GenHaze: Pioneering Controllable One-Step Realistic Haze Generation for Real-World Dehazing

Supplementary Material

This is supplementary material for ***GenHaze: Pioneering Controllable One-Step Realistic Haze Generation for Real-World Dehazing.***

We present the following materials in this supplementary material:

1. Constructing the Haze Database

To build a comprehensive haze database, we leverage the Multi-Modal Large Language Model (MLLM) to generate high-quality captions that describe haze-related characteristics in real-world images. The process focuses on the description of haze-specific attributes, such as thickness, color, light diffusion, texture, distance effects, contrast reduction, ambient mood, depth gradient, and atmospheric layering, while excluding unrelated scene elements. As illustrated in Fig.1, iterative refinement of haze description prompts is employed to optimize MLLM's performance. By using in-context learning techniques, we guide the model to identify relevant haze attributes and produce concise, precise captions limited to 50 words. For instance, the haze is described as dense and yellowish, diffusing light heavily to create a muted tone, with reduced contrast and softened edges. These captions reflect key haze properties like texture smoothness, atmospheric layering, and pronounced depth gradients. This approach ensures the generation of a robust database with accurately labeled haze descriptions, which can be further utilized to replace handcrafted text for better robustness and coverage.

Figure 1. **The details of constructing the haze database.** Iteratively refining the haze description task to enable the MLLM [14] to focus on specific haze attributes, leveraging similar in-context learning to guide the model in identifying relevant attributes and generating precise prompts.

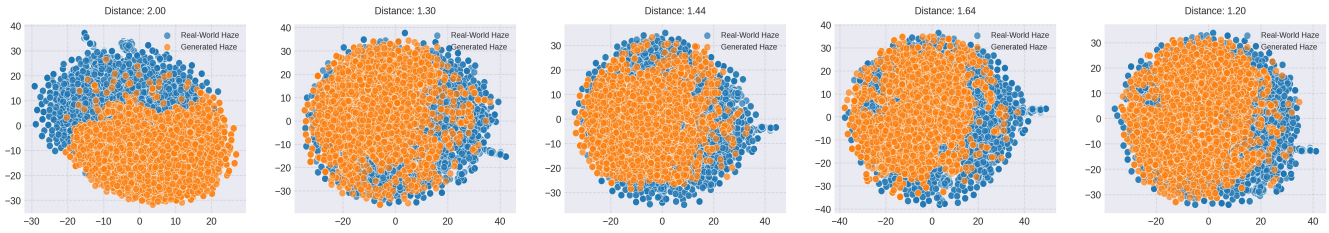
2. Theory for GenHaze and Better Fine-tuning

2.1. Why Do We Need GenHaze

Traditional methods for generating hazy images typically rely on the atmospheric scattering model [6, 11, 28]:

$$\mathcal{I}(x) = \mathcal{J}(x) \cdot t(x) + \mathcal{A} \cdot \big(1 - t(x)\big), \tag{1}$$

where $\mathcal{I}(x)$ is the hazy image, $\mathcal{J}(x)$ is the clear image, and $t(x)$ denotes the transmission. In these formulations, parameters such as the scattering coefficient and atmospheric light are often assumed to be spatially invariant. This static assumption fails to capture the spatially varying and dynamic characteristics of real-world haze—like local changes in density, color, and light scattering—which leads to a synthetic hazy dataset, $\mathcal{D}_p$, that diverges from the true distribution of real hazy images, $\mathcal{D}_t$. This discrepancy can



(a) UWNR (b) Low-Res* (c) RIDCP* (d) CycleGAN-Turbo (e) GenHaze

Figure 2. **tSNE visualizations of haze images generated by different methods and real-world haze images.** The haze generated by GenHaze is more controllable and closer to the real-world haze than other methods [3, 22, 28, 29].

be expressed as:

$$d_D(\mathcal{D}_p, \mathcal{D}_t) = \int \left| p_{\mathcal{D}_p}(I) - p_{\mathcal{D}_t}(I) \right| dI. \tag{2}$$

A large $d_D(\mathcal{D}_p, \mathcal{D}_t)$ implies that models trained on $\mathcal{D}_p$ may struggle to generalize to real hazy images. Domain adaptation theory [1] formalizes this by bounding the risk (error) of a model f on the real distribution $\mathcal{D}_t$:

$$\mathcal{R}_{\mathcal{D}_t}(f) \leq \mathcal{R}_{\mathcal{D}_p}(f) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_p, \mathcal{D}_t) + \lambda, \tag{3}$$

where $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_p, \mathcal{D}_t)$ quantifies the divergence between the synthetic and real domains, and λ is a constant representing ideal conditions. A high divergence directly translates into a higher risk bound for dehazing models when applied to real-world data.

While some methods try to adapt the scattering model by incorporating spatially varying parameters, they often encounter challenges with complicated pipeline and still fail to capture the full nuances of natural haze.

GenHaze takes a different route by leveraging a diffusion model conditioned on real haze features. Instead of relying solely on fixed, physically derived parameters, GenHaze extracts haze characteristics from real-world images and integrates these cues into the generation process. The goal is to produce a generated distribution $\mathcal{D}_g$ that is closer to the true distribution $\mathcal{D}_t$, such that:

$$d_D(\mathcal{D}_g, \mathcal{D}_t) < d_D(\mathcal{D}_p, \mathcal{D}_t), \tag{4}$$

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_g, \mathcal{D}_t) < d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_p, \mathcal{D}_t). \tag{5}$$

This reduction in divergence implies that models trained on $\mathcal{D}_g$ will have a lower risk when applied to real-world data, as indicated by the domain adaptation risk bound. Moreover, by transforming the training object, GenHaze bypasses the inefficiencies of multi-step diffusion, achieving realistic haze generation in a single step.

