2025 Design, Automation & Test in Europe Conference (DATE 2025)

AeroDiffusion: Complex Aerial Image Synthesis with Keypoint-Aware Text Descriptions and Feature-Augmented Diffusion Models

Douglas J. Townsell[1], Mimi Xie[2], Bin Wang[1], Fathi Amsaad[1], Varshitha Reddy Thanam[1], Wen Zhang*[1]
[1]Department of Computer Science and Engineering, Wright State University, Dayton, OH, USA
[2]Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, USA
{townsell.4, bin.wang, fathi.amsaad, thanam.2, wen.zhang}@wright.edu, mimi.xie@utsa.edu

***Abstract*—Aerial imagery provides crucial insights for various fields, including remote monitoring, environmental assessment, and autonomous navigation. However, the availability of aerial image datasets is limited due to privacy concerns and imbalanced data distribution, impeding the development of robust deep learning models. Recent advancements in text-guided image synthesis offer a promising approach to enrich and diversify these datasets. Despite progress, existing generative models face challenges in synthesizing realistic aerial images due to the lack of paired text-aerial datasets, the complexity of densely packed objects, and the limitations of modeling object relationships. In this paper, we introduce AeroDiffusion, a novel framework designed to overcome these challenges by leveraging large language models (LLMs) for keypoint-aware text description generation and a feature-augmented diffusion process for realistic image synthesis. Our approach integrates region-level feature extraction to preserve small objects and multi-modal feature alignment to improve textual descriptions of complex aerial scenes. AeroDiffusion is the first to extend deep generative models for high-resolution, text-guided aerial image generation, including the creation of images from novel viewpoints. We contribute a new paired text-aerial image dataset and demonstrate the effectiveness of our model, achieving an FID score of 78.15 across five benchmarks, significantly outperforming state-of-the-art models such as DDPM (217.95), Stable Diffusion (119.13), and ARLDM (111.59).**

I. Introduction

Aerial imagery, characterized by its top-down perspective, detailed semantic information, and extensive coverage, provides invaluable insights across a wide range of domains [1]–[5]. Drones, capable of capturing aerial images from various angles while maintaining mobility, have enabled the application of aerial image datasets in fields such as remote monitoring, environmental assessment, and autonomous navigation [6]–[9]. This capability enhances situational awareness by offering a comprehensive view from above, supporting critical real-time decision-making in diverse environments. However, due to privacy concerns, the availability of aerial image datasets is limited, which impedes the development of deep learning models for aerial image datasets and restricts the full potential of utilizing aerial images in downstream intelligent tasks. Moreover, the data distribution in aerial imagery is often poor. In autonomous driving, images are available from multiple viewpoints (e.g., 6 viewpoints in the NuScenes dataset [10]) and under various conditions, such as cloudy or sunny days, and daytime or nighttime images for the same location/scenario. In contrast, drone-captured image datasets are usually limited, with few images available per scenario, leading to unbalanced data distribution.

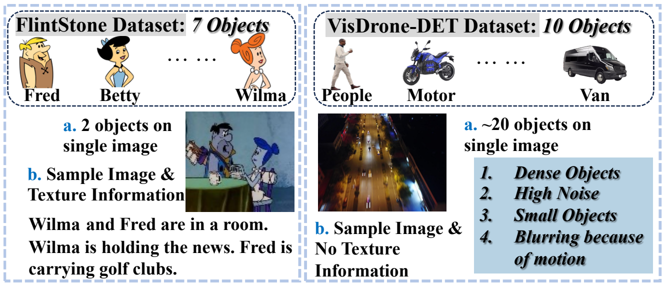

FlintStone Dataset: 7 Objects
Fred
Betty
… …
Wilma
a. 2 objects on single image
b. Sample Image & Texture Information
Wilma and Fred are in a room. Wilma is holding the news. Fred is carrying golf clubs.
VisDrone-DET Dataset: 10 Objects
People
Motor
… …
Van
a. ~20 objects on single image
b. Sample Image & No Texture Information
1. Dense Objects
2. High Noise
3. Small Objects
4. Blurring because of motion

Fig. 1: Comparing Classical Image Synthesis Dataset to Aerial Image Dataset: VisDrone-DET [14] vs. FlintStone [15].

Recent advancements in deep conditional generative models, particularly in text-guided image synthesis [11]–[13], have revolutionized image generation, enabling the production of high-quality images even from imperfect inputs. These models have significant potential for **text-guided aerial image synthesis**, offering not only the ability to enrich aerial image datasets but also to enhance the diversity of the data set, resulting in improved data distribution. For instance, in an aerial surveillance task, a training dataset might only include images of "building A from a top-down angle," "building A from a 45-degree angle" and "building B from a top-down angle." Conditional interpolation can generate the missing condition, "building B from a 45-degree angle.' By leveraging knowledge learned from existing conditions, deep generative models can synthesize images from various angles, helping to create a more comprehensive and diverse augmented dataset for model training.

