Structure-retained low-rank adapters for weather synthesis

Shunxin Wang Alexandros Stergiou Luuk Spreeuwers Estefania Talavera
Nicola Strisciuglio
shunxinwang1030@gmail.com

Abstract

Real-world deployment of computer vision models, particularly in outdoor scenarios, is often limited by environmental interferences, such as adverse weather. Improving model robustness across weather conditions is thus critical for reliability. However, sourcing high-quality data relevant to isolated weather effects is a significant challenge. Most datasets include changes in dynamic elements across observations of the same scene: subjects such as pedestrians and vehicles change between shots, while camera viewpoints vary. These aspects make it difficult to decouple solely weather-induced distribution shifts from the scene. In this work, we address this by a novel weather synthesis framework that utilizes structure-retaining Low-Rank Adapters (LoRA). We analyze the attention blocks in Stable Diffusion XL (SDXL) and identify where relevant structural information is mostly processed during image-to-image translation. Through an architectural analysis of SDXL, we learn structural LoRA matrices for the corresponding blocks and then leverage them for weather synthesis, eliminating the need for external conditioning such as ControlNet. Experimental results demonstrate that our method improves upon existing synthesis approaches, achieving greater realism while preserving structure. This provides a robust pipeline for generating high-fidelity training and evaluation data for outdoor computer vision applications.

1. Introduction

Computer vision models deployed in real-world scenarios often encounter robustness problems. In outdoor settings, interference comes not only from hardware, such as photon and thermal noise (especially in low-light settings), but also from unpredictable environmental changes, such as weather conditions. Improving model performance is important for ensuring the reliability of deployed models, especially in safety-critical applications such as autonomous driving. However, collecting high-quality, sufficiently varied image data that contains various weather changes remains a challenge. Sourcing real-world image data from the same scene in different weather conditions requires comprehensive control over dynamic objects. Existing datasets mostly contain image pairs under different conditions, e.g., sunny-rainy and sunny-foggy, without ensuring that the original objects present in the images (e.g., cars and passengers) or camera views are preserved [12, 39]. This induces not only weather-related distribution shifts, but also temporal and spatial variations.

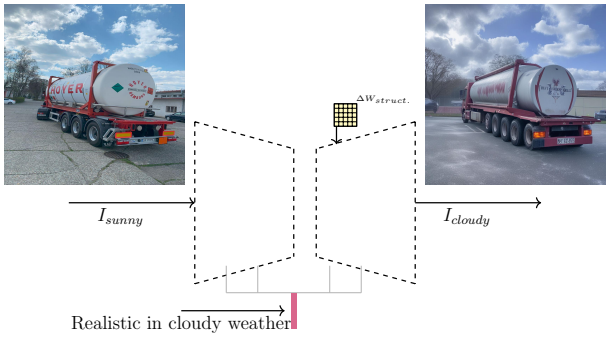


Figure 1. Text-guided weather manipulation with structure-LoRA for preserving the structural layout of input images.

Generative models have evolved to produce high-quality images while offering granular control over details like style (e.g., weather) and specific scene content [22, 24]. For style manipulation, previous works extracted style information from one reference image [14], or from a pair of images with and without the target style [15]. For text-to-image generation tasks, image style can be manipulated by changing style-related text descriptions. However, for image-to-image translation, manipulating the input image style using only text prompts is insufficient, either failing to produce strong changes or disrupting the image content [1, 32].

Synthesizing adverse weather conditions is a critical frontier for improving computer vision models deployed in outdoor scenarios, e.g., autonomous driving. Previous work relied on Generative Adversarial Networks (GANs), such as CycleGAN [40] that simulated diverse weather types. These methods often require high-quality paired training data, which is expensive to obtain. To circumvent this, [16]

turned to graphics engines and large-scale generative models. For instance, Stable Diffusion XL (SDXL) [21] is used to curate precisely controlled weather attributes within synthetic scenes. Despite the increased variety these models offer, the original scene structure is often not preserved during image-to-image translation. ControlNet [36] was proposed to preserve structure information of an input image, by injecting structured spatial conditions, e.g., edges and depth, into intermediate feature maps. But this often leads to a loss of fine-grained details, such as original textures and object identities [13]. Other approaches that adopt instruction-based editing, such as InstructPix2Pix [2], frequently struggle with spatial relational reasoning, i.e., the models either focus solely on local effects (e.g., rain in the sky) while neglecting global consequences (e.g., reflections on the road), or alter global image color rather than specific weather elements.

In this work, we perform an architectural analysis of SDXL on the sensitivity of attention blocks to structural information, and subsequently devise a structure-retaining Low-Rank Adaptation (LoRA) [11] strategy to synthesize weather while preserving structure details. Compared to other weather synthesis methods, our method improves the variety of weather effects while requiring light computations. Compared with other style-manipulation approaches, our method preserves more structural and content information. Effectively, we reduce the need to curate high-quality paired image data through manual collection or graphic engines.

Our contributions are as follows:

1. An architectural analysis and identification of SDXL attention blocks that capture more structural and dynamic information within scenes;
2. A weather synthesis framework using structural priors in the form of low-rank adapters (structure-LoRA), to retain the layout and dynamic elements of input images. Our method improves realism and weather-synthesis effects while better preserving scene dynamic objects and structural layout compared to existing approaches.

