}}_g$ and $\hat{\mathbf{Z}}_c$ denote the latent representations of the generation modality set M_g and the conditioning modality set M_c, respectively.

Meanwhile, to maintain consistency of the conditioning modalities during joint modeling, a reverse update is further introduced:

$$\mathbf{Z}_{c\leftarrow g} = \mathrm{Attn}\left(\hat{\mathbf{Z}}_c, \hat{\mathbf{Z}}_g\right). \tag{10}$$

After obtaining the bidirectional aligned representations between the generation modalities and the conditioning modalities, we further perform residual fusion to construct the final generation-conditioned joint alignment representation:

$$\mathbf{Z}_{\text{pair}} = \hat{\mathbf{Z}}_g + \lambda_1 \mathbf{Z}_{g\leftarrow c} + \lambda_2 \mathbf{Z}_{c\leftarrow g}, \tag{11}$$

where λ_1 and λ_2 are learnable weighting coefficients.

4.4.3. Gated Fusion

After obtaining the global alignment features $\mathbf{Z}_{\text{global}}$ and the generation-conditioned alignment features $\mathbf{Z}_{\text{pair}}$, we introduce a gated fusion mechanism to dynamically balance the relative importance of the global alignment branch and the generation-conditioned alignment branch.

Specifically, modality-level gating vectors are generated using the modulation parameters from layer normalization:

$$\mathbf{g}^{(m)} = \sigma\left(\mathbf{W}_g\left(\mathbf{s}^{(m)} + \mathbf{b}^{(m)}\right)\right), \quad \mathbf{g}^{(m)} \in \mathbb{R}^2, \tag{12}$$

where $\mathbf{s}^{(m)}$ and $\mathbf{b}^{(m)}$ denote the scale and bias parameters of the corresponding modality, respectively, and $\sigma(\cdot)$ is the Sigmoid function. The resulting gating coefficients are then used to perform weighted fusion of the two aligned representations:

$$\mathbf{Z}_{\text{fused}} = g_{\text{pair}} \cdot \mathbf{Z}_{\text{pair}} + g_{\text{global}} \cdot \mathbf{Z}_{\text{global}}. \tag{13}$$

This mechanism enables the model to adaptively emphasize either local cross-view alignment or global contextual modeling according to the current generation conditions, achieving a better balance between global structural consistency and fine-grained detail preservation in complex underwater scenes.

4.4.4. Cross-Attention

After completing internal multi-view alignment, we introduce the cross attention mechanism to inject textual semantic information into the fused multi-view latent representations.

Specifically, let the output of the text encoder be denoted as $\mathbf{Y}$, where the Gemma3 [27] is employed in our implementation. Given the fused multi-view latent representation $\mathbf{Z}_{\text{fused}}$, semantic alignment is achieved through a cross-attention operation:

$$\mathbf{Z}_{\text{text}} = \text{CrossAttn}\left(\mathbf{Z}_{\text{fused}}, \mathbf{Y}\right). \tag{14}$$

4.5. Loss Function

The training objective of UMDM-USG comprises a Mean Squared Error (MSE) term and a Representation Alignment Regularization (REPA) [36] term. The former ensures the accuracy of the generated results in either pixel space or latent space, while the latter imposes constraints at the representation level, guiding the model to learn intermediate representations that better conform to the statistical properties of underwater scenes.

4.5.1. Mean Squared Error

During the diffusion process, given the ground-truth multi-view observation x_0 and timestep t, Gaussian noise is injected via the forward process:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{15}$$

where $\bar{\alpha}_t$ denotes the noise scheduling coefficient. The model is trained to predict either the noise or the original sample using the MSE, and the basic training objective is formulated as

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{x_0,\epsilon,t}\left[|\epsilon - \epsilon_\theta(x_t, t, C)|_2^2\right], \tag{16}$$

where C denotes the conditional information composed of text and other conditional modalities.

4.5.2. Representation Alignment Regularization

Relying solely on the diffusion reconstruction loss may lead to slow convergence in the early stages of training and structural instability in the generated results. To address this issue, we use the Representation Alignment Regularization (REPA) [36]. This term imposes constraints on the intermediate hidden states, encouraging the diffusion features to align with representations that better capture the inherent characteristics of underwater scenes.