Despite these advances, the direct application of existing deep generative models to text-guided aerial image synthesis faces substantial challenges and often fails to generate realistic aerial images. **1)** A primary limitation is the absence of readily available paired text-aerial image datasets, which hinders effective model training. **2)** Additionally, aerial imagery typically contains small, densely packed objects that are prone to being discarded during the subsampling process in feature extraction. This issue is exacerbated by the complexity of aerial scenes, which can include up to $\sim$ 90 objects per image, in contrast to classical image datasets that generally

*Corresponding author: Wen Zhang (wen.zhang@wright.edu)

978-3-9826741-0-0/DATE25/© 2025 EDAA

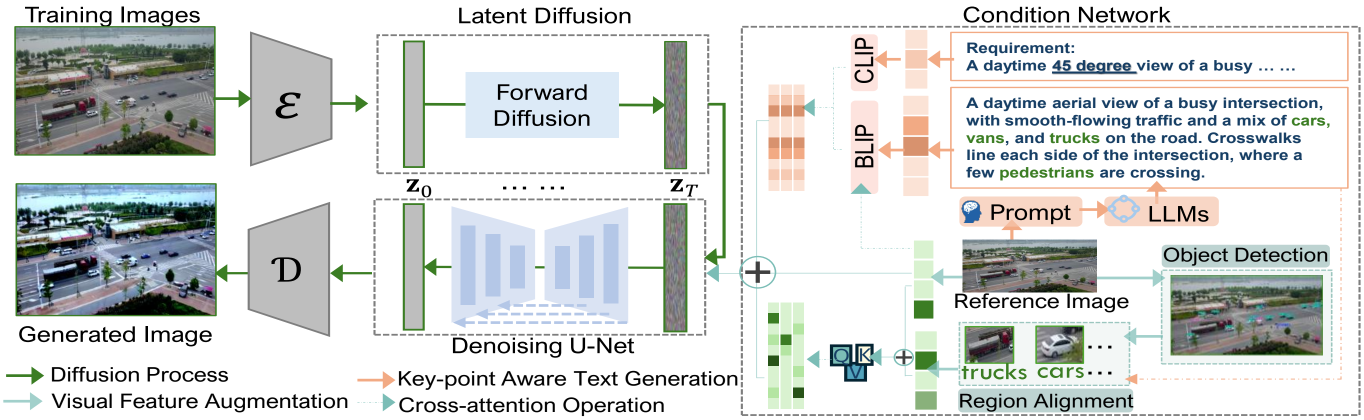

Training Images
Latent Diffusion
Condition Network
$\mathcal{E}$
Forward Diffusion
$\mathbf{z}_0$
… …
$\mathbf{z}_T$
$\mathcal{D}$
Denoising U-Net
Generated Image
CLIP
BLIP
Requirement:
A daytime 45 degree view of a busy … …
A daytime aerial view of a busy intersection, with smooth-flowing traffic and a mix of cars, vans, and trucks on the road. Crosswalks line each side of the intersection, where a few pedestrians are crossing.
Prompt
LLMs
Object Detection
Reference Image
Q K V
trucks cars…
Region Alignment
Diffusion Process
Visual Feature Augmentation
Key-point Aware Text Generation
Cross-attention Operation

Fig. 2: An overview of our proposed model, **AeroDiffusion**, for generating complex aerial images.

feature only 1 or 2 objects. **3)** While large language models (LLMs) [16] could potentially be leveraged to generate textual descriptions for aerial images, the inherent complexity of aerial scenes results in lengthy and frequently inaccurate descriptions. **4)** Moreover, the high object density within these images complicates modeling the correct spatial arrangement of objects. Existing methods continue to struggle with synthesizing multiple objects with accurate positional relationships [17]. This challenge is magnified in aerial imagery, where each image may contain between ~ 20 and ~ 90 objects, as shown in Fig. 1.

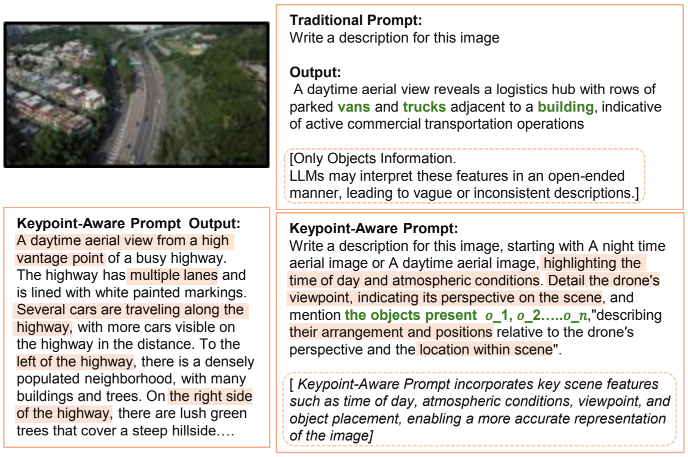
To address these challenges, this paper introduces **AeroDiffusion**, the first approach to achieve complex aerial image synthesis by leveraging large language models (LLMs) [16] to dynamically generate text descriptions and enhance the diffusion process with region-level feature augmentations. The core framework, depicted in Fig. 2, consists of two key components: keypoint-aware text description generation for complex aerial images and a feature-augmented diffusion process to produce realistic and explainable imagery. Specifically, LLMs are leveraged to dynamically generate effective textual descriptions of complex aerial images by learning image features through multi-modal feature alignment. This enables the generation of detailed text representations that capture deep-level image features. Subsequently, the text-guided aerial image synthesis is achieved through a feature-augmented diffusion process. To mitigate the loss of small objects during the subsampling process, an object retrieval mechanism is employed to identify and extract regions of interest (ROIs) from the images, ensuring critical features are preserved for subsequent processing. These retrieved regions are integrated with the source image features via multi-head attention, enhancing the representation of specific objects. Finally, the augmented features, combined with the synthesized text descriptions, form conditional vectors within the latent diffusion process, enabling precise and context-aware text-guided aerial image generation. The main contributions of this paper are summarized as follows:

