TCWEATHERGEN: TARGET-CONSISTENT WEATHER AUGMENTATION FOR AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion editors can convert labeled clear-weather driving images into adverse-weather training data, but often deform task-critical targets such as traffic signs, lights, pedestrians, and vehicles, invalidating their annotations despite plausible weather appearance. Our analysis of FLUX.2 identifies three causes: aligned source attention, which carries a critical source constraint, weakens during denoising; background condition keys divert attention from target-specific support; and the global flow-matching loss gives small regions little spatial weight. We therefore introduce **TCWeatherGen**, which combines **SegRGBHeat** for sparse semantic anchoring, **Source–Condition LoRA (SC-LoRA)** for object-aware pathway adaptation, and **Content-Aware Token Pruning (CATP)** for suppressing uninformative condition tokens. On ACDC weather transfer, our fine-tuned 4B student improves target ROI SSIM from 0.560 to $\mathbf{0.727}$ and FID from 83.6 to $\mathbf{57.1}$ against matched FLUX.2-32B-dev pseudo-targets. CATP reduces peak training and inference memory by 10 and 3 GB and inference time by 35.8%. TCWeatherGen augmentation further improves downstream mIoU and mAP_{50} from 0.484 and 0.239 to $\mathbf{0.598}$ and $\mathbf{0.339}$.

1 INTRODUCTION

Autonomous-driving (AD) perception must handle snow, rain, fog, and low light, yet widely used labeled datasets are dominated by clear daytime scenes (Cordts et al., 2016; Sakaridis et al., 2021). Because collecting adverse-weather annotations at scale is costly, diffusion transformers (DiTs) offer strong generative priors for data augmentation (Peebles & Xie, 2023; Black Forest Labs, 2025). Such augmentation must satisfy two requirements simultaneously: global weather appearance should change, while task-critical targets such as traffic lights, signs, pedestrians, riders, and vehicles should preserve their identity, position, and geometry; otherwise, the reused annotations become invalid. We call this requirement ***target consistency***, which ordinary DiTs often violate.

Analysis in FLUX.2 klein-base-4B (Black Forest Labs, 2025) identifies three causes: an aligned-support intervention shows that noisy-image-to-source attention carries critical preservation evidence that weakens during denoising; background condition keys compete with target-specific evidence; and global flow matching underweights small annotated regions. These findings implicate attention-routing failures and spatially imbalanced supervision.

Based on these findings, we introduce **TCWeatherGen** (Figure 1), a target-consistent weather augmentation framework with three complementary designs. **SegRGBHeat** provides sparse source RGB anchors at task-critical locations. **Source-Condition LoRA (SC-LoRA)** independently adapts source and condition QKV pathways, while object-aware supervision increases the training emphasis on valid target regions. **Content-Aware Token Pruning (CATP)** removes low-signal background condition tokens that compete with target-specific support, while also shortening the attention sequence.

In the teacher–student setup, our 4B student model fine-tuned with FLUX.2-32B-dev pseudo-targets improves target-region SSIM from 0.560 to 0.727 and FID from 83.6 to 57.1 over the base model. CATP further reduces training and inference memory by 10 and 3 GB, and shortens inference time by approximately 35.8%. Downstream segmentation and detection models trained with TCWeatherGen outputs reach 0.598 mIoU and 0.339 mAP_{50} on real ACDC adverse-weather images, compared with 0.484 and 0.239 under base model augmentation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

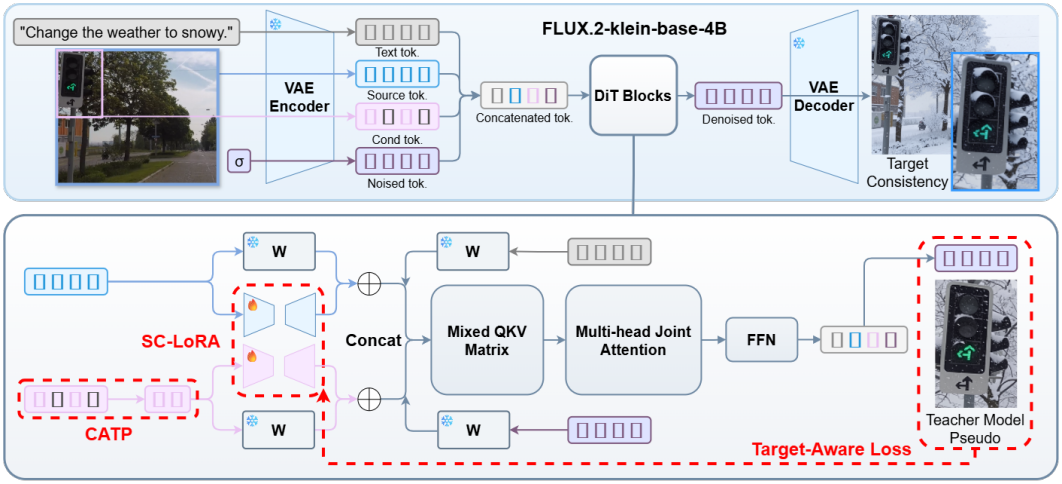

Figure 1: TCWeatherGen framework. Red frames denote our components. Snowflakes denote pretrained, and flames denote trainable modules. SegRGBHeat supplies aligned evidence, SC-LoRA sustains source and condition pathways, target-aware supervision emphasizes reliable regions, and CATP removes background condition keys.

Our contributions are summarized as follows:

- We diagnose target degradation in DiT weather editing through aligned-support intervention and denoising analysis, revealing source-route decay, condition-key competition, and area-dominated supervision.
- We introduce TCWeatherGen, which combines sparse semantic RGB anchors, independent source-condition adaptation, object-aware target supervision, and content-aware condition-token pruning.
- We show that TCWeatherGen improves target consistency, weather-transfer fidelity, and downstream perception performance, while CATP reduces computational cost.