2.2. Leveraging Diffusion Models for Haze Generation

Unlike conventional diffusion models that iteratively denoise over T steps, GenHaze employs a one-step generation strategy. Traditional diffusion models follow a forward process adding noise to a clean image $\mathbf{x}_0$:

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\epsilon, \tag{6}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and a reverse process gradually refines $\mathbf{x}_T$ back to $\mathbf{x}_0$. GenHaze, however, repurposes this paradigm for a "clean-to-haze" transformation. It fixes the timestep at $t = T$, treating the encoded clean image $\mathcal{E}(\mathbf{x})$ as the starting point:

$$\mathbf{z}_T = \mathcal{E}(\mathbf{x}), \tag{7}$$

and trains the model to predict the hazy image distribution directly in one forward pass:

$$p(\mathbf{x}_0|\mathcal{E}(\mathbf{x}), \mathcal{F}^d, e^{\text{text}}) = \mathcal{N}(\mu_\theta(\mathcal{E}(\mathbf{x}), \mathcal{F}^d, e^{\text{text}}), \sigma_\theta^2\mathbf{I}), \tag{8}$$

where $\mathbf{x}_0$ is the hazy image. Here, μ_θ and σ_θ are conditioned on $\mathcal{E}(\mathbf{x})$, $\mathcal{F}^d$, and e^{text}, enabling a direct transformation from clean to hazy without iterative sampling. This is theoretically grounded in the observation that at $t = T$ [18], the diffusion process typically yields near-pure noise, but by substituting $\mathcal{E}(\mathbf{x})$ (with a mean of zero, the same with Gaussian noise), the model learns a single-step mapping to the target distribution. The training loss reinforces this:

$$\mathcal{L}_{\text{1-step}} = \mathbb{E}_{\mathbf{x}} \left[\| \mathcal{D}(\mu_\theta(\mathcal{E}(\mathbf{x}), \mathcal{F}^d, e^{\text{text}}, T)) - \mathbf{x}_{\text{hazy}} \|_1 \right], \tag{9}$$

This eliminates multi-step overhead while preserving realism through conditioning on $\mathcal{F}^d$ and e^{text} [24], ensures that the model learns to predict realistic haze degradation in one inference step. It also reduces the iterative denoising of traditional diffusion (reducing the complexity from $\mathcal{O}(T)$ to $\mathcal{O}(1)$), making it computationally efficient while preserving the generative capacity of the model.

2.3. Theoretical Analysis of Genhaze Advantage

Traditional synthetic datasets, denoted as $\mathcal{D}_p$, differ significantly from real-world hazy datasets, $\mathcal{D}_t$. This gap is captured by the total variation distance:

$$d_D(\mathcal{D}_p, \mathcal{D}_t) = \int |p_{\mathcal{D}_p}(I) - p_{\mathcal{D}_t}(I)| \, dI, \tag{10}$$

where $p_{\mathcal{D}_p}(I)$ and $p_{\mathcal{D}_t}(I)$ represent the probability density functions of the synthetic and real hazy datasets, respectively. This distributional divergence has practical implications, elevating the risk bound in domain adaptation theory [1]:

$$\mathcal{R}_{\mathcal{D}_t}(f) \leq \mathcal{R}_{\mathcal{D}_p}(f) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_p, \mathcal{D}_t) + \lambda, \tag{11}$$

where $\mathcal{R}_{\mathcal{D}_t}(f)$ is the error of a dehazing model f on real data, $\mathcal{R}_{\mathcal{D}_p}(f)$ is the error on synthetic data, $d_{\mathcal{H}\Delta\mathcal{H}}$ quantifies domain discrepancy, and λ is a constant tied to ideal conditions. A large $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_p, \mathcal{D}_t)$ increases $\mathcal{R}_{\mathcal{D}_t}(f)$, hindering generalization to real-world haze.

GenHaze mitigates this by conditioning its generation on real haze features $\mathcal{F}^d$ and refined text embeddings e^{text}, extracted via a degradation encoder and a caption-based embedding process from actual hazy images, yielding a conditional distribution:

$$p(\mathbf{x}_0|\mathcal{E}(\mathbf{x}), \mathcal{F}^d, e^{\text{text}}) = \mathcal{N}(\mu_\theta(\mathcal{E}(\mathbf{x}), \mathcal{F}^d, e^{\text{text}}), \sigma_\theta^2 \mathbf{I}), \tag{12}$$

where $\mathcal{E}(\mathbf{x})$ is the encoded clean image in the latent space, e^{text} is a refined text embedding, and μ_θ and σ_θ are model-predicted parameters. This conditioning aligns the generated distribution $p_{\mathcal{D}_g}(I)$ with $p_{\mathcal{D}_t}(I)$, achieving:

$$d_D(\mathcal{D}_g, \mathcal{D}_t) < d_D(\mathcal{D}_p, \mathcal{D}_t). \tag{13}$$

The alignment stems from $\mathcal{F}^d$ encoding instance-specific haze attributes (e.g., density, color shifts) and the text embedding e^{text} capturing haze descriptions. These are embedded into the diffusion process via a haze modulation mechanism:

$$\hat{w}_i = s_i \cdot w_i, \quad \hat{w}_i' = \frac{\hat{w}_i}{\sqrt{\sum_i (\hat{w}_i)^2 + \epsilon}}, \tag{14}$$

where s_i, derived from $\mathcal{F}^d$, adjusts convolutional weights w_i to reflect real haze characteristics, and ϵ ensures stability. This reduces variance in $\mathcal{D}_g$, aligning it with $\mathcal{D}_t$ and lowering $\mathcal{R}_{\mathcal{D}_t}(f)$, thus enhancing dehazing performance [12].