2. Related works

Style transfer Image style transfer aims to change the style of a source image to match that of a reference image. With the development of vision-language models (VLMs), works have leveraged encoded semantic knowledge. In [9], it was shown that cross-attention layers in text-conditioned diffusion models extract spatial relationships between image pixels and prompt tokens. Manipulating cross-attention maps by changing only the text prompts (e.g., replacing a word or adding a specification) allows editing an image while preserving most of the original structural information. Similarly, the utility of self-attention for capturing spatial relationships (e.g., color layouts) and local visual features (e.g., textures)was investigated in [14]. Replacing attention matrices of style reference images with those of a content image can preserve the original content while manipulating the style of the original image. However, some content of the style reference image might leak to the generated images, since content and style information extracted in the attention layers are entangled. By enabling 'attention sharing' during the diffusion process, style-consistent images with the first (i.e., style reference) image in a batch were generated, without optimization [10] . Explicitly learning style information, InST [38] maps the style representation of a reference image to text tokens that do not exist in the normal language, e.g., `[C]`, and then provides such a text prompt to manipulate the style of the input images. Instead of text prompts as conditions, IP-Adapter [35] additionally injected image information extracted from a style/structure reference image into the attention layers of diffusion models. However, this approach is limited to either style manipulation or structure generation. For text-to-image (T2I) style transfer, the authors of [3] used a style encoder to extract multi-level style embeddings aligned with textual descriptions alongside an auxiliary content adapter to separate content and style.

To manipulate style information more precisely, several works have focused on decomposing style and content. This is done by disentangling and then replacing the style or content components of the images: the generated images may change in subject or style [30, 34]. The authors of [6] explored the sensitivity of each block of StableDiffusion XL to style or content information. This analysis guides the implicit separation of a single image style and content components by jointly learning the LoRA weights of the content-sensitive and style-sensitive blocks. UnZipLoRA [17] disentangles a single stylized image into content and style information as two independent LoRAs by encouraging orthogonality through regularization of LoRA weight matrices. PairCustomization [15] learns the stylistic differences from an image pair (a content image and a style reference image with similar structure) by jointly optimizing content-LoRA to reconstruct the content image and the style-LoRA to edit the style of the content image. However, these methods require per-pair or per-image optimization, which can be computationally expensive and impractical for real-time inference. The work of [25] introduces a hypernetwork that predicts the merging coefficients of content and style LoRAs, which generalizes well to unseen content and style LoRA pairs, thus eliminating the need for costly optimization. CSGO [33] trains a style-transfer model that takes style and content images as inputs, and learns to fuse content and style features by a feature injection block. This allows for style transfer during inference without fine-tuning or optimization.

Rather than learning LoRA weights for style and con-

tent information, the work of [4] found that using the same noise sampling distributions and loss weighting schedules as those in the original model's training might not address style features during diffusion. Namely, stylistic features are more prominent in the early denoising steps, which have a low signal-to-noise ratio, than in later steps. Based on this, [4] shifted the mean of the noise sampling distribution to induce more noise during fine-tuning, allowing for better style alignment.

Our structure-LoRA eliminates the need for a style reference image and modulates the weather of original images using only text prompts, thus avoiding content leakage. Importantly, structure-LoRA preserves the structure and content of input images, whereas existing approaches to style manipulation either require additional structural conditions or do not address structure preservation altogether.

Weather synthesis Adverse weather synthesis is important for computer vision applications, such as autonomous driving [20]. It facilitates evaluating model robustness under varying weather conditions and improving vision models in adverse weather. In [20] and [40], GANs and CycleGANs were applied respectively, to generate images with 14 adverse weather conditions. The work of [29] uses both physics-based and GAN-based methods to augment data with rain effects. However, these methods need high-quality paired image data for model training. The work of [16] curated data using graphics engines that precisely control weather attributes of the same scene, as well as a large image generative model such as StableDiffusion XL [21]. With the development of generative vision-language models, the authors of [19] leveraged DALL-E [23] to synthesize weather effects given pre-designed text prompts. To preserve the structural layout of images, the work of [8, 18] used ControlNet [36], which takes structure-condition images, e.g., Canny edge images. However, many precise details may be lost during the generative process, such as textures, color, and identity. Rather than using ControlNet, [26, 27] generated weather-modulated images with instruction editing models, such as InstructPix2Pix[2]. These models are trained on well-crafted synthetic image pairs with diverse and extreme weather conditions generated by advanced conditional diffusion models. However, local details of input images in inference time might not be preserved.

3. Preliminary

We investigate how changes to the condition embeddings affect the structural layout of the generated images.

SDXL [21] consists of a 70-attention block UNet, each with cross- and self-attention blocks. SDXL is trained on a latent diffusion objective $\mathcal{L}_{LDM} = \mathbb{E}_{z\sim\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2]$, where ϵ_θ predicts the noise $\epsilon \sim \mathcal{N}(0, 1)$ added at timestep t to latents $\mathbf{z} \in \mathbb{R}^{H/8\times W/8\times 4}$ corresponding to image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ encoded with $\mathcal{E}(\cdot)$.

ControlNet [36] creates a trainable copy θ' of the UNet encoder part of SDXL, keeping parameters frozen. Zero convolutions, 1×1 convolutions with zero-init weights and biases are added between the frozen and trainable copy. For a given UNet block $\mathcal{F}_\theta$, the output y_c is formulated as $y_c = \mathcal{F}_\theta(\mathbf{z}_F) + \mathcal{Z}(\mathcal{F}_{\theta'}(\mathbf{z}_F + \mathbf{c}))$, where $\mathbf{z}_F \in \mathbb{R}^{h\times w\times d}$ is the input feature map and $\mathbf{c} \in \mathbb{R}^{h\times w\times d}$ is the added condition after preprocessing. We show the partial view of SDXL with ControlNet in Fig. 2a.

