



International Journal of Information Quality

ISSN online: 1751-0465 - ISSN print: 1751-0457

<https://www.inderscience.com/ijiq>

Semantic-guided multi-scale diffusion network for immersive tourism content generation

Lun Wei, Afia Zafar, Muhammad Attique Khan

DOI: [10.1504/IJIQ.2026.10081506](https://doi.org/10.1504/IJIQ.2026.10081506)

Article History:

Received:	15 June 2026
Last revised:	24 July 2026
Accepted:	28 July 2026
Published online:	23 October 2026

Semantic-guided multi-scale diffusion network for immersive tourism content generation

Lun Wei*

College of Art and Design,
Ma'anshan University,
Ma'anshan, 243100, Anhui, China
Email: masu-wl@masu.edu.cn
*Corresponding author

Afia Zafar

National University of Computer and Emerging Sciences (Fast),
Islamabad, Pakistan
Email: afia.zafar@isb.nu.edu.pk

Muhammad Attique Khan

Department of Computer Science,
College of Computer Engineering and Science,
Prince Mohammad Bin Fahd University,
Eastern, Saudi Arabia
Email: attique.khan@ieee.org

Abstract: This paper proposes an immersive tourism content generation model based on a semantic-guided multi-scale diffusion network (SGMSD-Net) to address semantic consistency, panoramic structure preservation, and high-resolution detail restoration in digital cultural tourism engineering. Built on the diffusion generation framework, SGMSD-Net integrates semantic modulation, multi-scale diffusion, and multi-objective optimization. Semantic embeddings and feature-wise linear modulation (FiLM) are used to dynamically guide the generation process, while a hierarchical diffusion module progressively restores structural and textural details from coarse to fine scales. A joint loss combining semantic consistency and perceptual constraints is further introduced to improve realism and semantic alignment. Experiments on SUN360 and Matterport3D show that SGMSD-Net outperforms existing methods in FID, LPIPS, and CLIP score. Subjective evaluations also confirm its advantages in clarity, spatial consistency, and visual appeal, demonstrating its potential for intelligent cultural tourism, virtual tours, and digital heritage reconstruction.

Keywords: semantic-guided diffusion model; multi-scale image generation; immersive tourism; panoramic scene synthesis; digital cultural heritage reconstruction.

Reference to this paper should be made as follows: Wei, L., Zafar, A. and Khan, M.A. (2026) 'Semantic-guided multi-scale diffusion network for immersive tourism content generation', *Int. J. Information Quality*, Vol. 4, No. 5, pp.1–20.

Biographical notes: Lun Wei is an Associate Professor at Ma'anshan University, holding a Master's degree and based in Huoshan County, Anhui Province. Her research and teaching focus on visual communication design, curriculum ideology and politics, and cultural brand design. She has presided over multiple provincial-level projects, including key humanities and social sciences projects like 'Urban Visual Identity Elements Design in the Yangtze River Delta' (completed) and 'AIGC-Driven Wanjia Culture Brand Reconstruction' (in progress), as well as quality engineering projects on curriculum innovation. She has also participated in provincial key projects on cultural element transformation.

Afia Zafar is affiliated with the National University of Computer and Emerging Sciences (FAST) Islamabad, Pakistan. As a member of this prestigious Pakistani university renowned for computer science and technology education, she engages in academic and related work at the Islamabad campus, one of the core branches of FAST-NUCES.

Muhammad Attique Khan serves at the Department of Computer Science, College of Computer Engineering and Science, Prince Mohammad Bin Fahd University in the Eastern Province, Saudi Arabia. He conducts academic work and research in the computer science field at this well-known local university.

1 Introduction

With the continuous advancement of information technology and artificial intelligence, the tourism industry is gradually shifting from traditional offline models to intelligent service paradigms centred on digital experiences. In recent years, the development of tourism recommendation systems and content generation technologies has become a key focus of research in intelligent cultural tourism. Early tourism recommendation approaches primarily relied on algorithms based on geographical location, user ratings, or collaborative filtering, utilising users' historical behaviour data to recommend destinations or plan itineraries (Nitu et al., 2021). However, these methods mainly emphasised decision-making recommendations, neglecting the visual and perceptual dimensions of the tourism experience. As a result, they fall short of satisfying modern users' growing demand for immersive and interactive content in the information age. At the same time, the rise of tourism content generation technologies, particularly those grounded in image generation and virtual reality, has enabled users to explore and understand destination environments through high-fidelity visual content before their actual trips – ushering in the era of pre-experiential tourism. Against this backdrop, research on tourism content generation has increasingly drawn attention from both academia and industry. By generating high-quality tourism imagery or immersive panoramic content, such technologies not only enhance user engagement but also offer new technical support for destination marketing, cultural heritage preservation, and urban planning. For example, virtual reconstructions of historical landscapes and immersive visualisations of scenic areas have become significant innovations in the integration of culture and tourism. Nevertheless, existing mainstream content generation methods still exhibit notable limitations: their outputs are often guided by general visual priors and lack fine-grained modelling of the semantic characteristics specific to tourism scenes (Jang et al., 2023). Therefore, incorporating semantic understanding mechanisms into

generation models – enabling them to truly comprehend the semantic structures and cultural connotations of “tourism scenes” – has emerged as a critical research challenge. Ultimately, the significance of immersive tourism content generation lies not only in achieving visual realism but also in providing users with perceptible and interactive digital experiences, advancing tourism research from recommendation toward experience creation.

In recent years, deep learning and machine learning have achieved remarkable breakthroughs in image generation, semantic understanding, and cross-modal representation, providing strong technical foundations for tourism content generation. Image synthesis methods based on convolutional neural networks (CNNs) and their variants – are capable of generating realistic tourism landscape images from semantic labels or conditional maps, thereby effectively enhancing image quality and scene consistency. Meanwhile, the emergence of variational autoencoders (VAEs) and Diffusion Models has made the generation process more stable and controllable, enabling the progressive restoration of image details across multiple spatial scales. These approaches allow for more realistic modelling of lighting, texture, and structural patterns, extending their applications beyond tourism scene reconstruction to include virtual tours, 3D visual augmentation, and other immersive experiences (Huang, 2022). Furthermore, the development of Vision-Language Models has brought semantic-guided content generation to the forefront of research. By incorporating text embeddings or semantic labels, models can interpret user-intent descriptions such as “a seaside city at dawn” or “an ancient town street under the night sky” and generate corresponding visual content accordingly. This cross-modal semantic control mechanism elevates tourism content generation from simple image synthesis to semantic-driven visual expression. At the same time, deep learning has also played an important role in personalised content generation. By integrating user profiling and preference modelling, systems can produce customised tourism visuals aligned with individual interests (Chen et al., 2026). For instance, for users who favour natural scenery, the system can automatically generate landscapes featuring mountains, rivers, and lakes; while for cultural exploration-oriented users, it can synthesise imagery of ancient towns or historical architectures. Overall, the incorporation of deep learning methodologies has significantly enhanced the intelligence, diversity, and semantic expressiveness of tourism content generation. Through end-to-end learning that bridges low-level visual features and high-level semantics, these models demonstrate immense potential in advancing virtual tourism and digital cultural tourism experiences.