Since UMDM-USG generates multiple modalities simultaneously, this alignment is applied not only to RGB images but across all generated views. From a multi-view learning perspective, our model jointly learns a unified representation of underwater scenes across multiple structure-related views. Specifically, images, segmentation maps, depth maps, and normal maps are treated as different projections of the same underwater scene in the appearance, semantic, and geometric subspaces. We denote these modalities in a unified form as

$$I = \{I^{(m)} \mid m \in \text{Img}, \text{Mask}, \text{Depth}, \text{Normal}\}. \tag{17}$$

For each modality $I^{(m)}$, we employ a pre-trained visual feature extractor $f_d(\cdot)$ to obtain its patch-level representations $\mathbf{f}_d(I^{(m)})$. Meanwhile, we extract the hidden state $h_t^{(m)}$ from the m-th layer of the Multi-view Alignment Module within UMDM-USG. This hidden state is then mapped into the same representation space as the pre-trained features via a projection head $\phi(\cdot)$, yielding:

$$\mathbf{h}_\phi^{(m)} = \phi(h_t^{(m)}). \tag{18}$$

REPA [36] is defined by maximizing the patch-wise similarity between the model's hidden representations and the features extracted from real observations. The corresponding loss function is formulated as:

$$\mathcal{L}_{\text{proj}} = \frac{1}{|M|} \sum_{m \in M} \frac{1}{N} \sum_{i=1}^{N} \left| \mathbf{h}_{\phi}^{(m,i)} - \mathbf{f}_{d}^{(m,i)}(I^{(m)}) \right|_2^2, \tag{19}$$

where M denotes the set of modalities and N is the number of patches per modality. Through this unified representation alignment, the model not only learns the noise denoising process during the training process, but also explicitly captures the statistical properties of underwater scenes in both semantic and geometric structures, improving overall consistency, geometric plausibility, and cross-view coherence in multi-view generation results.

4.5.3. Overall Loss Function

Finally, the overall training objective of UMDM-USG is defined as a weighted sum of the MSE and the REPA terms:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{proj}}, \tag{20}$$

where λ is a weighting coefficient that balances generation accuracy and the strength of representation alignment.

5. Experiments

This section presents the experimental setup, comparative evaluations with existing methods, and ablation studies to validate the effectiveness of our UMDM-USG.

5.1. Experimental Setup

During the training stage, we consistently used the U-TMDN and split it according to data sources, with two-thirds of the data allocated for training. The resulting training set contained 35,484 underwater images along with their corresponding multi-view annotations.

We fine-tuned UMDM-USG based on the Sana-600M [35] pre-trained model, leveraging its robust text–image alignment to enhance the stability and semantic consistency

Table 1: Comparison of different generative models on the U-TMDN in image generation quality and semantic consistency. For each metric, the best and second best results are highlighted in **Red** and *Cyan* fonts, respectively.

Method	FID ↓	NIQE ↓	VQA Score ↑	HPSv2.1 ↑
SD3-Medium [38]	31.2	5.71	0.58	**26.92**
SD-XL [39]	23.0	4.70	0.58	23.87
Pixart-σ [40]	15.2	5.57	0.54	23.23
Infinity [41]	33.3	5.17	**0.64**	24.38
Jodi [37]	*9.99*	4.26	0.49	22.87
Sana-1600M [35]	10.9	4.37	0.49	22.67
TG-TSGNet [42]	18.7	4.69	0.47	22.40
Atlantis [11]	18.0	4.68	0.52	21.93
TIDE [12]	12.2	**3.92**	0.47	22.27
UMDM-USG	**6.15**	*4.13*	*0.60*	*24.79*

of downstream generation. Training was conducted for 100 epochs using the CAME-8bit optimizer with an initial learning rate of 4×10^{-5} and a batch size of 4, with FP16 mixed-precision training to reduce memory usage. All experiments were implemented on a single NVIDIA L40 GPU. For a fair comparison, all baseline generative models were also fine-tuned on the same U-TMDN training split under their respective default training configurations.

5.2. Comparative Experiments

We conducted comparative experiments with six representative generative models, including SD3-Medium [38], SD-XL [39], PixArt-σ [40], Infinity [41], Jodi [37], TG-TSGNet [42] and Sana-1600M [35], as well as two underwater-specific methods, i.e., Atlantis [11] and TIDE [12]. The evaluation was performed using Fréchet Inception Distance (FID) [43], Natural Image Quality Evaluator (NIQE) [44], the Visual Question Answering Score (VQA Score) [23], and the Human Preference Score v2 (HPSv2.1) [45]. The corresponding experimental results are reported in Table 1 and Fig. 8.