LoRA. Low-rank adapters are efficient modules that enable fine-tuning for large pre-trained models by optimizing rank decomposition matrices [11]. Given the weights of a pre-trained text-to-image diffusion model ($W_0 \in \mathbb{R}^{d\times k}$), LoRA learns an update matrix ΔW that can be decomposed into two low intrinsic rank matrices $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$, with the rank $r \ll min(d, k)$ (i.e., $\Delta W = BA$). During training, W_0 is frozen, while the two low intrinsic rank matrices A and B are updated. During inference, the model weights are fine-tuned by $W = W_0 + \Delta W$. This approach promotes efficiency by restricting the number of tunable parameters.

4. Method

4.1. SDXL architectural inspection

Prior research examined how manipulating style or content keywords in text prompts, injected into various UNet blocks, impacts the resulting visual style and content [6]. However, only manipulating text prompt injection and replacement does not allow exploring how the structure information of images can be affected.

In this work, we investigate SDXL with ControlNet, which injects structural priors to each SDXL block. The combined model then generates structure-preserved images guided by a text prompt. Following an approach similar to [6], we investigate how manipulating ControlNet-encoded condition embeddings injected into SDXL affects the generated structure information. Specifically, we use two strategies: (1) masking out one of the condition paths while keeping the others unchanged; (2) keeping only one condition path, removing all others. We replace the masked path(s) with zero values to serve as a null-condition (see Fig. 2). During image generation, we employ a prompt template `A photo of a [category name]` alongside ControlNet that uses the canny edges of the source image as the additional condition.

Dataset. To visualize and quantify how structural information is affected, we selected images from the Caltech 256 dataset [7], with simple backgrounds and clear object con-

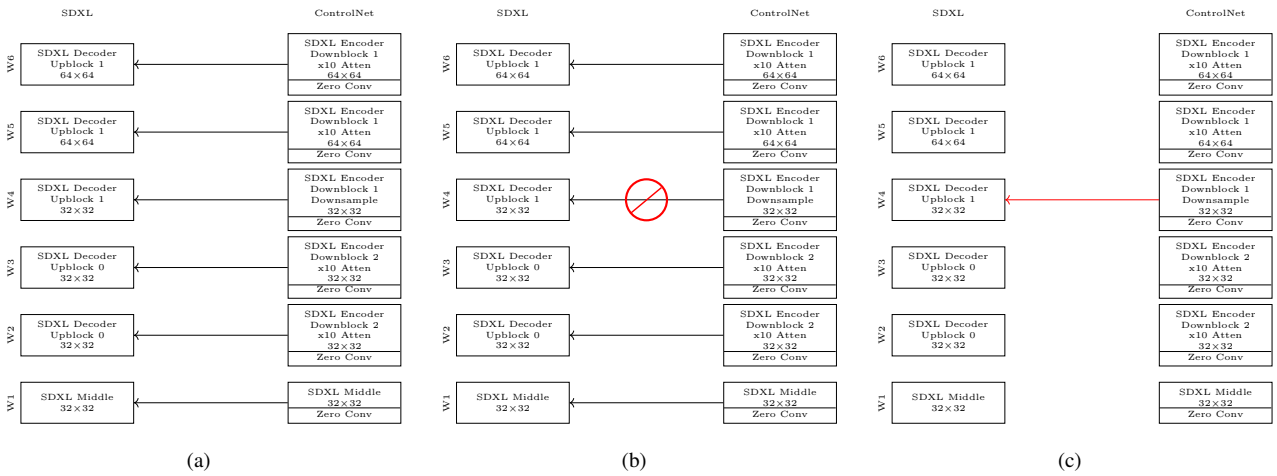


Figure 2. (a) Partial SDXL-ControlNet architecture. We use two strategies: (b) removing one of the embeddings and (c) retaining only one embedding to examine the influence of the condition embeddings fed to different blocks of SDXL on the structural information of the generated images. The weights of each block are denoted as Wi.

tours[1].

Visualization results. Masking out the condition embedding fed to the middle block (noted as W_1) of SDXL results in images that do not preserve the original structure (see Fig. 3a: the cormorant bird in the third image disappears, the teapot becomes unrecognizable, and the iguana is zoomed in). Similarly, we observe from Fig. 3b that retaining only the condition from the middle block enables SDXL to recover most structural details. Thus, for structure-preserving weather synthesis, we target optimizing the LoRA weights of the middle block of SDXL.

4.2. Structure-LoRA for weather synthesis

Guided by the insights gained, we design a structure-enhanced weather-synthesis adapter, offering a more lightweight alternative to traditional ControlNet-based conditioning. We follow DreamBooth LoRA [24] and fine-tune SDXL to reconstruct the input. We use a general prompt `A photo of [$v_i$]`. We avoid using `A photo of a [category name]`, since prompts with category names strongly bias the image generation towards the specific category, reducing the utility of the learned LoRAs. We follow the optimization setup of [6] to learn different LoRAs, each representing style, content, and structural information. During inference, given an input image for weather synthesis, we use a prompt template `realistic in [rainy/cloudy/snowy/foggy] weather` to guide the generation towards different weather effects, using specific learned LoRA weights (e.g. structure (struct.)-LoRA, content (cont.)-LoRA, style (sty.)-LoRA or a combination of them). The process is depicted in Fig. 4.