GenHaze overcomes two key limitations of traditional synthetic datasets by jointly leveraging instance-specific haze embeddings $\mathcal{F}^d$ and text embeddings e^{text}:

- **High Distributional Divergence and Embedding Alignment**: By modeling $p(\mathbf{x}_0|\mathcal{F}^d, e^{\text{text}})$, GenHaze captures the nuanced complexity of real haze. It does so not only by encoding the physical attributes through $\mathcal{F}^d$ but also by integrating haze descriptions through e^{text}. This reference-controllable dual conditioning surpasses physics-based models that rely on simplified assumptions. The reduced $d_D(\mathcal{D}_g, \mathcal{D}_t)$ minimizes domain adaptation risk as described by the adaptation bound, making $\mathcal{D}_g$ a robust training proxy for real-world dehazing [1].
- **Optimization of the Text Embedding and Joint Adaptation**: The text embedding e^{text} enhances adaptation by integrating general haze descriptions with the instance-specific haze attributes. GenHaze initializes a general embedding e^{general} as the average of CLIP-encoded captions from real haze images:

$$e^{\text{general}} = \frac{1}{N} \sum_{i=1}^{N} f^{\text{emb}}(f^{\text{cap}}(\mathcal{I}_i)), \tag{15}$$

where f^{cap} generates captions and f^{emb} maps them to embeddings. This provides a prior capturing broad haze attributes. To adapt to specific instances, GenHaze introduces a learnable prompt offset Δe:

$$e^{\text{text}} = e^{\text{general}} + \Delta e, \tag{16}$$

where Δe is optimized during training. The conditional distribution becomes:

$$\begin{aligned} &p(\mathbf{x}_0|\mathcal{E}(\mathbf{x}), \mathcal{F}^d, e^{\text{general}} + \Delta e) = \\ &\mathcal{N}\big(\mu_\theta(\mathcal{E}(\mathbf{x}), \mathcal{F}^d, e^{\text{general}} + \Delta e), \quad \sigma_\theta^2 \mathbf{I}\big). \end{aligned} \tag{17}$$

Theoretically, Δe refines e^{general} by minimizing the divergence between the generated and real hazy distributions. During training, the model updates θ (including μ_θ and σ_θ) and Δe to optimize:

$$\mathcal{L}_{\text{1-step}} = \mathbb{E}_{\mathbf{x}} \left[\|\mathcal{D}(\mu_\theta(\mathcal{E}(\mathbf{x}), \mathcal{F}^d, e^{\text{general}} + \Delta e, T)) - \mathbf{x}_{\text{hazy}}\|_1 \right]. \tag{18}$$

Since e^{text} conditions the generation via mechanisms like cross-attention, optimizing Δe adjusts the embedding to emphasize haze-specific semantics (e.g., thickness, opacity), further reducing distributional mismatch. This joint adaptation through both general and specific embedding aligns the generated $\mathcal{D}_g$ with $\mathcal{D}_t$, thereby effectively reducing the domain discrepancy $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_p, \mathcal{D}_t)$ and consequently lowering $\mathcal{R}_{\mathcal{D}_t}(f)$ as predicted by adaptation theory.

Probabilistically, conditioning on $\mathcal{F}^d$ and e^{text} minimizes:

$$\int \left| p_{\mathcal{D}_g}(I) - p_{\mathcal{D}_t}(I) \right| dI, \tag{19}$$

constraining $\mathcal{D}_g$ to real haze manifolds. The one-step approach and optimized e^{text} together enable GenHaze to efficiently transform clean images into hazy ones, minimizing $d_D(\mathcal{D}_g, \mathcal{D}_t)$, reducing $\mathcal{R}_{\mathcal{D}_t}(f)$, and improving dehazing generalization. By addressing distributional divergence and computational inefficiency, GenHaze provides a theoretically robust, practically efficient solution for realistic haze generation.

2.4. Validation with t-SNE and Distance Metrics

To validate that the generated distribution $\mathcal{D}_g$ of GenHaze aligns closely with the target real-world haze distribution $\mathcal{D}_t$, we employ both t-SNE visualizations and distance metrics, as shown in Fig.2. These methods offer qualitative and quantitative evidence that GenHaze achieves superior alignment compared to traditional approaches.

t-SNE is used to reduce the high-dimensional feature space of haze images into a 2D representation, which facilitates a visual comparison of the distributions. The t-SNE algorithm minimizes the Kullback-Leibler divergence between the joint probabilities of the high-dimensional data and those of the low-dimensional embedding:

$$\mathcal{L}_{\text{t-SNE}} = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{20}$$

where p_{ij} denotes the similarity between data points i and j in the original space (computed from pairwise Euclidean distances), and q_{ij} represents the similarity in the 2D embedding space, modeled using a t-distribution to mitigate the crowding problem. This approach preserves local structure, allowing us to assess how closely the generated haze images ($\mathcal{D}_g$) cluster with real haze images ($\mathcal{D}_t$) compared to methods such as RIDCP [28], Low-Res [3], and CycleGAN-Turbo [22]. As shown in Fig.2, GenHaze produces tighter clusters with the real-world haze samples, indicating a closer alignment.

Distance Metric. To quantitatively assess the alignment between the generated distribution $\mathcal{D}_g$ and the real-world haze distribution $\mathcal{D}_t$, we compute the average nearest-neighbor distance in the t-SNE embedding space. This metric measures the proximity of each generated sample to its closest real-world counterpart:

$$d_{\text{t-SNE}} = \frac{1}{N} \sum_{i=1}^{N} \min_j \|\mathbf{z}_i^{(\mathcal{D}_g)} - \mathbf{z}_j^{(\mathcal{D}_t)}\|_2, \tag{21}$$

where $\mathbf{z}_i^{(\mathcal{D}_g)}$ and $\mathbf{z}_j^{(\mathcal{D}_t)}$ denote the 2D t-SNE embeddings of the generated and real haze images, respectively, and N is the number of generated samples. A smaller $d_{\text{t-SNE}}$ indicates a closer local alignment between $\mathcal{D}_g$ and $\mathcal{D}_t$, which is also suggestive of a reduced global divergence (i.e. $d_D(\mathcal{D}_g, \mathcal{D}_t) = \int \left| p_{\mathcal{D}_g}(I) - p_{\mathcal{D}_t}(I) \right| dI$,) and aligns with our theoretical formulation.

The results shown in Fig.2 indicate that GenHaze achieves a $d_{\text{t-SNE}}$ of 1.20, outperforming RIDCP (1.44) and Low-Res (1.30). This smaller distance demonstrates that $\mathcal{D}_g$ is more closely aligned with $\mathcal{D}_t$ in the embedding space, underscoring GenHaze's capability to generate haze images that more accurately reflect real-world characteristics. Moreover, the tighter clustering and reduced $d_{\text{t-SNE}}$ confirm the effectiveness of the one-step generation strategy and the clean-to-haze protocol in producing realistic, controllable haze—capturing the subtle nuances of natural haze distributions better than competing methods.