(1) We propose **AeroDiffusion**, a novel framework capable of generating complex aerial scenes, even generating aerial images from different viewpoints based on a reference image. To the best of our knowledge, this is the first work to extend deep generative models for text-guided high-resolution aerial image synthesis using a latent diffusion model.
(2) We contribute a publicly available paired text-aerial image dataset, accessible at https://github.com/NolimitDougie/AeroDiffusion. Furthermore, we enhance latent diffusion by integrating keypoint-aware text descriptions, feature-augmented image representations, and multi-modal embedding fusion. These innovations enable the diffusion model to capture small object features, deep-level patterns, and dependencies between text and aerial images, thereby enhancing denoising and generating high-quality outputs.
(3) We validate the effectiveness of AeroDiffusion on five benchmarks, achieving an FID score of 78.15, significantly outperforming SOTA models such as DDPM (217.95), Stable Diffusion (119.13), ARLDM (111.59), Versatile Diffusion (124.12), and Make-a-Scene (114.75).

II. Related Work

Deep generative models. Deep generative models have shown great potential in image generation, with recent advancements in text-to-image synthesis categorized into *Generative Adversarial Networks (GANs)* [18], *Variational Autoencoders (VAEs)* [19], autoregressive models [20], and *Diffusion Models* [21]. While GANs produce high-resolution images, they struggle with semantic understanding, whereas VAEs offer more stable optimization. Diffusion models, such as Imagen [22], DALL-E 2 [23], and Stable Diffusion [24], have recently excelled in generating photorealistic images from text prompts.

Aerial Image Synthesis. Despite advancements in deep generative models, aerial image synthesis remains underexplored. Existing approaches primarily focus on object detection [14] or image-to-image translation [22], leveraging architectures such as CycleGAN [25] and Pix2Pix [26]. However, these methods often fail to capture the intricate spatial relationships and fine-grained details required for text-guided aerial image generation. Recent efforts like SatSynth [27] employ diffusion models for satellite imagery but are constrained to segmentation tasks, leaving the synthesis of complex aerial scenes with detailed object representation under challenging conditions largely unaddressed.



Traditional Prompt:
Write a description for this image

Output:
A daytime aerial view reveals a logistics hub with rows of parked **vans** and **trucks** adjacent to a **building**, indicative of active commercial transportation operations

[Only Objects Information.
LLMs may interpret these features in an open-ended manner, leading to vague or inconsistent descriptions.]

Keypoint-Aware Prompt:
Write a description for this image, starting with A night time aerial image or A daytime aerial image, highlighting the time of day and atmospheric conditions. Detail the drone's viewpoint, indicating its perspective on the scene, and mention **the objects present *o_1*, *o_2*.....*o_n***,"describing their arrangement and positions relative to the drone's perspective and the location within scene".

[*Keypoint-Aware Prompt incorporates key scene features such as time of day, atmospheric conditions, viewpoint, and object placement, enabling a more accurate representation of the image]*

Keypoint-Aware Prompt Output:
A daytime aerial view from a high vantage point of a busy highway. The highway has multiple lanes and is lined with white painted markings. Several cars are traveling along the highway, with more cars visible on the highway in the distance. To the left of the highway, there is a densely populated neighborhood, with many buildings and trees. On the right side of the highway, there are lush green trees that cover a steep hillside….

Fig. 3: An Example of Keypoint-Aware Text Generation.

III. Problem Statement

Given an aerial image dataset consisting solely of RGB images, denoted as $\mathcal{X} = \{\mathbf{X}_i\}_{i=1}^{n}$, where each $\mathbf{X}_i \in \mathbb{R}^{H\times W\times 3}$, our goal is to generate high-resolution aerial images that are spatially coherent and effectively capture small, densely packed objects. To achieve this, we propose a diffusion-based approach for aerial image synthesis using a sequence of conditional models, denoted as $\epsilon\theta(\mathbf{X}_t, t, \mathbf{C})$; $t = 1, \cdots, T$, where the image generation process is guided by a condition vector $\mathbf{C}$ that encodes semantic and feature-level information. Unlike traditional autoencoders that rely on simple prompts or a single image as the condition vector, we enhance $\mathbf{C}$ by refining both the textual descriptions of $\mathbf{X}$ through keypoint-aware text generation and augmenting region-specific features. These aligned text and feature representations provide richer semantic guidance throughout the generation process, ensuring greater fidelity and detail in the synthesized imagery.

IV. Proposed Model

As illustrated in Fig. 2, we begin by generating keypoint-aware text descriptions for each aerial image in the dataset, addressing both the absence of paired text-aerial image datasets and the challenge of accurately describing complex images for downstream image generation tasks. Next, we enhance the image component through region-level feature augmentation, refining image representations to overcome the difficulties posed by small and densely packed objects that are characteristic of aerial datasets. Finally, these high-quality text descriptions and enhanced image features are aligned at a deep level, enabling robust multimodal integration. This alignment supports the image generation process within the latent diffusion framework, ensuring the generation of realistic and contextually relevant aerial images.