5.2.1. Quantitative Evaluation

As can be seen, UMDM-USG achieved the best performance in terms of FID and ranked the second on NIQE, VQA Score, and HPSv2.1, outperforming the two underwater-specific generation methods, i.e., Atlantis [11] and TIDE [12] by a large margin. These results demonstrate the strong overall advantage of UMDM-USG in balancing visual quality and semantic consistency for underwater image generation.

In contrast, the large-scale models [38, 39, 40] primarily designed for general-domain image generation exhibited competitive performance on certain semantic-related metrics, but fell behind UMDM-USG in terms of FID and structural plausibility. This indicates their limited ability to simultaneously preserve visual fidelity and structural consistency in complex underwater environments.

UMDM-USG generally better maintained geometric plausibility and semantic coherence during the generation of complex underwater scenes. TIDE [12] achieved the lowest NIQE score, suggesting an advantage in modeling low-level natural image statistics. However, its relatively inferior performance on VQA Score and HPSv2.1 reveals limitations in high-level semantic alignment and human preference consistency.

5.2.2. Qualitative Analysis

To intuitively validate the performance gains observed in our quantitative evaluation, we conducted a qualitative analysis across a range of representative underwater scenarios, as illustrated in Fig. 8. The selected prompts covered diverse environments, including expansive coral reef structures, large aquatic fauna, shipwrecks, and dense schools of fish.

The visual results indicate that general text-to-image models such as SD3-Medium [38], SD-XL [39], and Infinity [41] generally exhibited loose structural coherence, unstable geometric relationships, or semantic deviations in underwater scenes. In complex scenes, these models often produced blurred object contours and confused the layering between foreground and background. Sana [35] showed improvements in overall color style and global layout, but still fell short in capturing fine-grained geometric structures and small objects.