Although deep learning has driven significant progress in tourism content generation, several limitations remain in existing research. First, most current generation models lack fine-grained control over semantic hierarchies. They tend to produce visually similar images without accurately capturing the intended semantic meaning of tourism scenes (Mou et al., 2022). Second, mainstream single-scale generation methods often suffer from texture blurring, uneven lighting, and structural distortion when generating high-resolution images, making them inadequate for meeting the quality requirements of immersive environments. Third, in tourism content generation tasks, existing approaches still exhibit limited capability in integrating multi-modal conditions – such as semantics, geographic cues, and depth information – resulting in outputs that lack spatial coherence and realism. To address these challenges, this paper proposes an immersive tourism content generation model based on a semantic-guided multi-scale diffusion network (SGMSD-Net). Through semantic modulation, hierarchical diffusion, and multi-objective

optimisation, the model achieves end-to-end modelling from semantic understanding to high-fidelity content generation.

The main innovations and contributions of this work are summarised as follows:

1 *Semantic-guided modulation mechanism*

A semantic-guided modulation framework is developed by integrating semantic embeddings with FiLM-based feature adaptation into the diffusion generation process. Different from conventional diffusion models that mainly rely on implicit semantic guidance, the proposed framework explicitly incorporates tourism-related semantic attributes into intermediate feature representations, enabling more controllable panoramic scene generation.

2 *Multi-scale diffusion generation framework*

The generation process is divided into three resolution stages, progressively synthesising high-resolution tourism images from coarse to fine. This hierarchical design allows structural, textural, and illumination features to be optimised at different levels, significantly enhancing image realism and spatial consistency.

3 *Comprehensive evaluation system and experimental validation*

Comparative experiments are conducted on the SUN360 and Matterport3D datasets. Using metrics such as FID, LPIPS, and CLIP Score, the proposed method is evaluated from three dimensions – visual quality, perceptual consistency, and semantic alignment – to comprehensively validate its effectiveness.

2 Related works

2.1 *Research on diffusion-based generative models*

Yang et al. (2023) proposed a systematic review framework that organises the methodological taxonomy and developmental trajectory of diffusion models across multiple modalities, including images, videos, and language. They identified three major research directions: efficient sampling, likelihood estimation improvement, and structured data modelling. Through comparative analysis, they examined differences among classifier-guided, classifier-free guidance, DDPM, Score-based, and Latent Diffusion methods, highlighting key challenges in text-to-image and discrete language generation tasks. Ke et al. (2024) introduced the meta diffusion model, a meta-learning-driven framework designed for text generation under strong constraints. By explicitly injecting hard constraints into the diffusion inversion process and employing meta-learning for rapid adaptation to different constraint sets, MDM achieves a balanced trade-off between feasibility and fluency. Gao et al. (2025) proposed TWDT, a “training-free” word-level controllable diffusion text generation method. Their approach directly applies semantic guidance to the latent space during inference, enabling fine-grained control over emotion, style, theme, and length, while preserving the non-autoregressive parallel sampling advantages of diffusion-based language models. Experimental results demonstrate consistent improvements in content alignment and readability across various conditional and stylistic benchmarks. The study also discusses the trade-off between control strength and fluency, providing a practical path for enhancing controllability in

real-world tasks such as instruction writing and style transfer without additional retraining. Li et al. (2023) proposed ET-DM, which integrates an Efficient Transformer with a diffusion model for text-to-image synthesis. The authors utilised efficient attention mechanisms and lightweight network structures to reduce inference steps and memory consumption, while embedding text-semantic alignment into the diffusion inversion process to improve detail fidelity and generation accuracy. Zhou et al. (2025) presented MaskDiffusion, addressing the issue of multi-entity or attribute mismatches in text-to-image generation. By introducing conditional masking and region-level guidance, their model explicitly constrains key objects and attributes throughout the diffusion inversion stages, thereby improving text-image consistency and mitigating attribute leakage. Daneshfar et al. (2024) conducted the first systematic review focusing specifically on diffusion-based image captioning. Using a task-oriented taxonomy, they summarised the roles of diffusion models in caption generation – covering conditional modelling, guidance and alignment strategies, decoding mechanisms, and evaluation protocols – and provided an analytical comparison with autoregressive and retrieval-based methods, outlining their respective strengths and weaknesses.

2.2 Research on tourism information generation and analysis

Hu et al. (2023) first extracted geographical location tags from travelogue texts, then retrieved corresponding web images based on these tags, and finally presented the related textual and visual information to users through an integrated interface. Wang et al. (2023) collected 340,000 check-in records from 86 scenic spots in Shenzhen via Sina Weibo, and mined data to analyse tourist origins, travel duration, check-in activity indices, and inter-attraction correlations. Shapoval et al. (2018), using data from the Japan National Tourism Organisation, applied decision tree analysis to investigate the behavioural patterns of inbound tourists in Japan. Similarly, Fu et al. (2025) employed frequent itemset mining on large-scale travel blogs to identify popular attractions and classic travel routes. Wang et al. (2022) conducted a comprehensive review of text-based tourism big data mining methods, detailing various text mining techniques such as topic detection, text classification, clustering, and sentiment analysis. Park et al. (2021) argued that tourism information mining can significantly improve the accuracy of tourist demand prediction, thereby providing valuable insights for policy makers, industry investors, and destination managers to support data-driven planning and investment decisions. Acharya and Mohbey (2023) highlighted that the widespread use of location-based social networks has generated vast amounts of user data. Based on this, they proposed a personalised recommendation algorithm that integrates users' historical behaviours, attraction locations, social relationships, and user similarity, and validated the model using real-world LBSN datasets. Collectively, these studies demonstrate that tourism content generation research has traditionally focused on information mining and multimodal analysis, integrating textual and visual data to form the foundation for content recommendation. However, with the continuous advancement of deep learning technologies, the cross-modal transformation between different types of data – such as generating visuals from text – has become increasingly feasible. Building upon this technological shift, this paper investigates the problem of immersive tourism content generation. By leveraging diffusion models, it aims to generate semantically aligned, high-fidelity visual scenes from textual descriptions, thereby providing users with a more immersive and engaging tourism experience.

3 Methodology

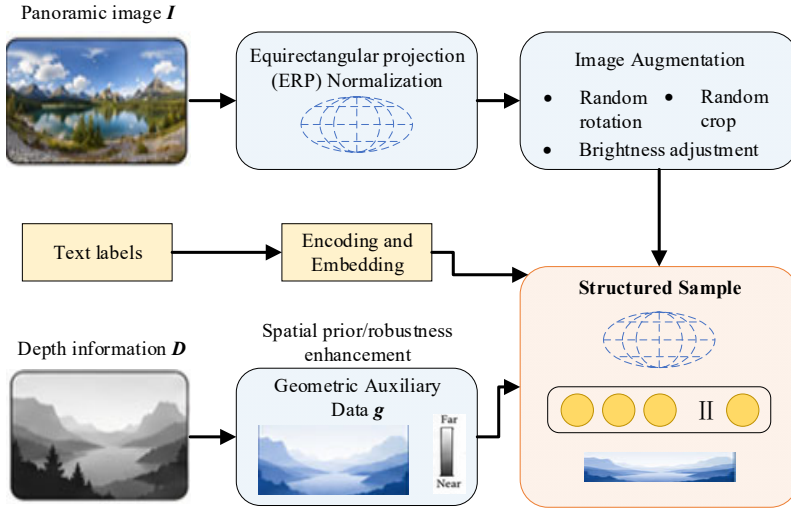
3.1 Data preprocessing

To train the generative model, this study employs tourism scene image datasets as the primary data source. The original panoramic images are standardised using an equirectangular projection to ensure geometric consistency across samples. Existing label information within the dataset is utilised to construct corresponding semantic vectors. Each sample in the dataset primarily contains three components: original panoramic image, associated textual label or scene description, and depth information of the scene. The textual descriptions or categorical labels are one-hot encoded and subsequently embedded into a semantic vector space (Zhang et al., 2024). After preprocessing, each sample can be formally represented as shown in equation (1):