3. More Details of Dehazing Fune-tuning

For the fine-tuning of dehazing baselines, we select all the dehazing baselines that are pre-trained on OTS [17] for fair comparison, T^3-DiffWeather, PSD, and KANet are chosen with their pre-trained weights for real-world scenes. Tab.1 summarizes the detailed configuration for each baseline. The initial learning rate (LR) is set for each model and is significantly smaller than the rates used for training the networks from scratch. This is because our paradigm is designed to unlock the potential of already-trained models in a straightforward manner, without requiring large adjustments that would necessitate a high learning rate. Most models use the Adam optimizer [8] with a weight decay of 0.0000, while Dehazeformer and TaylorFormer utilize the AdamW optimizer [19] for better regularization.

Fine-tuning is conducted over 1-2 epochs because our approach eliminates the need for extensive training on large datasets. GenHaze generates realistic and controllable haze images, allowing baselines to adapt efficiently with minimal training. In addition, no learning rate schedule is applied to demonstrate the robustness of our paradigm for each model. Notably, we retain the original loss functions specified in the respective papers and do not modify any model architectures. This highlights the effectiveness of our strategy in enhancing existing baselines without requiring architectural changes. All experiments are conducted using the PyTorch framework on A800 GPUs to ensure reproducibility and computational efficiency.

Table 1. Detailed fine-tuning configurations of dehazing baselines.

Model	Initial LR	Optimizer	Weight Decay	Epochs	LR Schedule
MSBDN [9]	0.0001	Adam	0.0000	1	None
FocalNet [7]	0.00006	Adam	0.0000	2	None
DeHamer [13]	0.0001	Adam	0.0000	2	None
Dehazeformer [25]	0.00012	AdamW	0.01	2	None
TaylorFormer [23]	0.0001	AdamW	0.01	2	None
T3-DiffWeather [4]	0.00007	Adam	0.0000	1	None
PSD [6]	0.0001	Adam	0.0000	1	None
KANet [11]	0.000015	Adam	0.0000	1	None

4. Evaluation Metrics

Unlike previous approaches [11, 28], we evaluate the effectiveness of GenHaze in enhancing dehazing performance from multiple perspectives. The improvements across various comprehensive non-reference metrics demonstrate that GenHaze can unlock the potential of existing baselines through a straightforward yet effective approach. Specifically, we adopt a set of non-reference metrics tailored for various aspects of image quality assessment, as detailed in Tab. 2. These metrics provide a robust framework for analyzing the effectiveness of our methods across aesthetic, perceptual, and structural dimensions:

- **FADE [10]**: This metric evaluates fog and haze quality through statistical analysis of foggy regions and density-based measures, making it ideal for haze-specific assessments.

- **BRISQUE [20]**: Focused on general distortions, BRISQUE analyzes scene statistics and locally normalized luminance to assess image quality.

- **NIMA [26]**: Based on a CNN-based architecture, NIMA predicts aesthetic scores, providing insights into the subjective quality of images.

- **NIQE [21]**: A training-free metric that uses statistics from the natural scene to evaluate the overall naturalness of an image.

- **PIQE [27]**: Conducting block-wise analysis and spatial feature extraction, PIQE effectively detects and measures local distortions.

- **MUSIQ [15]**: Leveraging multi-scale processing and transformer-based architectures, MUSIQ assesses perceptual quality, particularly for complex image features.

- **PaQ-2-PiQ [30]**: This deep learning-based metric combines patch-level and global quality analyses, offering a comprehensive perspective on image fidelity and quality.

These metrics are carefully selected to cover diverse evaluation needs, ensuring a thorough and objective comparison of performance across multiple datasets and tasks. By incorporating both specialized and general-purpose metrics, we provide a holistic assessment of image quality in dehazing scenarios.

5. Comparison of tSNE Visualization

Fig.2 presents tSNE visualizations comparing haze images generated by different methods with real-world haze images, where orange represents real-world haze and blue represents generated haze. The average cluster distances reflect the alignment of generated haze with real-world distributions. UWNR [29] and CycleGAN-Turbo [22] exhibit the largest distances (2.00 and 1.64, respectively), indicating significant deviations from real-world haze characteristics. RIDCP and Low-Res show moderate improvements, with distances of 1.44 and 1.30, capturing some haze properties but still lacking precision. GenHaze, however, achieves the smallest distance (1.20), demonstrating superior alignment with real-world haze. This tighter clustering highlights GenHaze's ability to generate highly realistic and controllable haze images that closely resemble natural haze, surpassing other methods in capturing real-world haze characteristics.

Table 2. Overview of Non-Reference Image Quality Assessment Metrics.

Metric	Key Features	Application Focus
FADE [10]	• Statistical analysis of foggy regions • Density-based evaluation	Fog/haze assessment
BRISQUE [20]	• Scene statistics analysis • Locally normalized luminance	General distortions
NIMA [26]	• CNN-based architecture • Aesthetic score prediction	Aesthetic quality
NIQE [21]	• Natural scene statistics • Training-free approach	Natural image quality
PIQE [27]	• Block-wise analysis • Spatial feature extraction	Local distortions
MUSIQ [15]	• Multi-scale processing • Transformer-based	Perceptual quality
PaQ-2-PiQ [30]	• Patch quality analysis • Deep learning based	Global-local quality

6. Time cost of Fine-tuning

Tab.3 provides a comparison of tuning time and performance gains for baseline models when incorporating GenHaze. The table illustrates the balance between inference time on the original baselines and the additional fine-tuning time required with GenHaze. For TaylorFormer [23], the inference time is approximately 17 minutes for the baseline, and fine-tuning with GenHaze requires 19 minutes. This tuning time significantly improves the performance, reducing FADE by over 0.9 and improving BRISQUE and NIQE scores by approximately 12 and 1, respectively. Similarly, for T^3-DiffWeather [5], the baseline inference time is 22 minutes, while the GenHaze fine-tuning process takes 25 minutes, yielding substantial improvements, including a FADE reduction exceeding 1.4 and notable gains in BRISQUE and NIQE. These results demonstrate that the fine-tuning time required for integrating GenHaze is relatively modest compared to the overall inference time of the baselines. At the same time, it delivers significant performance enhancements, underscoring the practical benefits of adopting GenHaze for efficient and high-quality dehazing.