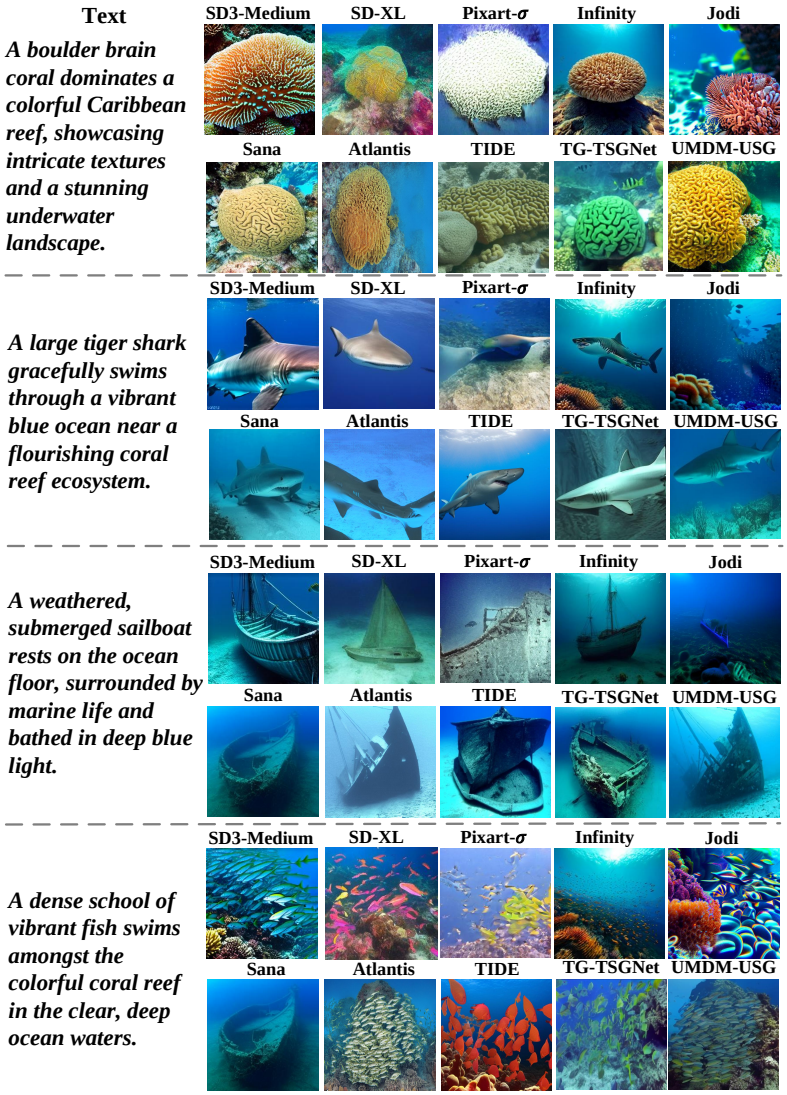

Figure 8: Qualitative comparison of different generative models in terms of a given textual description of underwater scene.

In contrast, UMDM-USG maintained overall visual realism while further enhancing structural clarity and multi-view consistency, enabling more accurate modeling of geometric and semantic information in complex underwater scenes. Besides, UMDM-USG showed more stable performance on small objects and densely populated regions, which can be attributed to the design of CGA-Attn and the incorporation of REPA [36].

5.3. Ablation Study

To investigate the effect of each component of UMDM-USG, we constructed a series of variants by removing or replacing individual components and evaluated them in the following ablation experiments.

5.3.1. Impact of CGA-Attn

To verify the role of CGA-Attn in multi-view generation, we further removed this module and trained the model using only GA-Attn and Cross Attention. As shown in Table 2, removing the CGA-Attn led to performance degradation across all evaluation metrics, indicating that this module can effectively enhance multi-view information alignment and improve the overall generation quality.

Table 2: Impact of the CGA-Attn mechanism on the generation performance of UMDM-USG.

Method	FID ↓	NIQE ↓	VQA Score ↑	HPSv2.1 ↑
UMDM-USG (w/o CGA-Attn)	10.21	4.42	0.51	22.19
UMDM-USG	**6.15**	**4.13**	**0.60**	**24.79**

5.3.2. Impact of REPA

We analyzed the effect of REPA [36] in UMDM-USG training. Specifically, we constructed different UMDM-USG variants by either not applying REPA or by applying alignment constraints to the hidden states of different layers in the multi-view alignment module. Table 3 shows that introducing REPA consistently improves the generation quality of UMDM-USG. Among the different layer settings, applying the alignment to the hidden states of 8 layers achieved the best overall performance, indicating that properly aligning intermediate representations can effectively enhance the model's generation capability.

5.3.3. Impact of Pre-trained Weight Initialization

To evaluate the effect of pre-trained weights on generation quality, we compared the models trained from random initialization against those initialized with the pre-trained

Table 3: Impact of REPA [36] on the generation quality of UMDM-USG.

Method	Layer	FID ↓	NIQE ↓	VQA Score ↑	HPSv2.1 ↑
UMDM-USG (w/o REPA [36])	-	8.45	4.21	0.49	21.80
UMDM-USG	4	7.02	4.19	0.60	24.21
UMDM-USG	8	**6.15**	**4.13**	0.60	**24.79**
UMDM-USG	16	6.98	4.40	**0.61**	23.77
UMDM-USG	24	7.89	4.33	0.59	23.04

Sana-600M [35] weights. Table 4 reports the comparative results. Models initialized with pre-trained weights consistently outperformed those trained from scratch across all evaluation metrics, indicating that pre-trained models provided better initial semantic and visual representations, which in turn facilitated improved generation quality.

Table 4: Comparison of the generation performance of UMDM-USG with random initialization and initialization using pre-trained Sana-600M [35] weights.

Method	Pre-trained Weights	FID ↓	NIQE ↓	VQA Score ↑	HPSv2.1 ↑
UMDM-USG	N/A	23.1	4.76	0.35	15.54
UMDM-USG	Sana-600M [35]	**6.15**	**4.13**	**0.60**	**24.79**

6. Downstream Tasks

To further validate the effectiveness of UMDM-USG in consistency modeling and data augmentation, we incorporated the generated multi-view data as auxiliary training samples for three representative underwater vision tasks, including underwater semantic segmentation, depth estimation, and surface normal estimation. Specifically, we augmented the training process with synthetic multi-view samples generated by UMDM-USG and baseline models [11, 12] when the original real annotated dataset was also used. We then evaluated the performance impact by comparing models trained on the augmented datasets against those trained exclusively on real-world data.