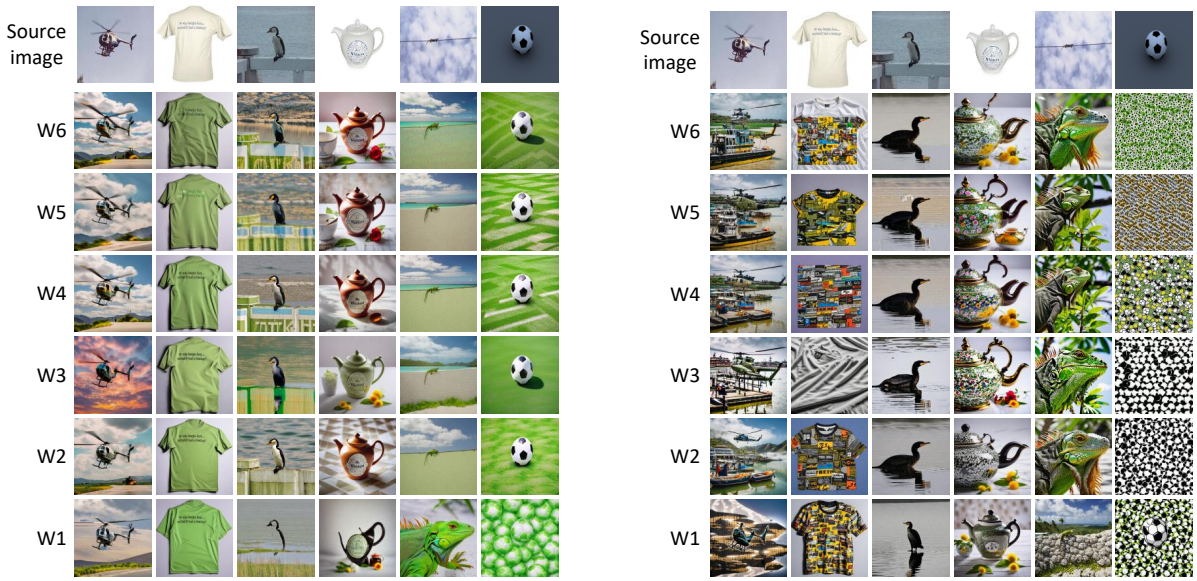
Metrics. We use the Structure Similarity Index Measure (SSIM) [31] and the Learned Perceptual Image Patch Similarity (LPIPS) [37] to evaluate the preservation and realism of the structure of the generated images. SSIM quantifies the structural fidelity between source and generated images, ranging from 0 to 1, with higher values indicating better structure preservation. LPIPS measures perceptual differences aligned with human visual judgment using deep network feature embeddings. Lower LPIPS scores indicate greater perceptual similarity to human observation and thus better image quality and realism.

To evaluate the success rate of weather effect synthesis, we deploy an *LLM-as-judge* approach, using the prediction output of Gemini 2.5 flash. Giving a text prompt `You are an image analysis model specializing in meteorological observations. Your task is to detect and describe visible weather effects in the provided image. Respond STRICTLY according to the JSON schema provided in Listing 1.` The detected dominant weather effect is taken as the prediction of the model.

5. Experiments

Baselines and training settings. We select B-LoRA [6] and PairCustomization [15] as baselines. These two parameter-efficient fine-tuning methods are designed for style manipulation using style and content disentangled

[1]The classes are: `human skeleton`, `helicopter`, `bathtub`, `tuning fork`, `T-shirt`, `ladder`, `fighter jet`, `revolver`, `bowling pin`, `cormorant`, `spoon`, `tricycle`, `knift`, `frisbee`, `sword`, `eyeglasses`, `diamond ring`, `tripod ipod`, `teadpot`, `iguana`, `yoyo`, `soccer ball`, `ak47`, and `penguin`.



(a) (b)

Figure 3. Qualitative results analyzing SDXL architecture using conditional embeddings masking out (a) strategy 1 and (b) strategy 2.

LoRA weights, offering a straightforward comparison to our structure-LoRA regarding structural preservation. B-LoRA and PairCustomization can generate images in two setups, which are (1) text-to-image generation (T2I) and (2) image-to-image generation (I2I). Our method uses I2I generation to preserve structural information as much as possible, while manipulating the weather style with fewer constraints from the original image input. For a comprehensive analysis of the performance, we compare our method to the baseline methods using both settings. During I2I, the hyperparameter 'strength' controls the guidance scale of text prompts. A lower strength indicates details of an input image are largely preserved, while a higher strength indicates significant changes applied to the input image guided by the text prompt.

In our baseline experiments, we learn the weather style LoRA from a reference image depicting rain, snow, fog, or clouds. B-LoRA extracts style LoRA weights from a reference image or a set of similarly styled reference images. PairCustomization learns style-LoRA and content-LoRA from the pairwise differences between images. For weather conditions, this can include `sunny` - [`rainy`/`snowy`/`foggy`/`cloudy`] image pairs in exactly the same scene. Natural image datasets containing such image pairs are hard to obtain. We thus use Gemini 2.5 Flash Image (Nano Banana) [5] to generate these image pairs with text prompts modifying only the weather description.

Dataset. We select haztruck dataset, which contains truck images with hazmat plates, sourced from the internet [28].