Table 3. More comparisons of tuning time and performance gains for baselines on the RTTS [16] dataset.

Settings	FADE ↓	BRISQUE ↓	NIQE↓	Inference Time /Fune-tuning Time
TaylorFormer [23]	1.9827	28.51	4.66	17min
+GenHaze	-0.9671	-11.84	-1.05	19min
T^3-DiffWeather [5]	2.3771	30.14	5.20	22min
+GenHaze	-1.4000	-8.13	-1.32	25min

7. Additional Ablation Studies

In this section, we supplement the ablation experiments on baseline fine-tuning of the dehazing method. Specifically, we use MSBDN [9] to conduct experiments on the RTTS [16] dataset. Other parameter settings remain the same as the previous configuration.

Table 4. Ablation studies of epochs in fune-tuning process.

Settings	FADE ↓	BRISQUE ↓	NIQE↓	PIQE ↓
0 (Baseline)	1.5818	28.51	4.66	44.96
1 (ours)	0.6881	**20.23**	**3.65**	**28.26**
2	**0.6734**	21.07	3.69	30.12
3	0.7136	23.45	3.99	32.45
4	0.7392	24.06	4.08	33.72

Number of epochs in fune-tuning. Tab.4 highlights the impact of the number of fine-tuning epochs on the performance of dehazing baselines adapted with GenHaze. The results demonstrate that fine-tuning for 1–2 epochs achieves the optimal balance between improving model performance and maintaining generalization. Specifically, fine-tuning for 1 epoch significantly reduces FADE from 1.5818 (baseline) to 0.6881, while also yielding substantial improvements in BRISQUE, NIQE, and PIQE metrics. Extending the fine-tuning to 2 epochs further enhances dehazing performance slightly, with FADE decreasing to 0.6734, indicating that the haze images generated by GenHaze effectively unlock the potential of existing baselines with minimal training effort.

However, fine-tuning beyond 2 epochs leads to a progressive decline in performance, as shown by the increasing FADE and PIQE scores for 3 and 4 epochs. This suggests that extended fine-tuning may disrupt the pre-trained capabilities of the baseline models, introducing unnecessary adjustments that negatively impact their performance. To address this, we set the number of fine-tuning epochs to 1–2 based on the characteristics of each baseline, ensuring both computational efficiency and effective performance enhancement without overfitting. ***While reducing the learning rate or increasing the training dataset size could potentially mitigate these issues, such modifications are beyond the scope of this work. Therefore, we do not explore these adjustments further in this study.***

Training from scratch v.s GenHaze. We provide additional details regarding the experimental setup shown in Tab.5. For training from scratch, the haze reference images are collected from the internet, while clean images are selected entirely from the clean subset of the Allweather dataset. The training dataset is fully synthesized using GenHaze's single-step reference-controlled generation strategy. All other training hyperparameters (e.g., optimizer type, learning rate, number of epochs, and loss functions) remain consistent with the baseline codes.

Table 5. Com. of the training from scratch v.s GenHaze.

Method	1. Baseline		1 + GenHaze			2. Trained on our syn. images			2 + GenHaze	
	FADE↓	BRISQUE↓	FADE↓	BRISQUE↓	Time	FADE↓	BRISQUE↓	Time	FADE↓	BRISQUE↓
MSBDN	1.5818	28.51	**0.6881**	**20.23**	≈14 min	0.8312	21.39	≈1200 min	**0.6524**	**18.89**
Dehamer	1.8926	33.07	**0.8263**	25.12	≈16 min	0.8937	**23.97**	≈1350 min	**0.7963**	**21.19**

As demonstrated in Tab.5, models trained from scratch on datasets generated by GenHaze significantly outperform those trained on existing synthetic datasets, confirming that GenHaze generates superior-quality data. However, training from scratch still slightly underperforms compared to the baseline enhanced by the GenHaze strategy, due to GenHaze's capability of targeted optimization for specific haze conditions at a much faster speed (85 times faster). Furthermore, applying GenHaze's single-step reference-controlled strategy to the scratch-trained models can potentially further improve dehazing performance.

Overall, Tab.5 highlights two key advantages of GenHaze:

i. It generates high-quality haze datasets, significantly enhancing dehazing performance when training models from scratch.

ii. Its single-step reference-controllability rapidly improves baseline performance through a plug-and-play strategy, even for strong baselines.

Dimensions of learnable prompts. Regarding the learnable prompts, our aim is for them to adaptively capture the characteristics of the current haze image based on the general text embeddings. Thus, the dimension of the learnable prompts is aligned with that of the general embedding encoded by CLIP during the offline stage, which is initialized at 1024. Nevertheless, in order to conduct a more comprehensive investigation, we explore various dimensions, including 512, 768, 1024, 1280, and 2048. To ensure compatibility, a 1x1 convolutional layer is employed for dimensional alignment. As shown in Tab.6, the misalignment of dimensions and the subsequent alignment process result in suboptimal outcomes, likely due to the introduction of unnecessary noise into the prompt embeddings.

Table 6. Abl. studies of physics-based generation in fune-tuning.

	w/o. Text Embedding	w/o. Learnable Prompts	512	768	**1024**	1280	2048
FID↓	58.89	57.99	57.61	57.57	**57.12**	57.45	57.59
sFID↓	107.56	106.53	105.92	105.99	**105.43**	105.79	105.87

8. Visual Comparisons of Haze Generation

The visual comparisons in Figs.3, 4, and 5 showcase the effectiveness of GenHaze in generating realistic hazy images compared to other existing methods, including RIDCP [28], Low-Res [3], InstructPix2Pix [2], CycleGAN-Turbo [22], and UWNR [29]. Each figure illustrates clean images (a), real-world haze images as references (b), and the results from GenHaze (c) alongside competing methods (d–h). In all examples, GenHaze demonstrates its superior ability

to generate haze effects that closely align with real-world haze in terms of density, texture, and consistency. Unlike methods such as Low-Res [3], which often produce noticeable artifacts and color shifts, GenHaze avoids such inconsistencies and retains a natural appearance. Furthermore, compared to RIDCP [28], which struggles with distribution alignment, GenHaze achieves higher fidelity and better matches the reference haze characteristics.