6.1. Experimental Setup

To ensure fair evaluation and prevent potential data leakage, all downstream task experiments were conducted strictly obeying the training split of each dataset and generated auxiliary data based on the training split.

Underwater Depth Estimation: Following the protocol in Atlantis [11], we evaluated the utility of the data that UMDM-USG generated using the D3 and D5 subsets of Sea-thru [6] and the SQUID dataset [46]. The Absolute Relative Error (A.Rel) was adopted as the primary evaluation metric to assess depth accuracy.

Underwater Semantic Segmentation: We evaluated segmentation performance on the UIIS [1] and USIS [2] datasets. The mean Intersection over Union (mIoU) was employed as the evaluation metric to quantify segmentation precision across diverse underwater object categories.

Underwater Surface Normal Estimation: Quantitative evaluation in this domain is inherently challenging due to the scarcity of public underwater benchmarks providing high-quality, ground-truth normal annotations. Existing datasets [6, 46] typically derived normals via numerical differentiation from depth maps, which often introduced noise and instability.

To ensure a rigorous evaluation, we followed the previous work [47] and adopted iBims-1 [48]. Although it is not an underwater-specific dataset, iBims-1 [48] provides high-precision ground-truth normals obtained via laser scanning and manual refinement, making it a standard for normal estimation [32, 33, 34]. The evaluation on iBims-1 [48] still offers a reasonable and reliable indication of the capability of the model in capturing geometric structures and maintaining normal consistency. During the evaluation process, both predicted and ground-truth normals were normalized to unit vectors. Performance was measured by the Mean Angular Error (MAE), representing the average angular deviation between the predicted and ground-truth vectors.

6.2. Underwater Depth Estimation

We selected four representative depth estimation methods for comparative evaluation, including AdaBins [49], NewCRFs [50], PixelFormer [51], and MIM [52]. As shown in Table 5, incorporating the depth data generated by UMDM-USG as auxiliary

training samples consistently led to improved or competitive performance across different depth estimation networks on the Sea-thru [6] and SQUID [46]. These results demonstrate the effectiveness of our model in synthesizing geometrically consistent depth information.

Table 5: Comparison of underwater depth estimation performance on the D3 and D5 subsets of the Sea-thru [6] dataset and the SQUID [46] dataset.

Method	Fine-Tuning Dataset	*A.Rel* ↓	
		Sea-thru [6]	SQUID [46]
AdaBins [49]	Atlantis [11]	1.33	0.28
	SynTIDE [12]	1.31	0.23
	UMDM-USG	**1.26**	**0.21**
NewCRFs [50]	Atlantis [11]	1.68	0.23
	SynTIDE [12]	1.50	0.23
	UMDM-USG	**1.44**	**0.20**
PixelFormer [51]	Atlantis [11]	1.34	0.18
	SynTIDE [12]	1.46	**0.16**
	UMDM-USG	**1.27**	0.19
MIM [52]	Atlantis [11]	1.37	0.26
	SynTIDE [12]	1.27	0.25
	UMDM-USG	**1.10**	**0.25**

6.3. Underwater Semantic Segmentation

We selected three mainstream semantic segmentation models for comparative evaluation, including SegFormer [53], Mask2Former [54], and ViT-Adapter [55]. To assess the impact of data augmentation, we compared three training strategies: (1) training exclusively on real annotated images, (2) joint training with synthetic data from SynTIDE [12], and (3) joint training with multi-view samples generated by the proposed UMDM-USG.

As shown in Table 6, incorporating segmentation maps generated by UMDM-USG as auxiliary training data consistently outperformed training with real images alone

Table 6: Comparison of underwater semantic segmentation performance on the UIIS [1] and USIS [2] datasets.

Method	Training Data			mIoU ↑	
	Real Images	SynTIDE [12]	UMDM-USG	UIIS [1]	USIS [2]
SegFormer [53]	✓			70.2	74.6
	✓	✓		**75.4**	76.1
	✓		✓	74.3	**77.6**
Mask2Former [54]	✓			72.7	76.4
	✓	✓		**74.2**	72.9
	✓		✓	73.7	**77.0**
ViT-Adapter [55]	✓			73.5	74.6
	✓	✓		**75.1**	76.7
	✓		✓	**75.1**	**77.9**

across all three segmentation models and both underwater datasets. This result demonstrates the effectiveness of generation-based data augmentation in alleviating annotation scarcity in underwater scenes and improving the generalization ability of segmentation models.