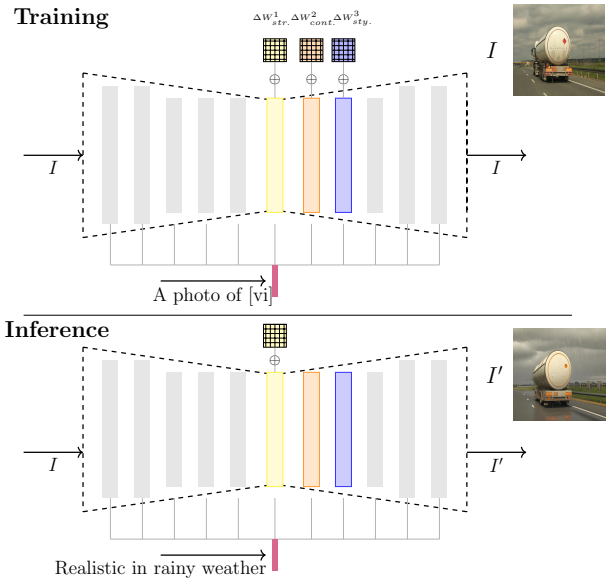


Figure 4. Learning LoRA matrices for structure, style and content information separately. During inference, we apply the structure LoRA matrix for preserving the structural layout of images.

The dataset is curated for hazmat plate detection and optical character recognition. After filtering low-quality images, we use 174 images out of 238 images. Example images are shown in Fig. 5.

(a) (b)

Figure 5. Example images from Haztruck dataset.

```
{
    "type": "OBJECT",
    "properties": {
        "sky_condition": {
            "type": "STRING",
            "enum": ["clear", "partly_cloudy", "overcast", "stormy",
            ↪  "foggy", "hazy"],
            "description": "The dominant sky condition."
        },
        "precipitation": {
            "type": "STRING",
            "enum": ["none", "rain", "snow", "hail", "sleet", "drizzle"],
            "description": "The type of precipitation observed."
        },
        "lighting_condition": {
            "type": "STRING",
            "enum": ["bright", "dim", "sunset", "night", "artificial"],
            "description": "The observed lighting condition."
        },
        "visibility": {
            "type": "STRING",
            "enum": ["good", "moderate", "poor"],
            "description": "The general visibility level."
        },
        "surface_condition": {
            "type": "STRING",
            "enum": ["dry", "wet", "snowy", "icy", "muddy"],
            "description": "The condition of the ground surface."
        },
        "wind_effects": {
            "type": "STRING",
            "enum": ["none", "visible motion blur", "bent trees", "blowing
            ↪  debris"],
            "description": "Observable signs of wind."
        },
        "dominant_weather_effect": {
            "type": "STRING",
            "enum": ["sunny", "cloudy", "rainy", "snowy", "foggy", "stormy",
            ↪  "other"],
            "description": "The single most dominant weather phenomenon."
        },
        "confidence_score": {
            "type": "NUMBER",
            "format": "float",
            "description": "Confidence of the analysis, between 0.0 and
            ↪  1.0."
        },
        "reasoning": {
            "type": "STRING",
            "description": "A brief explanation of the visual cues used for
            ↪  the analysis."
        }
    },
    "required": [
        "sky_condition", "precipitation", "lighting_condition",
        ↪  "visibility",
        "surface_condition", "wind_effects", "dominant_weather_effect",
        "confidence_score", "reasoning"
    ]
}
```

Listing 1. JSON schema for the weather condition object.

5.1. Quantitative results

We present the results of image generation quality in Fig. 6, as we evaluated structure preservation, image realism, and

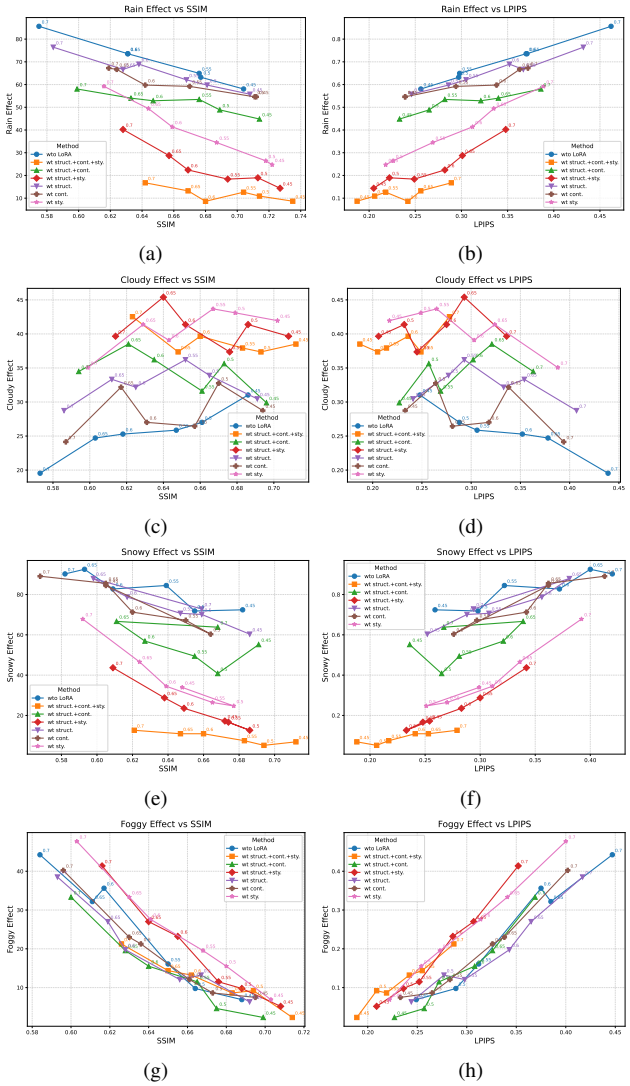


Figure 6. Quantitative results regarding structure preservation and image realism vs. weather effect.