$$S_i = \{I_i, T_i, D_i\} \quad (1)$$

where I_i denotes the original panoramic image, T_i represents the semantic embedding derived from the textual label or scene description, and D_i corresponds to the depth information associated with the scene. This structured preprocessing ensures that all multimodal inputs-visual, textual, and spatial-are properly aligned and normalised, providing a consistent foundation for the subsequent SGMSD-Net training process. The feature processing is detailed shown in Figure 1 as follows:

Figure 1 The data preprocessing before training (see online version for colours)



3.2 SGMSD-Net establishment

The ultimate goal of this study is to generate immersive images from users' textual descriptions. Therefore, during the design process, the primary inputs are the noisy image xxx and the corresponding semantic vector mentioned above, while the final output is the immersive image. The proposed SGMSD-Net mainly consists of three key modules: the

semantic encoding module, the multi-scale diffusion backbone module, and the temporal embedding and guidance module.

First, in the semantic encoding module, the text vector is processed through an MLP for mapping, as follows.

The semantic label is embedded into a semantic vector and mapped into channel modulation parameters to generate the feature scaling factor γ and the bias term β :

$$\gamma = \text{MLP}_\gamma(s_i) \quad (2)$$

$$\beta = \text{MLP}_\beta(s_i) \quad (3)$$

and applied to the intermediate layer features of the UNet through FiLM:

$$F' = \gamma \odot F + \beta \quad (4)$$

The semantic embedding controls the scaling and shifting of features, enabling different tourism semantics to produce distinct feature distributions, thereby guiding the generation direction. In the multi-scale diffusion process, a three-level resolution structure is adopted, progressively generating clearer images from low to high resolutions. Each level employs an independent UNet subnetwork as the noise predictor (Huang et al., 2024). The motivation of the multi-scale diffusion design is based on the observation that panoramic tourism images contain hierarchical visual information, where global scene layout, object-level structures, and local texture details exhibit different spatial characteristics. Directly learning all these representations within a single high-resolution diffusion process requires the model to simultaneously optimise low-frequency structural information and high-frequency details, which may increase the difficulty of feature learning. Therefore, SGMSD-Net decomposes the generation process into multiple resolution stages. The low-resolution diffusion stage first establishes global spatial organisation and semantic layout, while subsequent stages progressively refine structural boundaries and high-frequency visual details. This coarse-to-fine optimisation strategy enables more effective allocation of model capacity across different spatial frequencies and improves the consistency of panoramic scene generation.

$$\epsilon_\theta^{(k)} = f_{\text{UNet}}^{(k)}(x_t^{(k)}, s_i, t) \quad (5)$$

In the reverse diffusion sampling process:

$$x_{t-1}^{(k)} = \frac{1}{\sqrt{\alpha_t}} \left(x_t^{(k)} - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta^{(k)}(x_t^{(k)}, s_i, t) \right) + \sigma_t z_t \quad (6)$$

where $\epsilon_\theta^{(k)}$ is the noise of the network prediction, α_t represents the diffusion noise scheduling parameter; $\sigma_t z_t$ denotes the Gaussian sampling noise. The denoised output $\hat{x}_0^{(k)}$ is used as the upsampling condition for the next stage:

$$y^{(k+1)} = \text{U}_p(\hat{x}_0^{(k)}) \quad (7)$$

The final output is the highest-resolution generated image $\hat{x}_0^{(3)}$. The multi-scale generation strategy enables the model to learn global layouts and local textures progressively from coarse to fine, reducing the instability associated with directly generating high-resolution images. On the other hand, temporal embedding and guidance are incorporated into the model: the time step t is encoded using sinusoidal positional encoding $\psi(t)$ to help the model perceive the diffusion progress. During inference, classifier-free guidance is applied, as shown in equation (8):

$$\hat{\epsilon}_t = (1 + w)\epsilon_\theta(x_t, s_i, t) - w\epsilon_\theta(x_t, \phi, t) \quad (8)$$

where w represents the strength of semantic guidance. This strategy enhances semantic consistency without introducing an additional discriminator and allows the adjustment of generation diversity according to task requirements. Therefore, through the integration of semantic guidance and a multi-scale diffusion mechanism, the model can generate tourism scene images with complete structures, consistent semantics, and fine-grained textures, providing high-quality content for applications such as virtual tourism and digital tours. The motivation of the hierarchical diffusion strategy is based on the multi-scale characteristics of panoramic tourism images. Specifically, global scene layout and spatial organisation mainly correspond to low-frequency visual information, while object boundaries, texture patterns, and illumination variations are represented by high-frequency details. A single-stage high-resolution diffusion model needs to simultaneously optimise these heterogeneous representations, which increases the difficulty of feature learning. Therefore, SGMSD-Net decomposes the generation process into multiple resolution stages, allowing different diffusion stages to focus on specific levels of visual information and progressively refine the generated results.

3.3 Model framework and loss design

Based on the SGMSD-Net introduced in Section 3.2, the final output of the network is the realistic image. Considering the three-level diffusion process, three types of loss functions are designed in this study: noise prediction loss, perceptual consistency loss, and semantic consistency loss. The definitions of these three loss functions, along with the final total loss function, are as follows:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x, \epsilon, t} \epsilon - \epsilon_\theta(x_t, s_i, t)_2^2 \quad (9)$$

$$\mathcal{L}_{\text{perc}} = \phi(\hat{x}_0) - \phi(I_i)_2^2 \quad (10)$$

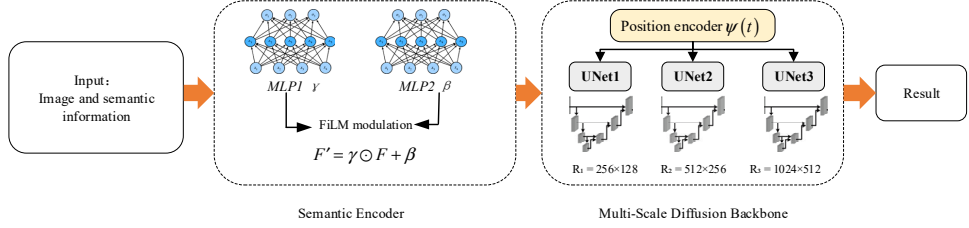
$$\mathcal{L}_{\text{sem}} = 1 - \cos(f(\hat{x}_0), s_i) \quad (11)$$

where $\mathcal{L}_{\text{diff}}$ is the noise prediction loss, is used to train the model for accurate noise prediction; $\mathcal{L}_{\text{perc}}$ represents the perceptual consistency loss, which ensures that the generated image maintains visual detail consistency with the real image, $\mathcal{L}_{\text{sem}}$ denotes the semantic consistency loss, which constrains the generated result to align with the semantic labels. The final total loss function is defined as shown in Equation (12):

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{diff}} + \lambda_2 \mathcal{L}_{\text{perc}} + \lambda_3 \mathcal{L}_{\text{sem}} \quad (12)$$

Based on the above information, the overall architecture of the proposed model is illustrated in Figure 2.