It should be noted that the performance advantage of SynTIDE [12] was partly attributable to its larger-scale synthetic training dataset, which contains approximately 50k samples. Nevertheless, the segmentation maps generated by UMDM-USG demonstrate comparable and competitive performance.

6.4. Underwater Normal Estimation

We evaluated three representative normal estimation methods, including StableNormal [32], DSINE [33], and Lotus [34], on the iBims-1 [48] dataset. The experimental results are reported in Table 7. Upon integrating the surface normal maps generated by UMDM-USG as auxiliary training data, all evaluated models exhibited a consistent reduction in MAE.

6.5. Summary

Comprehensive experimental results across segmentation, depth estimation, and surface normal estimation tasks confirmed that UMDM-USG not only achieved supe-

Table 7: Comparison of the MAE values obtained using different normal estimation methods [32, 33, 34] on the iBims-1 [48] dataset.

Method	Training Data		Mean Angular Error
	iBims-1 [48]	UMDM-USG	MAE ↓
StableNormal [32]	✓		17.2
	✓	✓	**17.0**
DSINE [33]	✓		18.7
	✓	✓	**17.4**
Lotus [34]	✓		17.1
	✓	✓	**16.3**

rior performance in underwater image synthesis but also served as a robust data augmentation engine. Its ability to generate physically plausible and alignment-consistent multi-view data improved the performance of diverse downstream underwater vision tasks. Although the construction of a certain portion of the U-TMDN is based on model-generated annotations, the consistent performance gains achieved across multiple downstream tasks demonstrate the effectiveness and reliability of these annotations.

7. Conclusion

In this paper, we presented UMDM-USG, a unified multi-view diffusion model for underwater scene generation that jointly models appearance, semantic, and geometric views under expressive textual guidance. By treating each modality as a distinct view of the same underwater scene, UMDM-USG adopts a unified diffusion architecture with explicit cross-view representation alignment, enabling coherent and scalable multi-view generation for complex underwater scenes. We also introduced U-TMDN, a large-scale, high-quality underwater multi-view dataset. Extensive experiments showed that UMDM-USG achieved superior generation performance and improved multiple downstream underwater vision tasks as a data augmentation source.

We acknowledge limitations of the current work. Approximately 47% of the segmentation masks in U-TMDN were generated through multi-model fusion rather than manual annotation, and the depth and normal maps were entirely model-predicted; er-

rors in these pseudo-labels might propagate into the generated outputs. In addition, the evaluation of surface normal estimation was conducted on iBims-1, an indoor dataset, due to the lack of public underwater benchmarks with high-quality ground-truth normals. Although this choice followed established practice, the domain gap between indoor and underwater scenes limited the interpretability of these results.

From a multi-view learning perspective, this work shows that principled cross-view alignment within a generative diffusion framework can yield strong generation quality and practical downstream benefits. We believe that UMDM-USG and U-TMDN provide a solid foundation for future research at the intersection of multi-view representation learning and conditional generation for underwater scene understanding.

CRediT authorship contribution statement

Yifan Zhu: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft; **Chengjia Wang:** Formal analysis, Methodology, Writing - Review & Editing; **Xinghui Dong:** Conceptualization, Funding acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing - Review & Editing.

Acknowledgement

This study was supported in part by the National Natural Science Foundation of China (NSFC) (No. 42576200) and in part by the Key Research and Development Program of Shandong Province, China (No. 2024ZLGX06).

References

[1] S. Lian, H. Li, R. Cong, S. Li, W. Zhang, S. Kwong, Watermask: Instance segmentation for underwater imagery, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1305–1315.

[2] S. Lian, Z. Zhang, H. Li, W. Li, L. T. Yang, S. Kwong, R. Cong, Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset (2024). arXiv:2406.06039.

[3] Z. Zheng, Y. Chen, H. Zeng, T.-A. Vu, B.-S. Hua, S.-K. Yeung, Marineinst: A foundation model for marine image analysis with instance visual description, in: European Conference on Computer Vision, Springer, 2024, pp. 239–257.

[4] C. Fu, R. Liu, X. Fan, P. Chen, H. Fu, W. Yuan, M. Zhu, Z. Luo, Rethinking general underwater object detection: Datasets, challenges, and solutions, Neurocomputing 517 (2023) 243–256.