A. Keypoint-Aware Text Generation

To integrate text guidance into **AeroDiffusion**, we synthesize detailed textual descriptions for each aerial image to facilitate text processing and embedding. Traditionally, labeling text descriptions for images relies on simple prompts like "write a text description for the input image." or only a caption for the image. This approach often leads to rough descriptions for complex images, failing to capture subtle differences between similar images (resulting in identical descriptions for similar images), and provides limited guidance for the image generation process, thereby reducing the controllability of the generated outputs. For example, as shown in Fig. 3, a description like "A daytime aerial view reveals a logistics hub with rows of parked vans and trucks adjacent to a building, indicative of active commercial transportation operations," lacks the level of detail needed to guide the generation process effectively. To address this, we first manually create specific prompts for the aerial image dataset. We then leverage the zero-shot capabilities of large language models (LLMs) [16] and employ chain-of-thought prompting [28] to more precisely guide the LLMs in capturing the specific details (Keypoints) we aim to extract from the images in our generated textual descriptions, as shown in Fig. 3. Specifically, given an image X_i, we carefully design a prompt template P_i to instruct the LLMs on articulating key aspects of the image, such as scene viewpoint, composition, lighting conditions (day/night), objects present, and their specific locations within the frame. Defining the object list of image X_i as $\mathcal{O}_i$. Using this template P_i, we prompt the LLMs via black-box APIs to dynamically extract the text description G_i for each image, as expressed by:

$$G_i = \mathrm{LLM}(\mathbf{X}_i, \mathcal{O}_i, P_i),\ \mathbf{X}_i \in \mathcal{X},\ G_i \in \mathcal{G}. \tag{1}$$

B. Region-Level Feature Augmentation

To address the challenges of small, densely packed objects and the complexities of spatial arrangements in aerial images, our core idea begins by identifying key objects in the image through object detection, designating them as regions of interest (ROIs). We then employ a cross-attention mechanism to align the visual features extracted from these ROIs with the text embeddings generated from the corresponding region-level descriptions. The region information $\mathcal{R}$ not only provides bounding box coordinates but also includes the indices of the text labels most relevant to each region, facilitating cross-modal interaction between the visual content and its associated textual description. Finally, these enhanced features are integrated with the original image feature map using multi-head attention, effectively augmenting the small objects representations.

Concrecetly, we first train YOLO [14], [29] on the aerial dataset to detect the ROIs. Given an image $\mathbf{X}_i \in \mathcal{X}$, the trained YOLO model detects objects within the image, generating a set of regions of interest, $\mathcal{R}$, where each $\mathbf{R} \in \mathcal{R}$ is resized to match the dimensions of the original image. Each region is associated with a bounding box $\mathcal{B}$, which defines the object's location, and the corresponding text labels L, representing the object categories. Once the ROIs are extracted, we employ a cross-attention mechanism to align the visual features derived from the regions, $\mathcal{B}$, with text embeddings generated from the object-level descriptions, L. This process produces a feature set $[\mathbf{f}_{\mathbf{X}_{i,1}}, \mathbf{f}_{\mathbf{X}_{i,2}}, \ldots, \mathbf{f}_{\mathbf{X}_{i,R}}]$ for the detected regions. The features from these regions are then concatenated with the original image features to form an aggregated feature set $F = [\mathbf{f}_{\mathbf{X}_i}, \mathbf{f}_{\mathbf{X}_{i,1}}, \mathbf{f}_{\mathbf{X}_{i,2}}, \ldots, \mathbf{f}_{\mathbf{X}_{i,R}}]$. To further enhance the representation of the image, we apply multi-head self-attention to the aggregated feature set $\mathbf{F}$. This process enables the model to dynamically determine the relevance and importance of different regions of the source image. The multi-head attention mechanism works by computing self-attention scores for each head, as follows:

$$\text{Attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \text{softmax}\left(\mathbf{Q}\mathbf{K}^{\mathbf{T}}/\sqrt{d_k}\right)\mathbf{V} \qquad (2)$$

where $\sqrt{d_k}$ serves as a scaling factor to stabilize gradients and $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ matrices are obtained from the input F through learned linear transformations:

$$\begin{cases} \mathbf{Q} = \mathbf{W}^{Q}[\mathbf{f}_{\mathbf{X}_i}, \mathbf{f}_{\mathbf{X}_{i,1}}, \mathbf{f}_{\mathbf{X}_{i,2}}, \ldots, \mathbf{f}_{\mathbf{X}_{i,R}}], \\ \mathbf{K} = \mathbf{W}^{K}[\mathbf{f}_{\mathbf{X}_i}, \mathbf{f}_{\mathbf{X}_{i,1}}, \mathbf{f}_{\mathbf{X}_{i,2}}, \ldots, \mathbf{f}_{\mathbf{X}_{i,R}}], \\ \mathbf{V} = \mathbf{W}^{V}[\mathbf{f}_{\mathbf{X}_i}, \mathbf{f}_{\mathbf{X}_{i,1}}, \mathbf{f}_{\mathbf{X}_{i,2}}, \ldots, \mathbf{f}_{\mathbf{X}_{i,R}}]. \end{cases} \qquad (3)$$

The attention-enhanced features $\hat{\mathbf{f}}_{\mathbf{X}}$ are derived from the attention tensor, forming a comprehensive representation of the enriched source image. These refined features are subsequently utilized to construct the condition vector, which guides the conditioning network.