In addition, the comparison also highlights GenHaze's robustness across diverse scenes, including urban areas, natural landscapes, and dynamic settings such as running events. For instance, in Figs.3 and 4, GenHaze accurately reproduces subtle variations in haze density and scattering effects, maintaining the original scene's structural integrity. Methods like UWNR [29] and CycleGAN-Turbo [22], on the other hand, often fail to maintain these subtle details, leading to oversmoothed or visually implausible results.

Moreover, unlike InstructPix2Pix [2] and other controllable pipelines, GenHaze achieves both one-step high-quality haze generation and controllability without requiring extensive parameter tuning. As seen in Fig.5, GenHaze successfully integrates diverse haze intensities while preserving the image's clarity and background details, which are often distorted by alternative methods.

9. Visual Comparisons of Image Dehazing.

Fig.6 and 7 showcase the effectiveness of GenHaze in unlocking the potential of existing dehazing baselines across various real-world hazy scenarios. These results underline the significant enhancements achieved in image dehazing when GenHaze is integrated into diverse baseline models.

In Fig.6, baselines such as MSBDN [9], Dehamer [13], Dehazeformer [25], TaylorFormer [23], T^3-DiffWeather [5], and PSD [6] exhibit varying levels of dehazing performance, but many struggle to fully recover details or maintain realistic color distributions, particularly in challenging haze conditions. By incorporating GenHaze, these models demonstrate remarkable improvements, with clearer outputs, enhanced structural consistency, and more natural color representation. For instance, intricate details in complex scenes, such as trees and human figures, are better preserved after the integration.

Fig.7 further highlights the ability of GenHaze to enhance baseline models across urban and other environments. Baselines like FocalNet [7], KANet [11], and Dehazeformer [25], which initially encounter issues such as oversaturation or uneven dehazing, achieve more robust and consistent results with the aid of GenHaze. These improvements are especially pronounced in dense haze and complex lighting conditions, where GenHaze helps produce more realistic and visually appealing outputs. These comparisons clearly demonstrate that GenHaze serves as an effective and efficient solution to amplify the capabilities of existing dehazing baselines, significantly improving their performance and adaptability across diverse haze scenarios.

10. Limitation and Future Work

While GenHaze effectively enables one-step, controllable haze generation guided by real reference images, its performance remains influenced by inherent limitations of latent diffusion models (LDMs). Future work includes adapting our Clean-to-Haze pipeline to alternative generative frameworks beyond latent diffusion models, such as masked autoregressive or token-based autoregressive models. In addition, we will leverage the powerful capabilities of GenHaze to build a more comprehensive, diverse, and realistic haze dataset to further enhance generalization and robustness and address a wider range of practical applications.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006. 2, 3

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 6, 7

[3] Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Haoze Sun, Xueyi Zou, Zhensong Zhang, Youliang Yan, and Lei Zhu. Low-res leads the way: Improving generalization for super-resolution by self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25857–25867, 2024. 2, 4, 6, 7

[4] Sixiang Chen, Tian Ye, Kai Zhang, Zhaohu Xing, Yunlong Lin, and Lei Zhu. Teaching tailored to talent: Adverse weather restoration via prompt pool and depth-anything constraint. *arXiv preprint arXiv:2409.15739*, 2024. 4

[5] Sixiang Chen, Tian Ye, Kai Zhang, Zhaohu Xing, Yunlong Lin, and Lei Zhu. Teaching tailored to talent: Adverse weather restoration via prompt pool and depth-anything constraint. In *European Conference on Computer Vision*, pages 95–115. Springer, 2025. 5, 7

[6] Zeyuan Chen, Yangchao Wang, Yang Yang, and Dong Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7180–7189, 2021. 1, 4, 7

[7] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Focal network for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13001–13011, 2023. 4, 7

[8] P Kingma Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014. 4

[9] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2157–2167, 2020. 4, 6, 7

[10] Raanan Fattal. Dehazing using color-lines. *ACM transactions on graphics (TOG)*, 34(1):1–14, 2014. 5, 13

[11] Yuxin Feng, Long Ma, Xiaozhe Meng, Fan Zhou, Risheng Liu, and Zhuo Su. Advancing real-world image dehazing: perspective, modules, and training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 4, 5, 7

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 3

[13] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5812–5820, 2022. 4, 7

[14] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[15] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 5

[16] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018. 5, 6, 12, 13

[17] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.*, 28(1):492–505, 2019. 4

[18] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5404–5411, 2024. 2

[19] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[20] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708, 2012. 5

[21] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 5

[22] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024. 2, 4, 5, 6, 7

[23] Yuwei Qiu, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, and Zhi Jin. Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12802–12813, 2023. 4, 5, 7

[24] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[25] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *arXiv preprint arXiv:2204.03883*, 2022. 4, 7

[26] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8): 3998–4011, 2018. 5

[27] Narasimhan Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*, pages 1–6. IEEE, 2015. 5

[28] Rui-Qi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, and Chongyi Li. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22282–22291, 2023. 1, 2, 4, 5, 6, 7

[29] Tian Ye, Sixiang Chen, Yun Liu, Yi Ye, Erkang Chen, and Yuche Li. Underwater light field retention: Neural rendering for underwater imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 488–497, 2022. 2, 5, 6, 7