[5] Y. Randall, T. Treibitz, Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets (2023). arXiv:2302.12772.

[6] D. Akkaynak, T. Treibitz, Sea-thru: A method for removing water from underwater images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[7] D. Levy, A. Peleg, N. Pearl, D. Rosenbaum, D. Akkaynak, S. Korman, T. Treibitz, Seathru-nerf: Neural radiance fields in scattering media, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 56–65.

[8] R. Liu, X. Fan, M. Zhu, M. Hou, Z. Luo, Real-world underwater enhancement: Challenges, benchmarks, and solutions (2019). arXiv:1901.05320.

[9] J. Wen, J. Long, X. Lu, C. Liu, X. Fang, Y. Xu, Partial multiview incomplete multilabel learning via uncertainty-driven reliable dynamic fusion, IEEE Transactions on Pattern Analysis and Machine Intelligence (2025).

[10] Z. Tang, Y. Fu, M. Li, H. Liang, Y. Tang, J. Wen, Mpr-net: Medicinal plant recognition network with dual-branch attention fusion, Pattern Recognition (2025) 112185.

[11] F. Zhang, S. You, Y. Li, Y. Fu, Atlantis: Enabling underwater depth estimation with stable diffusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11852–11861.

[12] H. Lin, D. Liang, Z. Qi, X. Bai, A unified image-dense annotation generation model for underwater scenes, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 961–970.

[13] M. J. Islam, Y. Xia, J. Sattar, Fast underwater image enhancement for improved visual perception, IEEE Robotics and Automation Letters (RA-L) 5 (2) (2020) 3227–3234.

[14] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, D. Tao, An underwater image enhancement benchmark dataset and beyond, IEEE Transactions on Image Processing (2020) 4376–4389.

[15] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, J. Sattar, Semantic segmentation of underwater imagery: Dataset and benchmark, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020, pp. 1769–1776.

[16] Y. Lyu, Y. Du, S. Tang, L. Hu, Stabilizing underwater acoustic data generation with generative adversarial network, Intelligent Marine Technology and Systems 3 (1) (2025) 12.

[17] J. Wen, Y. Liu, C. Huang, C. Liu, Y. Xu, X. Cao, Causal interventional prompt tuning for few-shot out-of-distribution generalization, IEEE Transactions on Pattern Analysis and Machine Intelligence 48 (2) (2026) 1978–1991.

[18] Y. Xie, M. Qiu, Y. Wang, S. Chen, M. Fang, M. Tang, W. Zhang, Mkgpl: graph prompt learning with multi-view knowledge for few-shot recognition, Pattern Recognition 172 (2026) 112737.

[19] A. Abdullah, T. Barua, R. Tibbetts, Z. Chen, M. J. Islam, I. Rekleitis, Caveseg: Deep semantic segmentation and scene parsing for autonomous underwater cave exploration, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 3781–3788.

[20] J. Han, M. Shoeiby, T. Malthus, E. Botha, J. Anstee, S. Anwar, R. Wei, M. A. Armin, H. Li, L. Petersson, Underwater image restoration via contrastive learning and a real-world dataset, Remote Sensing (2022).

[21] B. Lin, J. Dong, X. Dong, Perception-aware underwater image quality assessment: Dataset, perceptual quality scores, and assessment network, IEEE Transactions on Circuits and Systems for Video Technology (2025) 11113–11128.

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.

[23] Z. Lin, D. Pathak, B. Li, J. Li, X. Xia, G. Neubig, P. Zhang, D. Ramanan, Evaluating text-to-visual generation with image-to-text generation, in: European Conference on Computer Vision, Springer, 2024, pp. 366–384.

[24] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International Conference on Machine Learning, PMLR, 2023, pp. 19730–19742.

[25] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, D. Lin, Sharegpt4v: Improving large multi-modal models with better captions, in: European Conference on Computer Vision, Springer, 2024, pp. 370–387.

[26] Z. Zheng, J. Zhang, T.-A. Vu, S. Diao, Y. H. W. Tim, S.-K. Yeung, Marinegpt: Unlocking secrets of ocean to the public (2023). arXiv:2310.13596.

[27] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al., Gemma 3 technical report, arXiv preprint arXiv:2503.19786 (2025).