C. Text-Aerial Image Synthesizing

We adopt the latent diffusion model [24] as the underlying framework to realize text-guided aerial image synthesis. The framework is composed of three key components: a forward diffusion process, a conditioning network to construct the condition vector, and a denoising network to remove noise and synthesize the new image guided by the condition vector.

1) Forward Diffusion. The process begins with an initial training image $\mathbf{X}_i$, which is encoded by the encoder $\mathcal{E}$ to obtain its latent representation, $\mathbf{z}_0 = \mathcal{E}(\mathbf{X}_i)$. Gaussian noise is incrementally added to $\mathbf{z}_0$ over T diffusion steps, progressively generating noisy latent features $\mathbf{z}_T$ through the forward diffusion process, defined as:

$$\begin{aligned} q(\mathbf{z}_t|\mathbf{z}_{t-1}) &= \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}); \beta_{t-1} < \beta_t \quad (4) \\ q(\mathbf{z}_{1:T}|\mathbf{z}_0) &= \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{z}_{t-1}); \beta_t \in (0,1) \end{aligned}$$

Here, β_t represents the step size parameter for each diffusion step, ensuring the gradual addition of noise throughout the process[21]. $\mathbf{I}$ is the identity matrix.

2) Feature-Augmented Condition Network. AeroDiffusion is designed to enable controllable image generation by utilizing a conditional vector $\mathbf{C}$, which seamlessly integrates the source aerial image $\mathbf{X}_i$, its corresponding textual descriptions G_i, and feature-augmented representations $\hat{\mathbf{f}}_{\mathbf{X}_i}$. This integration ensures that the generated imagery accurately reflects the specified conditions and contextual details. **1)** To fuse visual and textual information from $\mathbf{X}_i$ and G_i at a deep level, we employ BLIP [30], a model designed specifically for multimodal data fusion. BLIP uses a Vision Transformer (ViT) [31] to process the image $\mathbf{X}_i$ and encodes the text description G_i using BERT [32]. The latent representations of both modalities are combined through a cross-attention mechanism, forming a unified representation $\mathbf{C}_{xg} = \text{BLIP}(\mathbf{X}_i, G_i)$. **2)** Additionally, let G'_i represent the text description for the generated image. To capture the nuanced textual semantics that guide image generation transitions, we leverage CLIP [33] to encode the text description G'_i into $\mathbf{C}_g = \text{CLIP}(G'_i)$. This step enhances the model's sensitivity to the text-driven requirements for the generated image, as well as to the textual description of the source image. Note that G_i corresponds to the text description of $\mathbf{X}_i$, while G'_i represents the text description we aim to generate alongside the image. The final conditional vector $\mathbf{C}$ is formed by concatenating these representations:

$$\mathbf{C} = [\mathbf{C}_{xg}; \mathbf{C}_g; \hat{\mathbf{f}}_{\mathbf{X}_i}] \qquad (5)$$

3) Denoising Network. To reconstruct desired data samples from noise, we utilize a conditional UNet denoising network[34], denoted as ϵ_θ. This network is tasked with reversing the forward diffusion process. Specifically, at each time step t, the network $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{C})$ learns to transform the noisy latent representation $\mathbf{z}_t$ into a cleaner version $\mathbf{z}_{t-1}$. The condition vector $\mathbf{C}$ provides additional contextual information that is integrated into the denoising process, enhancing the model's ability to capture complex patterns and dependencies for more effective noise removal. To facilitate the learning process, we integrate the condition vector $\mathbf{C}$ directly into each hidden layer of the autoencoder by concatenation. This approach modulates the feature extraction and refinement stages, enhancing the capacity of $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{C})$ to generate high-quality outputs. Building upon the aforementioned steps, we train AeroDiffusion by minimizing the following loss function:

$$\mathcal{L}_{\text{AeroDiffusion}} = \mathbb{E}_{\mathbf{z}_0, \epsilon\sim\mathcal{N}(0,\mathbf{I}), t, \mathbf{C}}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{C})\|_2^2\right] \qquad (6)$$

In this optimization process, both the parameters θ of the denoising network and those involved in generating the condition vector $\mathbf{C}$ are jointly updated.

V. Evaluation

We evaluate the performance of **AeroDiffusion** on the VisDrone-DET aerial image dataset [14], comparing it to other five state-of-the-art (SOTA) diffusion models.

A. Experimental Settings

Aerial Image Dataset. The *VisDrone-DET* [14] contains 10,209 images captured by drone-mounted cameras across 14 cities, covering various environments and conditions. The dataset includes over 2.6 million manually annotated bounding boxes for objects like pedestrians, cars, and bicycles.

Parameters for Text Generation. The key-point text generation, powered by LLMs, generates descriptions for the *VisDrone-DET* dataset [14], detailing scene composition, camera viewpoints, object orientations, and spatial arrangements. The temperature is set to 1.2 for balanced randomness, and the maximum token length is limited to 120 to ensure concise and relevant descriptions.

Parameters for Aerial Image Generation. We downsample the source aerial image to $512 \times 512 \times 3$ before compressing it into the latent space using our trained auto-encoder. AeroDiffusion is trained for 50 epochs on 6,471 images from the VisDrone-DET dataset. For evaluation, we randomly sampled 3,200 images from the test set to generate new images.