Figure 2 The framework for SGMSD-Net (see online version for colours)



As shown in Figure 2, the framework first receives an input random noise image along with the corresponding semantic label information. The semantic labels are processed through an embedding layer and a MLP to generate semantic vectors, which are then used to produce feature modulation parameters for semantic guidance within the feature layers of the UNet. Next, under the control of time-step encoding, the noisy image passes sequentially through three multi-scale diffusion modules of different resolutions. The model progressively denoises and refines the image from low to high resolution, achieving a coarse-to-fine generation process. The output from each stage is upsampled and used as a prior condition for the next stage, ensuring structural continuity and rich detail. Finally, the model outputs a high-resolution immersive tourism image, in which semantic information and visual features remain consistent. This enables realistic reconstruction of specific landscapes, scenes, or environments, providing a robust foundation for high-quality content generation in virtual tourism and digital cultural tourism applications.

4 Experiment results and analysis

4.1 Datasets and experiment setup

In this study, to meet the needs of immersive tourism content generation, we used two public datasets, SUN360 (Chen et al., 2023) and Matterport3D (Ramakrishnan et al., 2021), for model testing. SUN360 is a large-scale panoramic image dataset containing more than 26,000 high-resolution indoor and outdoor panoramic images, covering various scene types such as natural landscapes, urban streets, and building interiors. The data are stored in an equirectangular projection format, supporting full-view observation. This dataset provides scene categories and semantic labels, making it suitable for panoramic image generation, immersive visual understanding, scene reconstruction, and tourism content generation. Matterport3D is a high-quality real-world three-dimensional scene dataset that includes over 10,000 panoramic viewpoints from more than 90 architectural environments. Each sample contains RGB images, depth maps, normal maps, and camera pose information, which can be used for three-dimensional

reconstruction and spatial understanding. The dataset covers various indoor structures such as residential buildings, hotels, and shopping malls, featuring high geometric precision and rich semantic annotations, making it ideal for research on immersive content generation, visual navigation, and structural perception.

In model performance evaluation, we used three metrics: FID, LPIPS, and CLIP Score. FID measures the difference in feature distribution between generated and real images. It uses the Inception network to extract high-level image features and then calculates the Fréchet distance between the mean and covariance of the generated and real image feature distributions. A lower FID indicates that the generated images are more perceptually similar to real ones, making it one of the most commonly used quality metrics in image generation tasks. LPIPS calculates perceptual similarity in deep network feature space to assess the visual consistency between images. Compared with pixel-level metrics, LPIPS better aligns with human perception. A lower LPIPS value indicates that the generated image is closer to the real image in terms of semantics and textural details, and it is commonly used to evaluate the quality of diffusion and adversarial generation models. CLIP Score measures the similarity between generated images and input semantic descriptions based on the CLIP model. By comparing the cosine similarity of images and texts in a unified vision-language feature space, it evaluates the degree of semantic alignment. A higher CLIP Score indicates that the generated content better matches the semantic conditions, making it suitable for quantitative evaluation in semantic-guided generation tasks. In addition to CLIP Score, Semantic Attribute Accuracy (SAA) is introduced to further evaluate the fine-grained semantic controllability of the generated images. Different from CLIP Score, which measures the overall similarity between images and text descriptions, SAA evaluates whether specific semantic attributes contained in the input conditions are correctly reflected in the generated results. Specifically, a pretrained attribute classifier is employed to identify semantic attributes from generated images, and the accuracy between the predicted attributes and the input semantic labels is calculated. In this study, scene category, landmark type, and weather condition are considered as the main semantic attributes.

For comparison, four representative generative models, Pix2PixHD (Yu et al., 2024), SPADE (Gerogiannis et al., 2023), DDPM (Alblwi et al., 2025), and stable diffusion (Du et al., 2023), were selected as baselines to verify the advantages of the proposed SGMSD-Net in semantic consistency and immersive image quality. This study uses FID, LPIPS, and CLIP Score to comprehensively evaluate model performance from three aspects: visual quality, perceptual consistency, and semantic alignment. As this study involves large-scale parameter computation and image generation, the experimental environment we built is shown in Table 1 and the training configuration of SGMSD-Net is summarised in Table 2.

Table 1 The experiment environment setup

<i>Environment</i>	<i>Information</i>
CPU	I7-14700F
GPUs	RTX 4080
Language	Python 3.5
Framework	Pytorch

Table 2 Training configuration of SGMSD-Net

Configuration	Setting
Diffusion stages	3
Diffusion steps	1,000
Sampling steps	50
Optimiser	Adam
Learning rate	1×10^{-4}
Batch size	8
Training epochs	200
Semantic embedding dimension	256

These baselines cover different categories of image generation approaches, including conditional adversarial generation methods, semantic layout-guided generation methods, and diffusion-based generation frameworks. Although recent large-scale diffusion models have demonstrated strong general image synthesis capabilities, this study focuses on tourism-specific panoramic generation, where semantic consistency, spatial structure preservation, and scene-level controllability are the primary evaluation objectives.

4.2 Method comparison and result analysis

After completing the model construction and training, experiments were conducted on the two datasets, and the main results are shown in Figures 3 and 4.

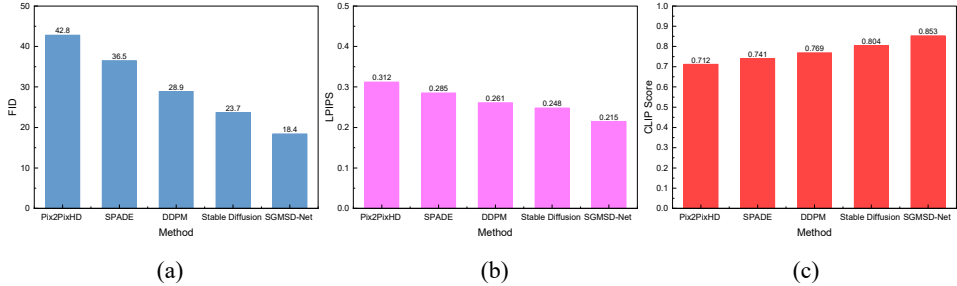
Figure 3 The method comparison results on SUN360 datasets (see online version for colours)

Figure 3 presents the performance comparison results of different methods on the SUN360 dataset. As shown in Figure 3(a), SGMSD-Net achieves the lowest FID value among all methods, indicating that the distribution of the generated images is closest to that of real panoramic tourism images. This result demonstrates that the proposed method can generate more realistic visual content with fewer distributional differences from the target data. In Figure 3(b), SGMSD-Net also obtains the lowest LPIPS score, suggesting that its generated images have better perceptual consistency and stronger detail restoration ability. Compared with Pix2PixHD and SPADE, which tend to produce blurred textures or insufficient structural details, SGMSD-Net can better preserve local visual features. As shown in Figure 3(c), the proposed method achieves the highest CLIP Score, proving that the generated results are more consistent with the input semantic

descriptions. This advantage mainly benefits from the semantic modulation mechanism, which strengthens the coupling between semantic embeddings and visual generation.