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the International Conference on Machine Learning, Vol. 139 of Proceedings of Machine Learning Research, PMLR, 2021, pp. 8748–8763.

[29] Y. Shang, W. Wang, C. Huang, X. Dong, Spda-sam: A self-prompted depth-aware segment anything model for instance segmentation (2026). arXiv:2602.06335.

[30] Z. Wang, S. Chen, L. Yang, J. Wang, Z. Zhang, H. Zhao, Z. Zhao, Depth anything with any prior (2025). arXiv:2505.10565.

[31] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, D. Novotny, Vggt: Visual geometry grounded transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 5294–5306.

[32] C. Ye, L. Qiu, X. Gu, Q. Zuo, Y. Wu, Z. Dong, L. Bo, Y. Xiu, X. Han, Stablenormal: Reducing diffusion variance for stable and sharp normal, ACM Transactions on Graphics 43 (6) (2024).

[33] G. Bae, A. J. Davison, Rethinking inductive biases for surface normal estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9535–9545.

[34] J. He, H. Li, W. Yin, Y. Liang, L. Li, K. Zhou, H. Zhang, B. Liu, Y.-C. Chen, Lotus: Diffusion-based visual foundation model for high-quality dense prediction (2025). arXiv:2409.18124.

[35] E. Xie, J. Chen, J. Chen, H. Cai, H. Tang, Y. Lin, Z. Zhang, M. Li, L. Zhu, Y. Lu, et al., Sana: Efficient high-resolution text-to-image synthesis with linear diffusion transformers, in: The Thirteenth International Conference on Learning Representations, 2025.

[36] S. Yu, S. Kwak, H. Jang, J. Jeong, J. Huang, J. Shin, S. Xie, Representation alignment for generation: Training diffusion transformers is easier than you think, arXiv preprint arXiv:2410.06940 (2024).

[37] Y. Xu, Z. He, M. Kan, S. Shan, X. Chen, Jodi: Unification of visual generation and understanding via joint modeling (2025). arXiv:2505.19084.

[38] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey,

A. Goodwin, Y. Marek, R. Rombach, Scaling rectified flow transformers for high-resolution image synthesis (2024). arXiv:2403.03206.

[39] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023). arXiv:2307.01952.

[40] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, Z. Li, Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation, in: European Conference on Computer Vision, Springer, 2025, pp. 74–91.

[41] J. Han, J. Liu, Y. Jiang, B. Yan, Y. Zhang, Z. Yuan, B. Peng, X. Liu, Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 15733–15744.

[42] Y. Zhu, Y. Wang, X. Dong, Tg-tsgnet: A text-guided arbitrary-resolution terrain scene generation network, IEEE Transactions on Image Processing 34 (2025) 8614–8626.

[43] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in Neural Information Processing Systems 30 (2017).

[44] A. Mittal, A. K. Moorthy, A. C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Transactions on Image Processing 21 (12) (2012) 4695–4708.

[45] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, H. Li, Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis (2023). arXiv:2306.09341.

[46] D. Berman, D. Levy, S. Avidan, T. Treibitz, Underwater single image color restoration using haze-lines and a new quantitative dataset (2019). arXiv:1811.01343.

[47] A. Saleh, M. Olsen, B. Senadji, M. R. Azghadi, A practical approach to underwater depth and surface normals estimation (2025). arXiv:2410.02072.

[48] T. Koch, L. Liebel, F. Fraundorfer, M. Körner, Evaluation of cnn-based single-image depth estimation methods (2018). arXiv:1805.01328.

[49] S. F. Bhat, I. Alhashim, P. Wonka, Adabins: Depth estimation using adaptive bins, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4009–4018.

[50] W. Yuan, X. Gu, Z. Dai, S. Zhu, P. Tan, Neural window fully-connected crfs for monocular depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3916–3925.

[51] A. Agarwal, C. Arora, Attention attention everywhere: Monocular depth prediction with skip attention, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5861–5870.

[52] Z. Xie, Z. Geng, J. Hu, Z. Zhang, H. Hu, Y. Cao, Revealing the dark secrets of masked image modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14475–14485.

[53] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, Segformer: Simple and efficient design for semantic segmentation with transformers, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, Inc., 2021, pp. 12077–12090.

[54] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1290–1299.

[55] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, Y. Qiao, Vision transformer adapter for dense predictions (2023). arXiv:2205.08534.