TABLE I: Performance Comparison of SOTA Models for Aerial Image Synthesis. Results are based on 3,200 generated sample images. **Bold** values represent the best performance, while underlined values indicate the second-best.

Models	**FID ↓**	**PSNR ↑**	**KID ↓**
DDPM	217.95	**10.38**	0.18
Stable Diffusion	119.13	4.85	0.07
ARLDM	111.59	5.61	**0.04**
Versatile Diffusion	124.12	5.70	0.06
Make-a-scene	114.74	5.74	0.06
Average	137.51	6.46	0.08
AeroDiffusion (ours)	**78.15**	5.98	**0.04**

Training utilized the Adam optimizer [35] with a learning rate of 1×10^{-5} and weight decay of 10^{-5}. The DDPM scheduler [36] added Gaussian noise over 1000 time steps, with a β range of 0.001 to 0.012, while DDIM [37] denoised in 250 inference steps with a guidance scale of 7.0. Training was conducted on eight Nvidia A100 GPUs, with evaluations on a single Nvidia A100 GPU.

Baselines. To thoroughly evaluate AeroDiffusion's performance, we benchmark it against five SOTA models, consisting of one probabilistic model, *DDPM*[38], and four conditional diffusion models: *Stable Diffusion*[24], *ARLDM*[20], *Versatile Diffusion*[39], and *Make-a-Scene* [40].

Evaluation Metrics. We utilize established evaluation metrics for image synthesis, including FID (Fréchet Inception Distance) [41], PSNR (Peak Signal-to-Noise Ratio), and KID (Kernel Inception Distance). FID and KID assess image quality and diversity, with KID offering an unbiased estimate of distribution similarity. PSNR evaluates reconstruction fidelity by measuring the ratio between signal and noise. Additionally, we measure the CLIP score [33] to assess how well the generated text aligns with the visual content, using a pre-trained CLIP model. Optimal metric directions are indicated in our evaluation tables.

B. Quantitative Assessment: Performance Overview

Aerial Image Synthesis. We evaluate AeroDiffusion using FID, PSNR, and KID. As shown in Table I, AeroDiffusion achieves a state-of-the-art FID of 78.15 and a KID of 0.04, reducing FID by 43.2% and KID by 50.0% compared to the average, outperforming models like Stable Diffusion and Versatile Diffusion. AeroDiffusion records a PSNR of 5.98, while DDPM achieves a higher PSNR of 10.38 due to its operation in pixel space, which retains finer details. However, as shown in Fig. 4, a higher PSNR does not always correspond to better perceptual quality, as DDPM struggles with object generation despite its higher PSNR. Overall, AeroDiffusion balances all metrics, delivering superior image quality.

Keypoint-Aware Text Generation. In Table II, we evaluate the performance of keypoint-aware text generation for aerial image synthesis across various LLMs [16]. AeroDiffusion outperforms the others in both CLIP Score and FID, achieving a CLIP Score of 32.82, surpassing the second-highest score of 30.12 from Gemini [42]. In terms of FID, it measures the overall quality of the generated images based on different text descriptions from the baseline LLMs. The combination of a higher CLIP Score and superior FID suggests a substantial improvement in our model's ability to generate text that is more effectively aligned with the keypoints and visual features of the provided aerial images.

TABLE II: Evaluation for Keypoint-Aware Text Generation.

LLM	**CLIP SCORE ↑**	**FID ↓**
Gemini	30.12	86.22
GPT-4o	29.22	92.11
BLIP	25.64	126.38
AeroDiffusion	**32.82**	**78.16**