Figure 4 The method comparison results on Matterport3D datasets (see online version for colours)

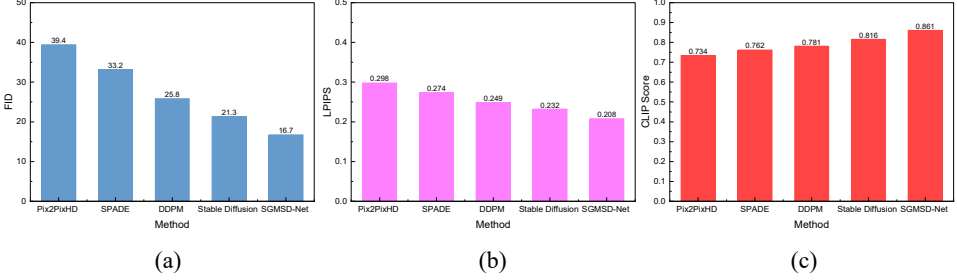


Figure 4 presents the comparison results on the Matterport3D dataset. According to Figure 4(a), SGMSD-Net again obtains the best FID performance, indicating its strong ability to generate realistic indoor immersive scenes with complex spatial structures. In Figure 4(b), the lower LPIPS value of SGMSD-Net shows that the proposed model can better restore perceptual details such as indoor textures, lighting, and spatial boundaries. This is particularly important for immersive tourism content generation, where structural continuity and scene realism directly affect user experience. As illustrated in Figure 4(c), SGMSD-Net achieves the highest CLIP Score, further confirming its superiority in semantic alignment. Compared with the baseline diffusion models, SGMSD-Net can more accurately generate visual content corresponding to specific tourism-related semantic conditions.

To further verify the semantic controllability of SGMSD-Net beyond global semantic similarity evaluation, the Semantic Attribute Accuracy (SAA) results are reported in Table 3. SGMSD-Net achieves an average SAA of 92.9%, outperforming Pix2PixHD, SPADE, DDPM, and Stable Diffusion by 12.6%, 11.0%, 8.3%, and 5.9%, respectively. This improvement indicates that the proposed FiLM-based semantic modulation mechanism can effectively transfer high-level semantic attributes into visual generation patterns. In particular, the model achieves higher accuracy in scene category and landmark recognition, demonstrating its capability to preserve key tourism-related semantic information during the generation process. Meanwhile, the weather attribute accuracy remains slightly lower because environmental conditions generally involve more subtle visual variations, such as illumination and atmospheric effects.

Table 3 Semantic attribute accuracy comparison

Method	Scene (%)	Landmark (%)	Weather (%)	Average SAA (%)
Pix2PixHD	82.6	79.8	78.5	80.3
SPADE	84.1	81.5	80.2	81.9
DDPM	86.7	84.3	82.8	84.6
Stable diffusion	88.9	86.5	85.7	87.0
SGMSD-Net	94.2	92.8	91.6	92.9

4.3 Application analysis

After completing the analysis of objective evaluation metrics, a subjective evaluation experiment was further conducted to assess the perceptual quality of generated images. Ten volunteers were recruited, including 6 males and 4 females, with ages ranging from 22 to 45 years. All participants had normal or corrected-to-normal vision and were familiar with digital image viewing tasks, but none of them participated in the model development process. To reduce potential selection bias, images for subjective evaluation were randomly selected from the test sets of SUN360 and Matterport3D. For each dataset, the same number of generated images from different methods was presented to participants in a randomised order. Participants were asked to independently score each image from three aspects: visual realism, semantic consistency, and overall perceptual quality, using a five-point scale. The final score of each method was calculated by averaging the ratings from all participants.

Figure 5 The picture quality distribution on SUN360 datasets (see online version for colours)

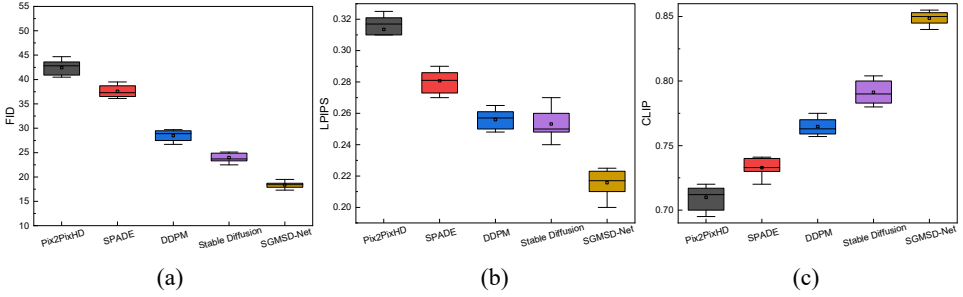


Figure 6 The picture quality distribution on Matterport3D datasets (see online version for colours)

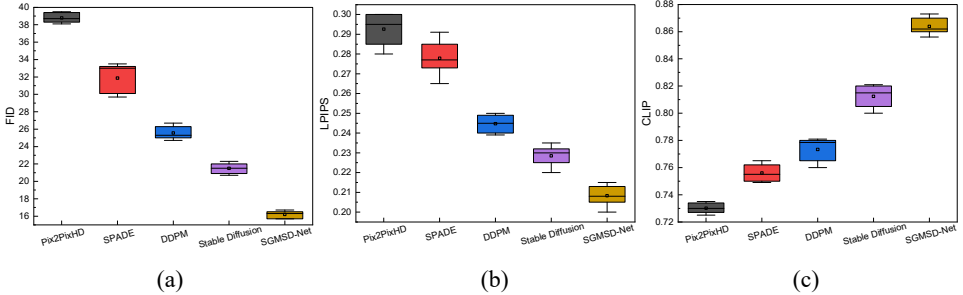


Figure 5 presents the quality distribution results on the SUN360 dataset. As shown in Figure 5(a), SGMSD-Net obtains the lowest FID distribution among all comparison methods, and its boxplot is more concentrated, indicating that the generated images are closer to real images in terms of feature distribution and that the model has better generation stability. In Figure 5(b), SGMSD-Net achieves the lowest LPIPS distribution, suggesting that the generated images are more consistent with real images in perceptual quality, texture details, and local visual structures. Compared with Pix2PixHD and SPADE, the proposed method produces fewer perceptual deviations and more stable

visual details. As shown in Figure 5(c), SGMSD-Net achieves the highest CLIP Score distribution, demonstrating that its generated results have stronger semantic consistency with the input descriptions. This result further verifies the effectiveness of the semantic-guided modulation mechanism in improving text-image alignment.