[30] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3585, 2020. 5

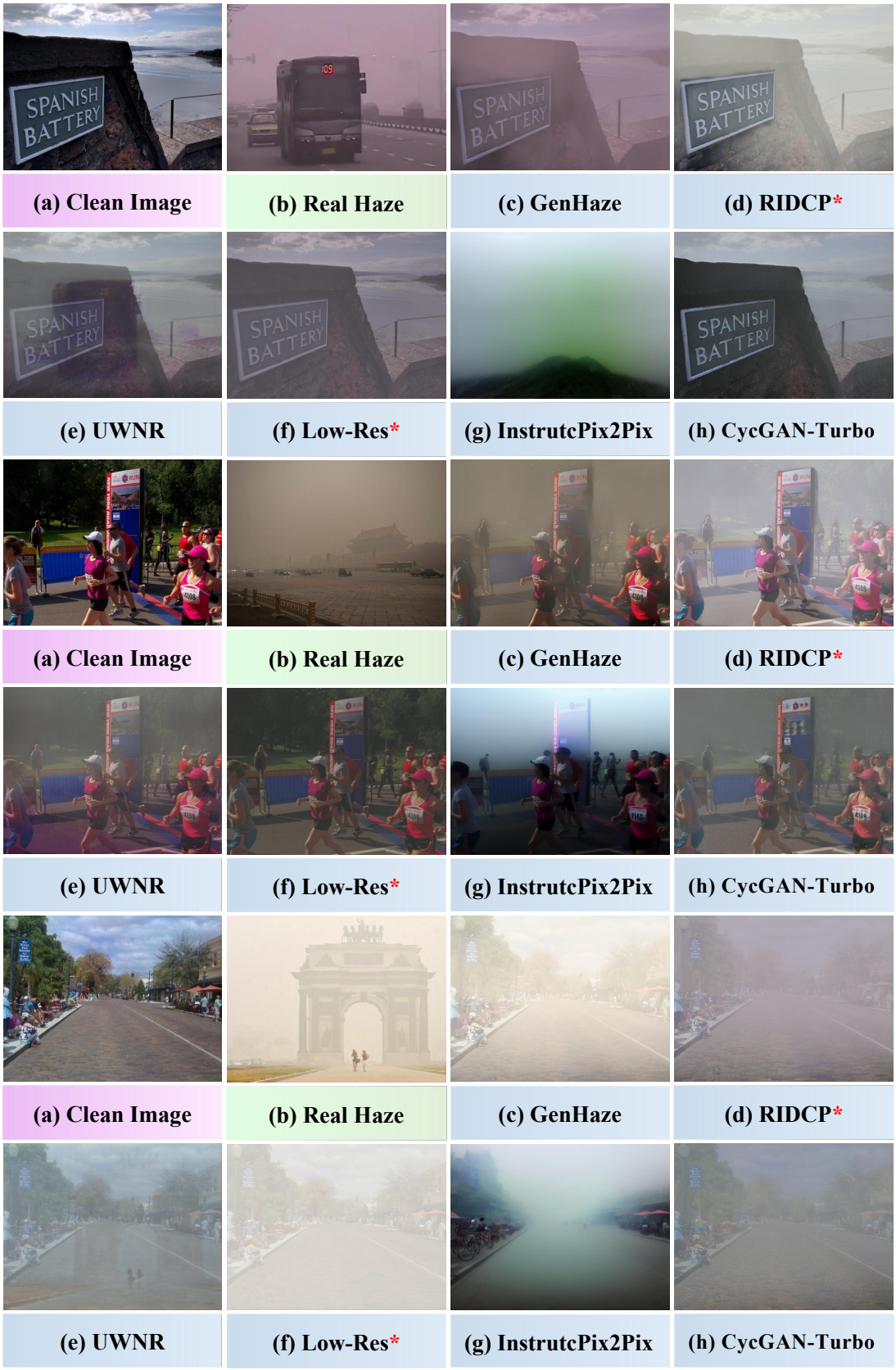


Figure 3. More visual comparisons of haze generation. Our Genhaze generates high-quality hazy images that closely match (b) real-world haze images.