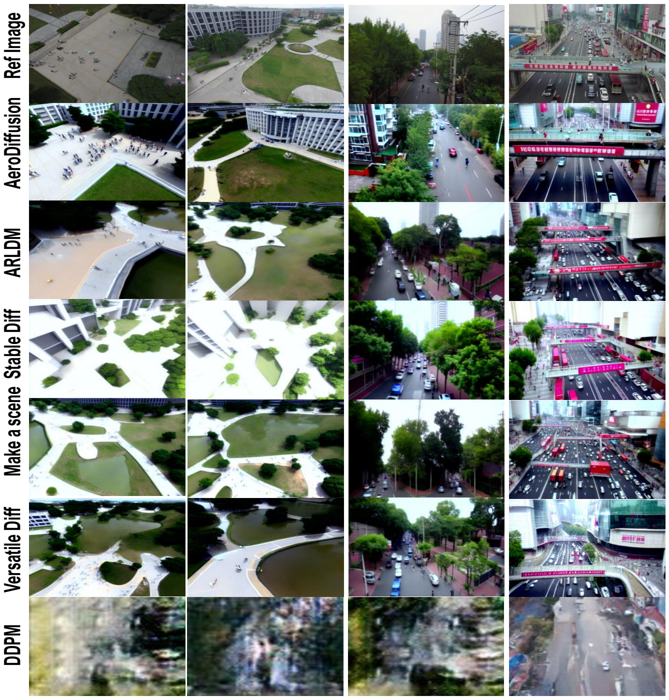


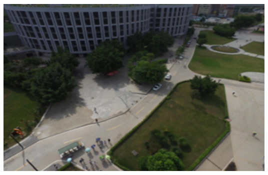
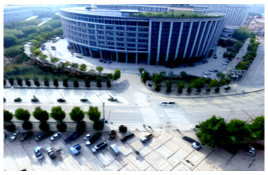
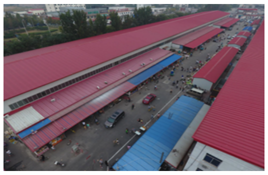
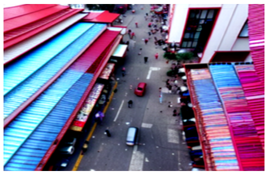
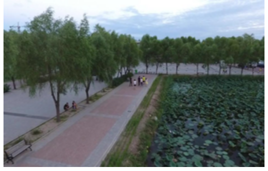
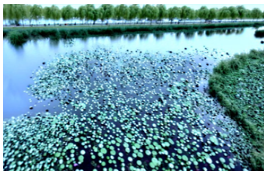
Fig. 4: Generated sample **daytime** images on VisDrone-DET.

C. Qualitative Assessment: Generated Image Visualization

To further validate the generation quality, we present visualizations of specific examples of generated aerial images, categorized into three types. First, we showcase four images generated by our model in comparison with five baseline models. Next, we display images generated from different viewpoint transitions of a reference image, followed by three nighttime examples. Due to space constraints, we have omitted visualizations of the baseline model outputs for the viewpoint transition and nighttime examples, as our model's superior performance has already been demonstrated through the quantitative results in the tables and other figures.

Complex Aerial Image Synthesis. Fig. 4 visualizes four samples alongside their original/reference images. Notably, AeroDiffusion generates highly realistic images, nearly identical to the original scenes, while the benchmark models only approximate the shapes of roads and buildings. The DDPM model, despite achieving the highest PSNR in Table I, fails to detect object features, highlighting that PSNR alone does not indicate the best performance. For the third example, all models generate a single road with traffic, but only AeroDiffu-

TABLE III: Viewpoint Transition Image Synthesis

Ref Image Description	Ref Image	Generated Image	G'_i: Requirement
G_i: A daytime aerial image of a paved campus with trees, grassy areas, a walkway, and several people walking around **captured from a low angle to the side**. There are a few cars parked on the side of the road.			A daytime aerial image showcasing a campus scene captured from **a high front angle above** a paved intersection where two roads meet. The drone's perspective allows **for a bird's eye view** of the scene, providing a sense of depth and revealing the layout of the campus.
G_i: A daytime aerial image captures a bustling market scene **from a drone hovering high above**. The sky is a muted blue, and the long shadows cast by the red-roofed buildings $\cdots$			A daytime aerial image captures a bustling market scene from a drone **hovering at a lower altitude, positioned slightly above and looking directly at the center of a narrow street**. The sky is a muted blue, and the long $\cdots$
G_i: A daytime aerial image taken under a slightly cloudy sky, showcasing a park scene. The drone is positioned **high above the park, looking down at a slightly angled perspective**. The image captures a paved walkway, lined with trees $\cdots$			A daytime aerial image of a tranquil park scene, with **the drone positioned high above the pond, offering a slightly angled view of the landscape**. The image highlights a paved walkway lined with evenly spaced trees $\cdots$

sion accurately captures the single skycross road in the fourth example, whereas others mistakenly generate three. Additionally, AeroDiffusion excels in generating small objects, such as people in the plaza, typically discarded during subsampling. Overall, the visualizations showcase AeroDiffusion's strong controllability and consistency, even in handling small objects.

Viewpoint Transition Image Synthesis. As illustrated in Table III, AeroDiffusion showcases its ability to diversify complex scene generation by producing images from new viewpoints, leveraging different CLIP texture information G'_i. This feature is crucial for applications that require a comprehensive understanding of scenes from multiple angles. By adjusting CLIP descriptions to new perspectives, AeroDiffusion not only generates closer viewpoints of the original image but also predicts extended perspectives (second and third examples). The third example, in particular, highlights an extended perspective, offering a new side view of the pond while capturing people walking along the sidewalk. This capability enhances the model's applicability in real-world scenarios, demonstrating its sophisticated grasp of spatial dynamics and scene structure.

Nighttime Image Synthesis. In addition to viewpoint adaptations, we incorporated detailed textual descriptions of lighting conditions to further enhance the quality of nighttime image generation. This refinement allowed AeroDiffusion to accurately render light shadows cast by objects, such as cars, closely mimicking real-world light projection relative to object positioning. As shown in Fig. 5, AeroDiffusion successfully generates highly realistic and contextually coherent nighttime aerial imagery (High-noise condition).

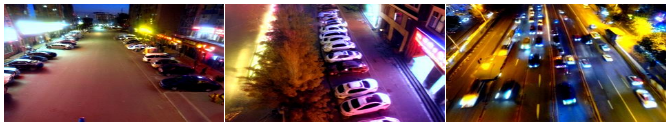


Fig. 5: Generated samples at **nighttime**.

TABLE IV: Ablation study: OD denotes object detection for feature augmentation.