Figure 6 shows the quality distribution results on the Matterport3D dataset. In Figure 6(a), SGMSD-Net maintains the lowest FID distribution, indicating that the proposed method can generate more realistic immersive indoor scenes even under complex spatial structures. As shown in Figure 6(b), the LPIPS distribution of SGMSD-Net is also the lowest, which proves that the model performs better in restoring indoor textures, lighting details, and spatial boundaries. This advantage is particularly important for virtual tourism and immersive scene construction, where spatial continuity directly influences the user’s visual experience. In Figure 6(c), SGMSD-Net obtains the highest CLIP Score distribution, showing that the generated images are more accurately aligned with the corresponding semantic conditions. These results indicate that SGMSD-Net achieves consistent performance across the evaluated panoramic scene datasets, demonstrating its effectiveness in generating high-quality tourism images under different scene conditions.

After completing the analysis of objective evaluation metrics, ten volunteers were recruited to conduct a subjective assessment of the quality of the generated images. The evaluation results of different models for various images are shown in Figure 7.

Figure 7 Comparison of user scores (see online version for colours)

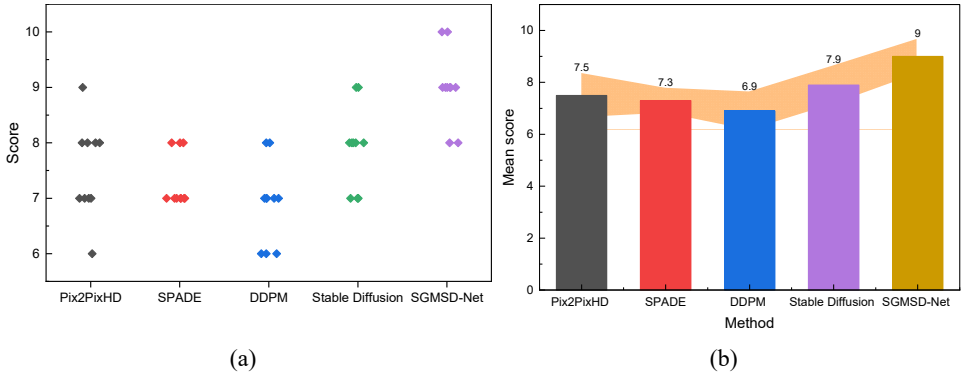


Figure 7(a) presents the distribution of individual user scores for different methods. It can be observed that SGMSD-Net receives higher and more concentrated scores than the baseline methods, indicating that users generally perceive its generated images as clearer, more realistic, and more semantically consistent. In contrast, the scores of Pix2PixHD, SPADE, and DDPM fluctuate more obviously, suggesting that their generated results are less stable in subjective visual quality. Figure 7(b) further compares the mean user scores of different methods. SGMSD-Net achieves the highest average score of 9.0, outperforming Stable Diffusion, DDPM, SPADE, and Pix2PixHD. This result confirms that the proposed method not only improves objective evaluation metrics but also provides a better subjective visual experience.

4.4 Ablation study of multi-scale diffusion design

To investigate the effectiveness of the proposed hierarchical diffusion strategy, an ablation experiment is conducted by varying the number of diffusion stages while keeping other components unchanged. As shown in Table 4 and Table 5, the three-stage diffusion architecture consistently achieves the best performance on both SUN360 and Matterport3D datasets. Compared with the single-stage diffusion model, SGMSD-Net reduces the FID by 20.8% and 21.7% on SUN360 and Matterport3D, respectively, while improving the CLIP Score by 2.4% and 2.8%.

Table 4 Ablation study of diffusion stages on SUN360

<i>Model</i>	<i>Diffusion stage</i>	<i>FID ↓</i>	<i>LPIPS ↓</i>	<i>CLIP score ↑</i>
SGMSD-Net (Single-stage)	1	18.72	0.162	0.873
SGMSD-Net (Two-stage)	2	16.45	0.151	0.884
SGMSD-Net (Three-stage)	3	14.83	0.139	0.897

Table 5 Ablation study of diffusion stages on Matterport3D datasets

<i>Model</i>	<i>Diffusion stage</i>	<i>FID ↓</i>	<i>LPIPS ↓</i>	<i>CLIP score ↑</i>
SGMSD-Net (Single-stage)	1	21.36	0.178	0.851
SGMSD-Net (Two-stage)	2	18.94	0.165	0.865
SGMSD-Net (Three-stage)	3	16.72	0.152	0.879

The results indicate that the performance improvement does not simply originate from increasing model complexity, but mainly benefits from the coarse-to-fine generation mechanism. The first diffusion stage focuses on learning global scene layout and spatial structure, while subsequent stages progressively refine object boundaries, textures, and illumination details. By decomposing high-resolution image synthesis into multiple hierarchical generation steps, the proposed strategy provides more effective feature representations for complex panoramic tourism scenes.

4.5 Computational cost analysis

To evaluate the computational overhead introduced by the proposed multi-scale diffusion architecture, the parameter number, floating-point operations (FLOPs), inference time, and GPU memory consumption are reported in Table 6. All experiments are conducted under the same hardware configuration. SGMSD-Net introduces additional computational cost compared with conventional DDPM due to the hierarchical diffusion process. However, compared with large-scale diffusion models such as Stable Diffusion, SGMSD-Net maintains a relatively moderate computational requirement while achieving better semantic alignment and panoramic generation quality.

Table 6 Computational cost comparison

<i>Method</i>	<i>Params (M)</i>	<i>FLOPs (G)</i>	<i>Inference time (s/image)</i>	<i>GPU memory (GB)</i>
DDPM	124.6	218.5	8.7	12.4
Stable diffusion	859.5	512.8	12.6	15.8
SGMSD-Net	356.2	286.4	10.3	13.9

Although the three-stage diffusion structure requires multiple denoising operations, the hierarchical generation strategy avoids directly optimising all high-resolution details in a single step. By progressively refining global structures and local textures, the proposed framework achieves a better balance between generation quality and computational efficiency.

5 Discussion

The proposed SGMSD-Net integrates a semantic encoding module, a multi-scale diffusion generation module, and a semantic-guided loss mechanism to address weak semantic control, insufficient visual detail, and poor panoramic consistency in tourism image generation. Experimental results on SUN360 and Matterport3D show that SGMSD-Net outperforms Pix2PixHD, SPADE, DDPM, and Stable Diffusion in FID, LPIPS, and CLIP Score, indicating improvements in visual realism, perceptual consistency, and semantic alignment within the evaluated datasets.

The semantic encoding module maps structured information, including scene categories, landmark types, and weather conditions, into continuous semantic vectors through an embedding layer and a multilayer perceptron. The generated modulation parameters γ and β are then applied to intermediate UNet features through FiLM. This mechanism provides direct channel-level control over the generation process and enables different semantic inputs to produce corresponding visual characteristics. The average semantic attribute accuracy of 92.9% further demonstrates that the model can preserve fine-grained attributes rather than merely achieving global image-text similarity. Related research has also shown that language-guided constraints can help maintain semantic information during visual optimisation (Liu et al., 2025), supporting the combination of perceptual and semantic consistency objectives adopted in SGMSD-Net.

FiLM itself is not a newly proposed conditioning mechanism. The contribution of this study lies in adapting it to immersive tourism panoramic generation by explicitly modelling structured tourism-related conditions. Compared with cross-attention, FiLM introduces relatively low computational overhead and is suitable for categorical and attribute-based semantic inputs. However, its ability to process long descriptions, ambiguous expressions, and complex relationships remains limited. Future work may therefore combine FiLM-based attribute control with cross-attention-based multimodal interaction to support more detailed natural-language instructions.