the effectiveness of the weather effect. Annotations in the figure refer to strength values. Generation without LoRA yields the best overall weather-synthesis effect, except for `cloudy` and `foggy` conditions. However, it shows lower SSIM than using a LoRA-enabled model at the same strength setting, indicating worse structure preservation. Using only text prompts without LoRA weights trained for structure preservation indeed limits the controllability of adding weather effects to a given scene and its applicability for systematic robustness tests of downstream tasks. The LPIPS values obtained by SDXL without LoRA are generally higher than those with LoRA at the same strength setting, indicating that the generated images are statistically less perceptually similar, i.e., less realistic. Among the results with different LoRA weights, images

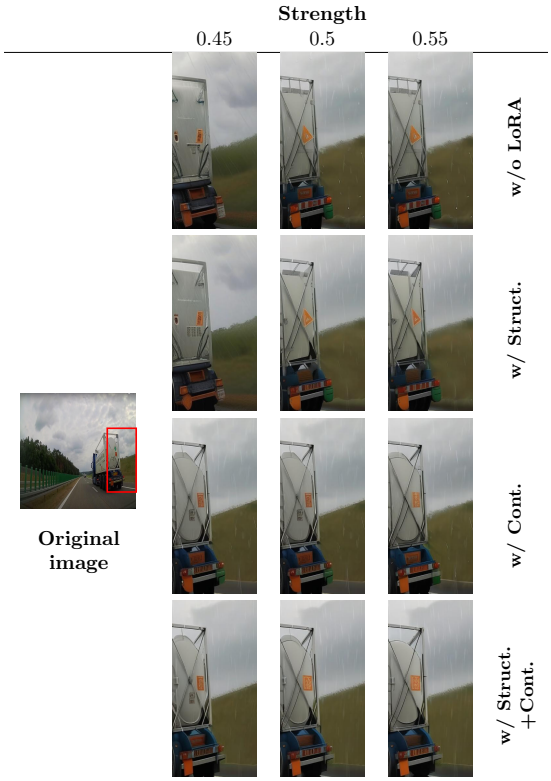


Figure 7. I2I results using strength $= 0.45, 0.5, 0.55$, with LoRA setting: w/o LoRA , w/ stlye, w/ content, w/ style+content LoRA, respectively. Structure-LoRA preserves structure while enable more obvious weather synthesis effect guided by text prompt.

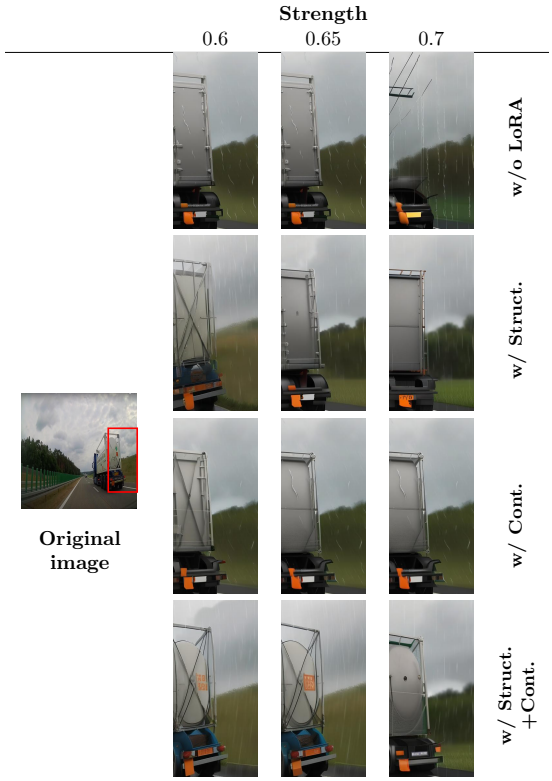


Figure 8. I2I results using strength $= 0.6, 0.65, 0.7$, with LoRA setting: w/o LoRA , w/ stlye, w/ content, w/ style+content LoRA, respectively. Structure-LoRA preserves structure while enable more obvious weather synthesis effect guided by text prompt.

generated with our structure-LoRA weights show a greater degree of weather-effect manipulation. Compared with images generated using content-LoRA weights, those generated with structure-LoRA weights produce better weather synthesis, with a slight loss in structure preservation. This might be that content-LoRA, which represents object information, also contains partial structural information, but its use might limit style/weather manipulation. The weather synthesis success rates when using style-LoRA for cloudy and foggy conditions are generally higher than those using only structure-LoRA. We inspect the original images and find that many of them exhibit a mild, cloudy, or foggy effect, which has been encoded in style-LoRA. But the realism of the generated images using style-LoRA is much lower (see the high LPIPS value when strength=0.7 for cloudy effect synthesis). Combining style and structure LoRAs results in better cloudy and foggy synthesis effect while retaining structure information.

5.2. Qualitative results

In Figs. 7 and 8, we show rain synthesized images using different combinations of LoRA (w/ structure, w/ content, and w/ structure plus content) or without using LoRA. Images generated with lower strength values have almost no observable synthesized weather effect, while a higher value results in synthesized rain streak and reflection on the road. The usage of structure-LoRA preserves most structure information, compared to that without using LoRA (generating images under strength=0.7 configuration replaces the original truck by a car). Observed from the results of LoRA w/ content and LoRA w/ structure plus content, the structure-LoRA better preserves overall structure details of the original image under high strength configuration while allowing for more obvious rainy synthesis effect. Images of other synthesized weather effects are in Fig. 9.