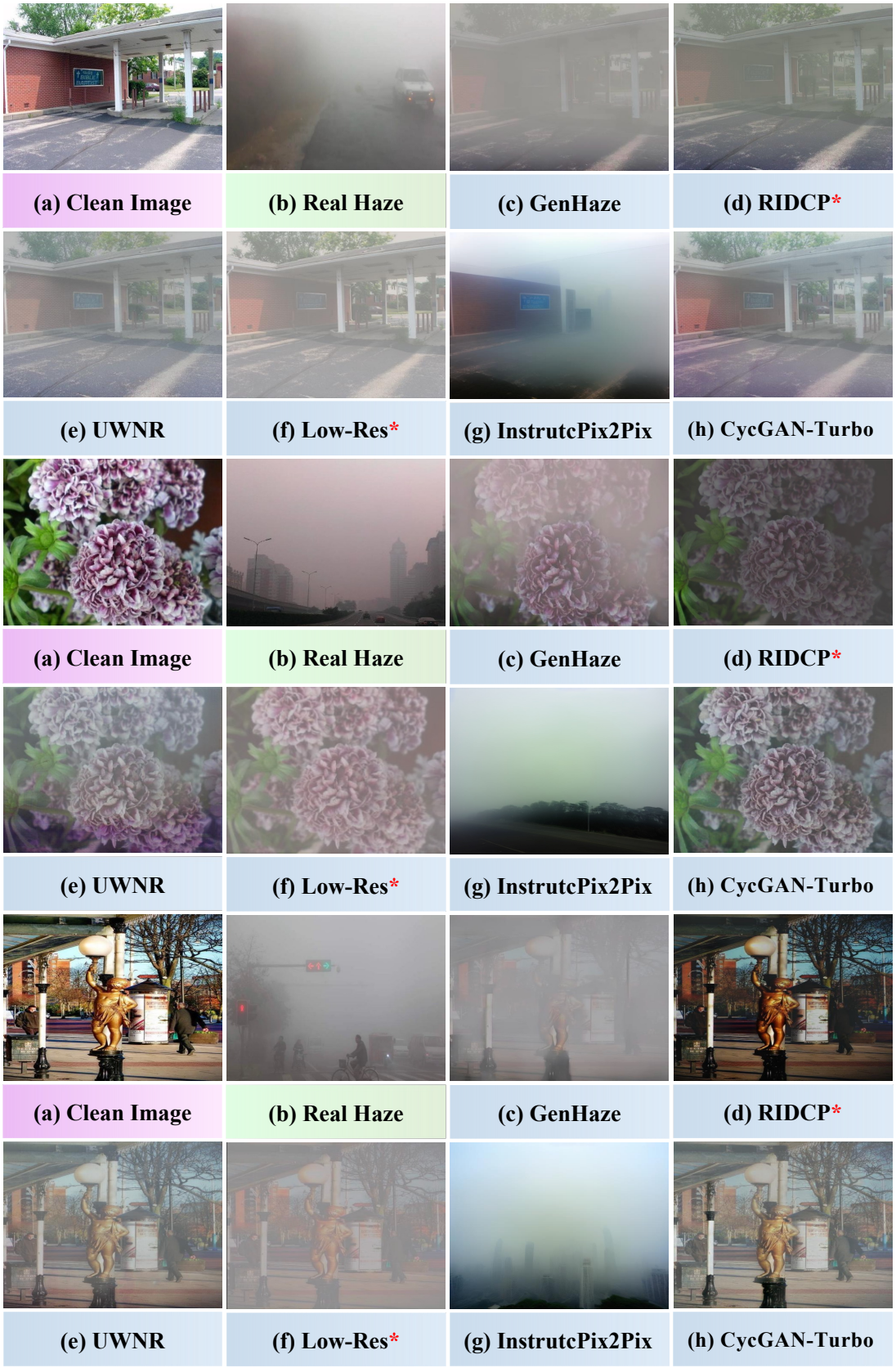


Figure 4. More visual comparisons of haze generation. Our Genhaze generates high-quality hazy images that closely match (b) real-world haze images.

Figure 5. More visual comparisons of haze generation. Our Genhaze generates high-quality hazy images that closely match (b) real-world haze images.

Figure 6. More visual comparisons of multiple baselines on real-world haze samples [16], before and after integrating GenHaze.

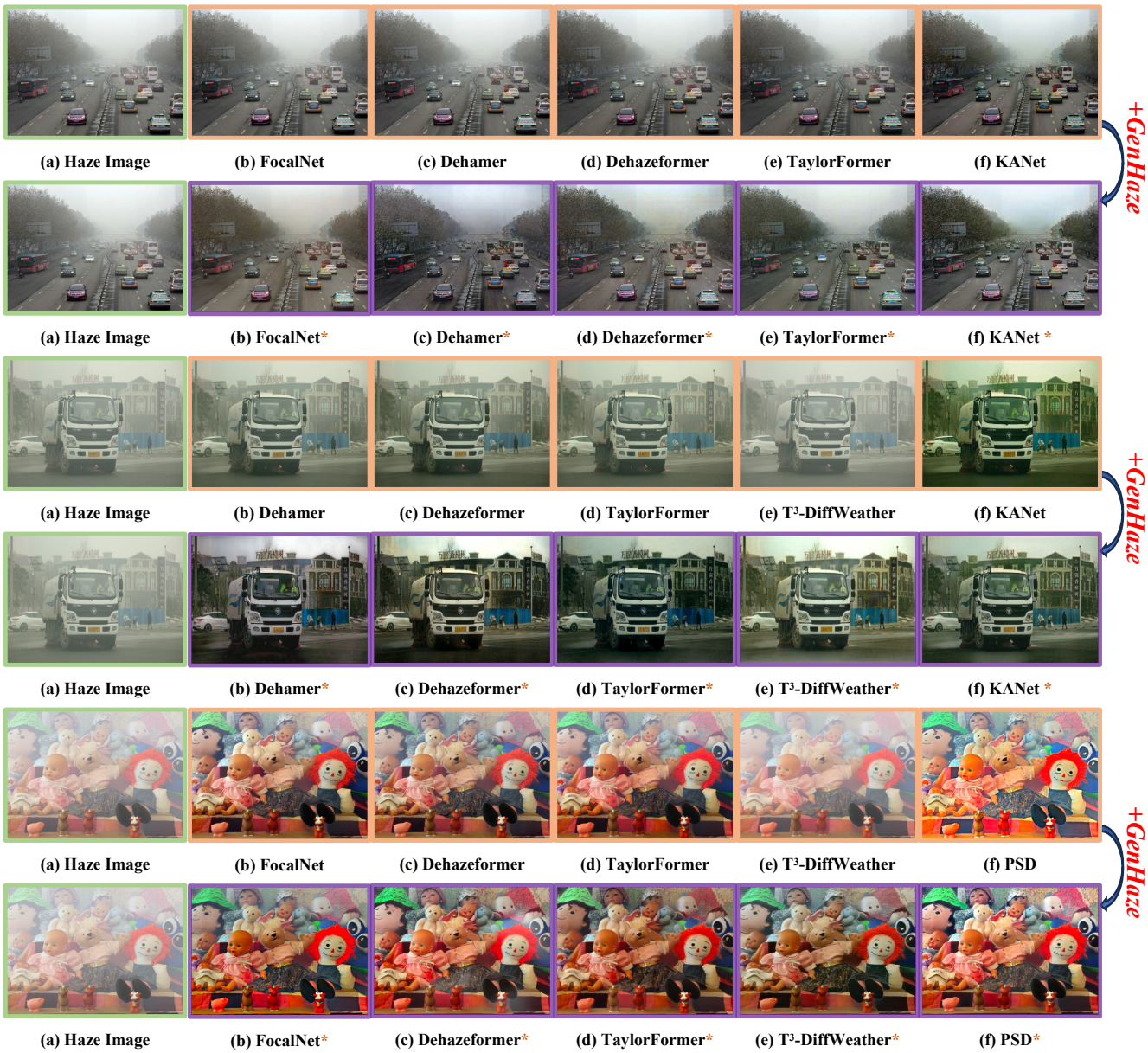


Figure 7. More visual comparisons of multiple baselines on real-world haze samples [10, 16], before and after integrating GenHaze.