The multi-scale diffusion module divides image generation into three resolution stages. The low-resolution stage establishes the global layout and colour distribution, the intermediate stage refines spatial structures and object boundaries, and the high-resolution stage restores textures and illumination details. This coarse-to-fine process reduces the difficulty of simultaneously learning global and local information in a single high-resolution diffusion model. The ablation results show progressive improvements from the single-stage to the three-stage configuration, confirming that the performance gain mainly results from hierarchical visual refinement. Nevertheless, the use of multiple diffusion stages also increases parameter size, inference time, and memory consumption. Therefore, the proposed framework is currently more suitable for offline virtual-tour production, exhibition design, and cultural heritage visualisation than for real-time or resource-constrained applications.

The joint loss function further improves generation quality by combining diffusion noise prediction loss, perceptual consistency loss, and semantic consistency loss. These objectives constrain the generated images at the noise, feature, and semantic levels, helping the model maintain both realistic visual details and thematic correspondence. This design is particularly important for tourism scenes, where visually attractive images may still be unsuitable when landmarks, environmental conditions, or cultural characteristics do not match the input descriptions.

Beyond algorithmic performance, SGMSD-Net provides a practical approach to virtual tourism and digital cultural tourism content production. Users can generate panoramic scenes from relatively simple semantic descriptions, reducing the technical requirements for tourism organisations, museums, and non-professional content creators. Smart tourism research has demonstrated the value of human-computer interaction and user-oriented intelligent services (Yan et al., 2025). In this context, SGMSD-Net may extend tourism systems from destination recommendation to personalised visual-content generation. The generated scenes may also be integrated with virtual reality and augmented reality systems for virtual tours, urban planning, and interactive exhibitions. Existing research on VR-based virtual environments indicates that realistic visual scenes can enhance immersive perception and interaction (Wu et al., 2025), providing support for the further integration of SGMSD-Net with immersive display technologies.

Several limitations should nevertheless be recognised. First, SUN360 and Matterport3D have limited geographical and cultural coverage, and the current results cannot fully demonstrate cross-domain generalisation. Second, incomplete or ambiguous semantic labels may cause inconsistent landmarks, environmental conditions, or spatial structures. Semantic normalisation and prompt-enhancement strategies should therefore be introduced. Third, the subjective evaluation involves a relatively small number of participants and should be expanded with more diverse users and inter-rater reliability analysis. Finally, highly realistic generated tourism images may raise concerns regarding authenticity, cultural misrepresentation, and inappropriate use. Synthetic content should be clearly identified, and human review or cultural-expert verification should be incorporated into practical applications.

6 Conclusions

This paper addresses the challenges of insufficient semantic control, lack of image detail, and limited generalisation across multiple scenes in immersive tourism content generation by proposing a SGMSD-Net. Built upon a diffusion generation framework, the model integrates a semantic modulation mechanism with a multi-scale structured generation strategy, achieving end-to-end optimisation from semantic understanding to high-resolution image synthesis. Specifically, the model first extracts semantic features of tourism scenes through a semantic embedding module, employing FiLM modulation to dynamically incorporate these features into a multi-scale diffusion network, thereby achieving deep coupling between semantic and visual representations. Next, a hierarchical multi-scale diffusion structure progressively reconstructs scene layout, detailed textures, and lighting characteristics from low to high resolution, ensuring spatial coherence and visual realism. Finally, a joint optimisation of semantic consistency loss and perceptual loss enables multi-level refinement from the pixel layer to the semantic layer.

Experimental results demonstrate that SGMSD-Net achieves superior performance on two benchmark datasets: it outperforms Pix2PixHD, SPADE, DDPM, and Stable Diffusion across all three metrics – FID, LPIPS, and CLIP Score. The generated images exhibit significant advantages in semantic alignment, detail fidelity, and immersive realism. Subjective user evaluations further confirm the superiority of SGMSD-Net in both visual experience and content consistency. These findings indicate that SGMSD-Net provides a new technical pathway for high-quality and semantically controllable tourism image generation, laying a solid foundation for intelligent cultural tourism and digital content creation.

Despite its strong performance, SGMSD-Net still has certain limitations. First, its detail reconstruction capability and generation efficiency require further improvement when dealing with ultra-high-resolution scenes. Second, its generalisation ability under extreme lighting conditions and complex spatial structures remains limited; in low-light or highly variable multi-view environments, structural distortion or semantic drift may still occur. Moreover, the current model primarily relies on fixed semantic labels for guidance and lacks comprehensive understanding and dynamic control of multimodal semantics. Future research may proceed in the following directions:

- 1 integrating large-scale multimodal pre-trained models to enhance semantic representation and cross-modal understanding
- 2 exploring controllable diffusion mechanisms by introducing interpretable conditional controls and user-interactive interfaces for flexible adjustment of image structure, style, and semantic hierarchy
- 3 investigating collaborative generation methods that combine diffusion-based synthesis with 3D reconstruction, virtual touring, and spatial perception modelling, to achieve high-fidelity restoration from 2D images to panoramic scenes.

Through these extensions, SGMSD-Net has the potential to further advance the deep application of artificial intelligence in digital cultural tourism, cultural heritage preservation, and immersive visual communication, providing a new research paradigm and technological foundation for the interpretability, personalisation, and practical deployment of intelligent visual generation systems. In addition, future studies will further investigate the integration of more recent large-scale diffusion architectures and controllable generation frameworks to evaluate the scalability of SGMSD-Net under broader generation scenarios.

Funding

This work was supported by the major project of Anhui Provincial Universities Scientific Research, ‘Research on Visual Reconstruction and New Ecology of Intelligent Communication of Anhui River Culture (Maanshan Section) Brand Based on AIGC Technology’ (research code: 2024AH040372), and the key project of Anhui Provincial Quality Engineering Education and Teaching Management, ‘Research on the Construction of Visual Communication Design Specialty in Private Universities in the Context of Evaluation and Certification and Practice’ (research code: 2022jxgl068).

Declarations

All authors declare that they have no conflicts of interest.