Comparisons with baselines. The quantitative comparisons between our method with B-LoRA and PairCustomization are presented in Tab. 1. Our method balances the synthesis of weather effects with the preservation of the original structure of the input images. We achieve the second-best weather-synthesis performance, except for cloudy weather, while our method preserves the most structural information of the original images (highest SSIM). The realism of our generated images ranks second com-

Method	LPIPS↓	SSIM↑	Rainy↑	Cloudy↑	Snowy↑	Foggy↑	Avg.
Pair (I2I)	0.541	0.547	**95.38**	12.14	**95.95**	29.48	**58.24**
BLoRA (T2I) w/ C	0.726	0.318	60.92	77.59	0	20.69	39.8
BLoRA (T2I) w/o C	0.738	0.316	47.7	**100**	0.57	0	37.07
BLoRA (I2I) w/o C	**0.431**	0.563	53.45	39.66	2.3	**43.1**	34.63
Structure LoRA	0.388	**0.601**	76.44	28.74	72.83	38.51	54.13

Table 1. Performance of different methods: we report LPIPS, SSIM, and weather synthesis success rate. Lower LPIPS and higher SSIM indicate better quality, while higher weather synthesis success rate indicates better weather manipulation.

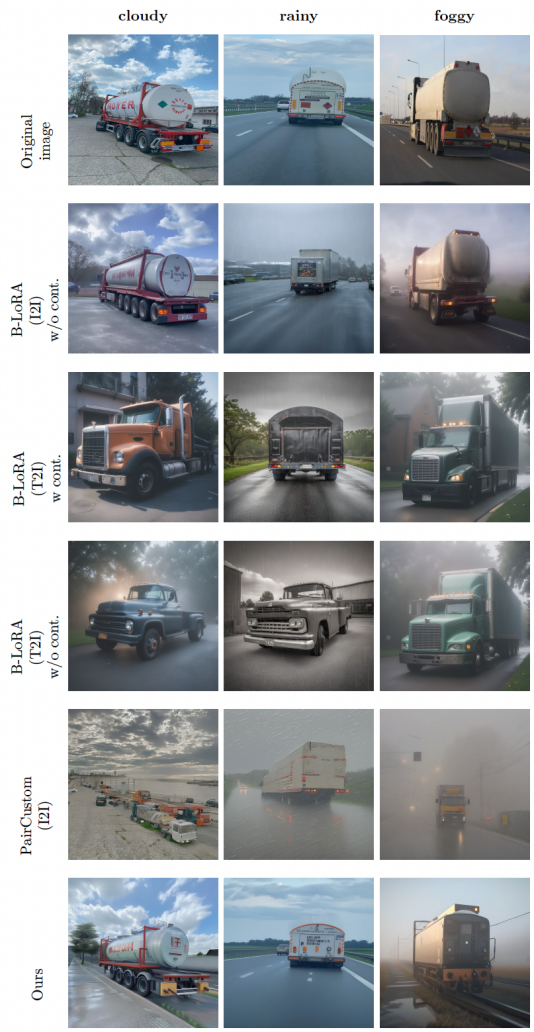


Figure 9. Comparison with B-LoRA and PairCustomization. For B-LoRA, we learn style-LoRA from a synthetic pair of sunny-[cloudy/rainy/foggy] images. We generate images using both I2I and T2I. As text provides limited structural information, we provide content-LoRA (which contains partial structural information) for T2I setup for a fairer comparison. For PairCustomization, we use I2I generation with the learned style-LoRA from the reference image pair.

pared with that of B-LoRA (I2I) w/o C.

We qualitatively compare our weather synthesized images with those generated using B-LoRA and PairCustomization in Fig. 9. Under I2I generation setting, our method maintains more structure information, while adding more realistic weather effects to the images. B-LoRA instead results in almost no synthesized weather effect, and PairCustomization completely changes the structure information of the original images. Using B-LoRA T2I generation option, the generated images demonstrate different structure and content information even using the content LoRA of the original image. These results show that our method preserves more structure information with less compensation regarding the magnitude of style manipulation.

5.3. Limitations

Although structure-LoRA preserves more structural information of images during I2I translation, not all the details of the original images are maintained. For example, the number plate and hazmat sign in the images are important for identifying a truck and the hazardous material being transported. To recover these details, incorporating text inpainting could be a reasonable approach. Moreover, we observe that if an input image contains contrasting weather condition to the target weather condition, the model might fail to synthesize the targeted weather effect. This might be attributed to the insufficient guidance of a given text prompt to cover the original weather effect when the semantic gap between them is large. For future work, disentangling explicitly the weather effect from the original image content would be a promising direction.

6. Conclusions

Through an architectural analysis of SDXL-ControlNet, we uncover that structural condition extracted by the middle block of ControlNet is highly correlated to structure preservation in image generation. We propose a weather synthesis method that learns the corresponding LoRA matrices to image structure, enabling the manipulation of weather effects while retaining the original image structure and content. Our method requires light computations and training time, and improves weather effect variety.