Our LLMs	OD	BLIP	FID ↓	PSNR ↑	KID ↓
			132.60	4.80	0.09
		✓	119.13	4.85	0.07
✓		✓	108.23	4.92	0.05
✓	✓	✓	78.15	5.98	0.04

D. Ablation Study

We conducted ablation studies to evaluate the impact of our proposed components, including keypoint-aware text generation and region-level feature augmentation. Starting with a fine-tuned Stable Diffusion model, we first integrated BLIP to validate deep text-visual feature fusion, then replaced the generated text with descriptions from Gemeni and our own keypoint-aware text. By leveraging ROIs, we used our object retrieval component to isolate critical regions, enabling the model to zoom in and improve the representation of smaller objects by focusing on high-importance areas. This approach led to a significant improvement of **54.45** in the FID compared to the base Stable Diffusion. The improvements are evident when evaluating individual component, as shown in Table IV.

VI. Conclusion

In conclusion, this paper presents **AeroDiffusion**, a novel framework for generating high-resolution, text-guided aerial images with enhanced controllability. By integrating keypoint-aware text generation and region-level feature augmentation, our approach addresses the challenges of complex scene synthesis, particularly in generating images from multiple viewpoints. Through extensive evaluations, including ablation studies, AeroDiffusion demonstrates significant improvements in image quality, outperforming SOTA models across multiple benchmarks.

References

[1] Y. Chang, Y. Cheng, U. Manzoor, and J. Murray, "A review of uav autonomous navigation in gps-denied environments," *Robotics and Autonomous Systems*, p. 104533, 2023.

[2] W. Wei, C. Pan, S. Islam, J. Banerjee, S. Palanisamy, and M. Xie, "Intermittent ota code update framework for tiny energy harvesting devices," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024.

[3] W. Zhang, M. Xie, C. Scott, and C. Pan, "Sparsity-aware intelligent spatiotemporal data sensing for energy harvesting iot system," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 11, pp. 4492–4503, 2022.

[4] W. Zhang, C. Pan, T. Liu, J. Zhang, M. Sookhak, and M. Xie, "Intelligent networking for energy harvesting powered iot systems," *ACM Transactions on Sensor Networks*, vol. 20, no. 2, pp. 1–31, 2024.

[5] W. Zhang, W. Wang, M. Sookhak, and C. Pan, "Joint-optimization of node placement and uav's trajectory for self-sustaining air-ground iot system," in *2022 23rd International Symposium on Quality Electronic Design (ISQED)*. IEEE, 2022, pp. 1–6.

[6] G. Rohi, G. Ofualagba *et al.*, "Autonomous monitoring, analysis, and countering of air pollution using environmental drones," *Heliyon*, vol. 6, no. 1, 2020.

[7] W. Zhang, L. Li, N. Zhang, T. Han, and S. Wang, "Air-ground integrated mobile edge networks: A survey," *IEEE Access*, vol. 8, pp. 125 998–126 018, 2020.

[8] W. Wei, S. Islam, J. Banerjee, S. Zhou, C. Pan, C. Ding, and M. Xie, "An intermittent ota approach to update the dl weights on energy harvesting devices," in *2022 23rd International Symposium on Quality Electronic Design (ISQED)*. IEEE, 2022, pp. 1–6.

[9] J. Banerjee, S. Islam, W. Wei, C. Pan, and M. Xie, "Autotile: Autonomous task-tiling for deep inference on battery-less embedded system," in *Proceedings of the Great Lakes Symposium on VLSI 2024*, 2024, pp. 323–327.

[10] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[11] J. Pan, G. Chen, Y. Liu, J. Wang, C. Bian, P. Zhu, and Z. Zhang, "Tell me the evidence? dual visual-linguistic interaction for answer grounding," *arXiv preprint arXiv:2207.05703*, 2022.

[12] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[13] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "Ediffsr: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[14] Y. Cao, Z. He, L. Wang, W. Wang, Y. Yuan, D. Zhang, J. Zhang, P. Zhu, L. Van Gool, J. Han *et al.*, "Visdrone-det2021: The vision meets drone object detection challenge results," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 2847–2854.

[15] A. Maharana and M. Bansal, "Integrating visuospatial, linguistic and commonsense structure into story visualization," *arXiv preprint arXiv:2110.10834*, 2021.

[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[17] W.-D. K. Ma, A. Lahiri, J. P. Lewis, T. Leung, and W. B. Kleijn, "Directed diffusion: Direct control of object placement through attention guidance," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4098–4106.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[20] X. Pan, P. Qin, Y. Li, H. Xue, and W. Chen, "Synthesizing coherent story with auto-regressive latent diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 2920–2930.

[21] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[22] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[23] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[27] A. Toker, M. Eisenberger, D. Cremers, and L. Leal-Taixé, "Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation," *arXiv preprint arXiv:2403.16605*, 2024.

[28] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[29] J. Wan, B. Zhang, Y. Zhao, Y. Du, and Z. Tong, "Vistrongerdet: Stronger visual information for object detection in visdrone images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2820–2829.

[30] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[31] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[32] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[35] P. Zhou, X. Xie, Z. Lin, and S. Yan, "Towards understanding convergence and generalization of adamw," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[37] T. Garber and T. Tirer, "Image restoration by denoising diffusion models with iteratively preconditioned guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 245–25 254.

[38] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[39] X. Xu, Z. Wang, G. Zhang, K. Wang, and H. Shi, "Versatile diffusion: Text, images and variations all in one diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7754–7765.

[40] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman, "Make-a-scene: Scene-based text-to-image generation with human priors," in *European Conference on Computer Vision*. Springer, 2022, pp. 89–106.

[41] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[42] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.