References

- Acharya, M. and Mohbey, K.K. (2023) ‘Trust-aware spatial-temporal feature estimation for next POI recommendation in location-based social networks’, *Social Network Analysis and Mining*, Vol. 13, No. 1, p. 102, DOI: 10.1007/s13278-023-01106-8.
- Alblwi, A., Makkawy, S. and Barner, K.E. (2025) ‘D-DDPM: deep denoising diffusion probabilistic models for lesion segmentation and data generation in ultrasound imaging’, *IEEE Access*, Vol. 13, pp.41194–41209, DOI: 10.1109/ACCESS.2025.3548128.
- Chen, H., Li, S. and Li, J. (2023) ‘An end-to-end network for upright adjustment of panoramic images’, *Procedia Computer Science*, Vol. 222, pp.435–447, DOI: 10.1016/j.procs.2023.08.182.
- Chen, W., Zhu, Y., Chang, Y., Luo, K., Wen, H., Li, L., Yu, Y., Wen, Q., Chen, C., Zheng, K., Gao, Y., Zheng, Y., Zhou, X. and Liang, Y. (2026) ‘Trajectory data management and mining: a survey from deep learning to the LLM era’, *arXiv*, DOI: 10.48550/arXiv.2403.14151.
- Daneshfar, F., Bartani, A. and Lotfi, P. (2024) ‘Image captioning by diffusion models: a survey’, *Engineering Applications of Artificial Intelligence*, Vol. 138, p.109288, DOI: 10.1016/j.engappai.2024.109288.
- Du, C., Li, Y., Qiu, Z. and Xu, C. (2023) ‘Stable diffusion is unstable’, in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS ’23)*, Curran Associates Inc., Red Hook, NY, USA, pp.58648–58669.
- Fu, L., Yi, Y., Liu, L. and Chen, R. (2025) ‘Designing a novel technique for evaluation of tourism informatization in scenic spots from a big data perspective’, *PeerJ Computer Science*, Vol. 11, p.e2807, DOI: 10.7717/peerj-cs.2807.
- Gao, N., Lu, Y., Chen, P., Sun, G., Liang, R. and Zhang, Y. (2025) ‘TWDT: training-free word-level controllable diffusion model for text generation’, *Knowledge-Based Systems*, Vol. 330, p.114437, DOI: 10.1016/j.knosys.2025.114437.
- Gerogiannis, G., Yesil, S., Lenadora, D., Cao, D., Mendis, C. and Torrellas, J. (2023) ‘SPADE: a flexible and scalable accelerator for SpMM and SDDMM’, in *Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA’23)*, Association for Computing Machinery, New York, NY, USA, pp.1–15, DOI: 10.1145/3579371.3589054.
- Hu, X., Zhou, Z., Li, H., Hu, Y., Gu, F., Kersten, J., Fan, H. and Klan, F. (2023) ‘Location reference recognition from texts: a survey and comparison’, *ACM Computing Surveys*, Vol. 56, No. 5, p. 112:1–112:37, DOI: 10.1145/3625819.
- Huang, L., Miron, A., Hone, K. and Li, Y. (2024) ‘Segmenting medical images: from UNet to Res-UNet and nnUNet’, in *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, Piscataway, NJ, USA, pp.483–489, DOI: 10.1109/CBMS61543.2024.00086.
- Huang, X. (2022) ‘Personalized travel route recommendation model of intelligent service robot using deep learning in big data environment’, *Journal of Robotics*, Vol. 2022, No. 1, p.7778592, DOI: 10.1155/2022/7778592.
- Jang, W., Kim, S., Chun, J.W., Jung, A-R. and Kim, H. (2023) ‘Role of recommendation sizes and travel involvement in evaluating travel destination recommendation services: comparison between artificial intelligence and travel experts’, *Journal of Hospitality and Tourism Technology*, Vol. 14, No. 3, pp.401–415, DOI: 10.1108/JHTT-01-2022-0013.
- Ke, W., Guo, Y., Liu, Q., Chen, W., Wang, P., Luo, H. and Luo, Z. (2024) ‘MDM: meta diffusion model for hard-constrained text generation’, *Knowledge-Based Systems*, Vol. 283, p.111147, DOI: 10.1016/j.knosys.2023.111147.

- Li, H., Xu, F. and Lin, Z. (2023) 'ET-DM: text to image via diffusion model with efficient transformer', *Displays*, Vol. 80, p.102568, DOI: 10.1016/j.displa.2023.102568.
- Liu, Q., Lu, F., Wang, N., Zhang, J. and Hu, Y. (2025) 'Transformer-based infrared and visible image fusion with language-driven loss', *International Journal of Ad Hoc and Ubiquitous Computing*, Vol. 49, No. 3, pp.143–152, DOI: 10.1504/IJAHUC.2025.147567.
- Mou, N., Jiang, Q., Zhang, L., Niu, J., Zheng, Y., Wang, Y. and Yang, T. (2022) 'Personalized tourist route recommendation model with a trajectory understanding via neural networks', *International Journal of Digital Earth*, Vol. 15, No. 1, pp.1738–1759, DOI: 10.1080/17538947.2022.2130456.
- Nitu, P., Coelho, J. and Madiraju, P. (2021) 'Improvising personalized travel recommendation system with recency effects', *Big Data Mining and Analytics*, Vol. 4, No. 3, pp.139–154, DOI: 10.26599/BDMA.2020.9020026.
- Park, E., Park, J. and Hu, M. (2021) 'Tourism demand forecasting with online news data mining', *Annals of Tourism Research*, Vol. 90, p.103273, DOI: 10.1016/j.annals.2021.103273.
- Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., Savva, M., Zhao, Y. and Batra, D. (2021) 'Habitat-Matterport 3D Dataset (HM3D): 1000 large-scale 3D environments for embodied AI', *arXiv*, DOI: 10.48550/arXiv.2109.08238.
- Shapoval, V., Wang, M., Hara, T. and Shioya, H. (2018) 'Data mining in tourism data analysis: inbound visitors to Japan', *Journal of Travel Research*, Vol. 57, No. 3, pp.310–323, DOI: 10.1177/0047287517696960.
- Wang, G., Wang, H. and Wang, L. (2022) 'Research trends in tourism and hospitality from 1991 to 2020: an integrated approach of corpus linguistics and bibliometrics', *Journal of Hospitality and Tourism Insights*, Vol. 6, No. 2, pp.509–529, DOI: 10.1108/JHTI-09-2021-0260.
- Wang, T., He, J. and Ning, Z. (2023) 'Observations on the development of townships around mega-tourist attraction caused by land use change – the case of Zhangjiajie, China', *Journal of Asian Architecture and Building Engineering*, Vol. 22, No. 5, pp.3006–3027, DOI: 10.1080/13467581.2023.2167494.
- Wu, S., Chen, H. and Wei, N. (2025) 'Visual safety in the meta-universe using VR technology', *International Journal of Ad Hoc and Ubiquitous Computing*, Vol. 50, No. 2, pp.61–72, DOI: 10.1504/IJAHUC.2025.149457.
- Yan, J., Wang, M. and Xue, R. (2025) 'Hybrid recommendation system based on human-computer interaction technology and smart tourism', *International Journal of Ad Hoc and Ubiquitous Computing*, Vol. 48, No. 1, pp.1–12, DOI: 10.1504/IJAHUC.2025.143545.
- Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B. and Yang, M.-H. (2023) 'Diffusion models: a comprehensive survey of methods and applications', *ACM Computing Surveys*, Vol. 56, No. 4, pp.105:1–105:39, DOI: 10.1145/3626235.
- Yu, J., Lu, K., Du, S., Xu, L., Chang, P., Liu, H., Lan, B. and Liu, T. (2024) 'MvAV-pix2pixHD: multi-view aerial view image translation', in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Piscataway, NJ, USA, pp.3066–3075, DOI: 10.1109/CVPRW63382.2024.00312.
- Zhang, J., Ma, M., Gao, X. and Chen, G. (2024) 'Encoder-decoder based route generation model for flexible travel recommendation', *IEEE Transactions on Services Computing*, Vol. 17, No. 3, pp.905–920, DOI: 10.1109/TSC.2024.3376231.
- Zhou, Y., Zhou, D., Wang, Y., Feng, J. and Hou, Q. (2025) 'MaskDiffusion: boosting text-to-image consistency with conditional mask', *International Journal of Computer Vision*, Vol. 133, No. 5, pp.2805–2824, DOI: 10.1007/s11263-024-02294-2.