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision – ECCV 2022*, pages 707–723, Cham, 2022. Springer Nature Switzerland. 1

[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 (CVPR)*, pages 18392–18402, 2023. 2, 3

[3] Dar-Yen Chen, Hamish Tennent, and Ching-Wen Hsu. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8619–8628, 2024. 2

[4] Jooyoung Choi, Chaehun Shin, Yeongtak Oh, Heeseung Kim, Jungbeom Lee, and Sungroh Yoon. Style-friendly snr sampler for style-driven generation, 2025. 3

[5] Alisa Fortin, Guillaume Vernade, Kat Kampf, and Ammaar Reshi. Introducing gemini 2.5 flash image, our state-of-the-art image model. *Google Developers Blog*, 2025. 5

[6] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *Computer Vision – ECCV 2024*, pages 181–198, Cham, 2025. Springer Nature Switzerland. 2, 3, 4

[7] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech 256, 2022. 3

[8] Chenchen He, Xia Yuan, Yi Zheng, Xi Wu, Jing Hu, and Ting Wang. Rendering extreme weather phenomena using generative ai base on controlnet. In *2024 IEEE 8th International Conference on Vision, Image and Signal Processing (ICVISP)*, pages 1–6, 2024. 3

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022. 2

[10] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4775–4785, 2024. 2

[11] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2, 3

[12] Xiaowei Hu, Lei Zhu, Tianyu Wang, Chi-Wing Fu, and Pheng-Ann Heng. Single-image real-time rain removal based on depth-guided non-local features. *IEEE Transactions on Image Processing*, 30:1759–1770, 2021. 1

[13] S. Huang, Q. Li, J. Liao, et al. Controllable image synthesis methods, applications and challenges: a comprehensive survey. *Artificial Intelligence Review*, 57:336, 2024. 2

[14] Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. Visual style prompting with swapping self-attention, 2024. 1, 2

[15] Maxwell Jones, Sheng-Yu Wang, Nupur Kumari, David Bau, and Jun-Yan Zhu. Customizing text-to-image models with a single image pair. In *SIGGRAPH Asia 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2, 4

[16] Chih-Hao Lin, Zian Wang, Ruofan Liang, Yuxuan Zhang, Sanja Fidler, Shenlong Wang, and Zan Gojcic. Controllable weather synthesis and removal with video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13580–13591, 2025. 1, 3

[17] Chang Liu, Viraj Shah, Aiyu Cui, and Svetlana Lazebnik. Unziplora: Separating content and style from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16776–16785, 2025. 2

[18] Thibaut Loiseau, Tuan-Hung Vu, Mickael Chen, Patrick Pérez, and Matthieu Cord. Reliability in semantic segmentation: Can we use synthetic data? In *Computer Vision – ECCV 2024*, pages 442–459, Cham, 2025. Springer Nature Switzerland. 3

[19] Aboli Marathe, Deva Ramanan, Rahee Walambe, and Ketan Kotecha. Wedge: A multi-weather autonomous driving dataset built from generative vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3318–3327, 2023. 3

[20] Valentina Muşat, Ivan Fursa, Paul Newman, Fabio Cuzzolin, and Andrew Bradley. Multi-weather city: Adverse weather stacking for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2906–2915, 2021. 3

[21] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[22] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[23] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[24] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023. 1, 4

[25] Donald Shenaj, Ondrej Bohdal, Mete Ozay, Pietro Zanuttigh, and Umberto Michieli. Lora.rar: Learning to merge loras via hypernetworks for subject-style conditioned image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16132–16142, 2025. 2

[26] Edward Sun. Conddg: Enhancing autonomous navigation in adverse weather with conditional diffusion synthetic data. In

2024 6th International Conference on Control and Robotics (ICCR), pages 63–68, 2024. 3

[27] Naufal Suryanto, Andro Aprila Adiputra, Ahmada Yusril Kadiptya, Thi-Thu-Huong Le, Derry Pratama, Yongsu Kim, and Howon Kim. Cityscape-adverse: Benchmarking robustness of semantic segmentation with realistic scene modifications via diffusion-based image editing. *IEEE Access*, 13: 69921–69940, 2025. 3

[28] Melissa Tijink, Stanislav Levendeev, Ewaldo Nieuwenhuis, Luuk Spreeuwers, Nicola Strisciuglio, and Estefanía Talavera. Vision on the move: Automated hazardous material plate detection in freight transport. In *Computer Analysis of Images and Patterns*, pages 259–270, Cham, 2026. Springer Nature Switzerland. 5

[29] M. Tremblay, S. S. Halder, and etc. de Charette, R. Rain rendering for evaluating and improving robustness to bad weather. *International Journal of Computer Vision*, 129: 341–360, 2021. 3

[30] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 2

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

[32] Yichun Wu, Huihuang Zhao, Wenhui Chen, Yunfei Yang, and Jiayi Bu. Textstyler: A clip-based approach to text-guided style transfer. *Computers & Graphics*, 119:103887, 2024. 1

[33] Peng Xing, Haofan Wang, Yanpeng Sun, wangqixun, Baixu, Hao Ai, Jen-Yuan Huang, and Zechao Li. CSGO: Content-style composition in text-to-image generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[34] Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Oliver Deussen, Weiming Dong, Jintao Li, and Tong-Yee Lee. Break-for-make: Modular low-rank adaptations for composable content-style customization, 2024. 2

[35] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 2

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 2, 3

[37] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[38] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10146–10156, 2023. 2

[39] Xian Zhong, Shidong Tu, Xianzheng Ma, Kui Jiang, Wenxin Huang, and Zheng Wang. Rainy wcity: A real rainfall dataset with diverse conditions for semantic driving scene understanding. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1743–1749. International Joint Conferences on Artificial Intelligence Organization, 2022. Main Track. 1

[40] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 3