2 Related Work

Diffusion-based weather augmentation for driving. Diffusion editors enable adverse-weather augmentation from labeled clear scenes (Brooks et al., 2023). ControlUDA generates cross-weather data for semantic segmentation (Shen et al., 2024), while SynDiff-AD synthesizes underrepresented conditions from semantic layouts (Goel et al., 2024). Cyclone learns from unpaired driving data through cycle consistency (Nguyen et al., 2026), and IntrinsicWeather edits material, geometry, and illumination factors (Zhu et al., 2026).

Fine-structure preservation in image translation. Latent compression makes fine structures particularly vulnerable. WarpI2I allocates more latent capacity to salient regions through reversible spatial warping (Zheng et al., 2026), while DrivePTS preserves foreground details through region-specific frequency supervision (Wang et al., 2026).

Condition control and efficient adaptation in DiTs. ControlNet and Uni-ControlNet inject edges, depth, and segmentation into U-Net diffusion models (Zhang et al., 2023; Zhao et al., 2023). For DiTs, OminiControl introduces image conditions as tokens in the pretrained model (Tan et al., 2025). EasyControl applies LoRA to the condition branch for efficient and flexible control (Zhang et al., 2025), whereas SC-LoRA decouples source preservation from condition control. OminiControl2 accelerates token-based conditioning through token pruning and compacting (Tan et al., 2026). Our CATP specializes token selection for SegRGBHeat support, removing background condition keys that compete with target-specific evidence during weather editing.

3 Analysis of Target Degradation

We identify three mechanisms that degrade task-critical targets during diffusion editing: decay of aligned source routing, leakage of condition attention into background keys, and area-dominated supervision.

3.1 Source Routing Carries Target Correspondence

The base model processes text, source-image, condition, and noisy-image tokens in joint attention. We refer to noisy-image queries within an annotated target region as ***target queries***, and examine their attention to the corresponding source region to assess how this attention contributes to target consistency.

Noisy-image–source support intervention experiment. Let $\mathcal{I}$ and $\mathcal{S}$ denote the noisy-image and source-image token sets, respectively. For an annotated target region $R \subset \mathcal{I}$, let $\Omega_R \subset \mathcal{S}$ be its spatially corresponding source support on the shared latent grid. Let L_{ij} denote the joint-attention logit from query i to key j. Before softmax, we remove all edges from queries in R to keys in Ω_R:

$$\widetilde{L}_{ij} = L_{ij} - \beta\,\mathbf{1}[i \in R,\, j \in \Omega_R], \qquad \text{with } \beta = +\infty. \tag{1}$$

The intervention is applied to every attention head, both single and double blocks, and all steps. Equal-cardinality controls instead close a translated support (Shifted support off) or random keys from $\mathcal{S} \setminus \Omega_R$ (Random support off).

The missing targets in Figure 2 show an explicit example in the annotated foreground under the aligned intervention. On 695 experimental frames, the average change from the base model output is 91.891 ROI RMSE for aligned support, compared with 1.544 and 1.603 for shifted and random support. This intervention validates that the aligned noisy-image-to-source route carries the source constraint needed to preserve annotated targets.



Figure 2: Example target-source support intervention with a fixed random seed 1642145648.

3.2 Progressive Loss of Source Constraint

At denoising step t, let $A_{ij}^{(t)}$ be the attention weight from query i to key j. For each target query $i \in R$, define its total source attention $s_i^{(t)} = \sum_{j \in \mathcal{S}} A_{ij}^{(t)}$ and aligned source attention $p_i^{(t)} = \sum_{j \in \Omega_R} A_{ij}^{(t)}$. We average these quantities over the target to obtain total source mass S_R, aligned source mass P_R, and the mean aligned-source attention ratio $\mathcal{C}_R$:

$$\big(S_R(t), P_R(t), \mathcal{C}_R(t)\big) = \frac{1}{|R|} \sum_{i \in R} \left(s_i^{(t)}, p_i^{(t)}, \frac{p_i^{(t)}}{s_i^{(t)}} \right). \tag{2}$$

Across 695 snow-transfer samples, S_R and P_R decrease by approximately 90% during the denoising, while $\mathcal{C}_R$ falls from 0.62 to 0.51 (Figure 3). Thus, the aligned source signal decays because target queries read less source content overall and allocate a smaller share of source attention to matching spatial support.

3.3 Condition-Attention Leakage into Background Keys

Condition tokens provide a second route for aligned evidence. Because target-support and background condition keys share the attention softmax, background attention can displace target-specific

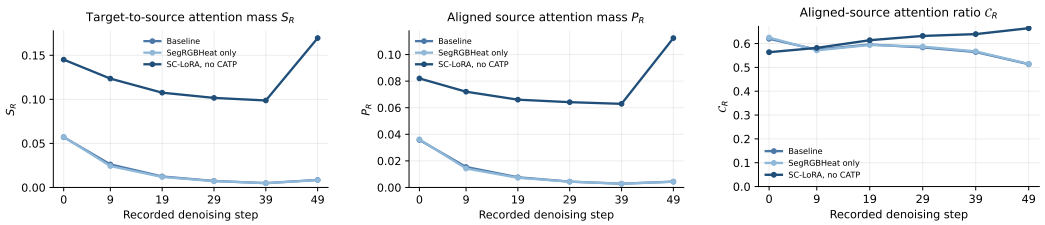

Figure 3: Source-routing metrics over 50 denoising steps in the last block, averaged over 8,212 targets from 695 snow frames.

support; we call this ***condition-attention leakage***, quantified by the fraction of condition attention from target queries that is assigned to the target instance:

$$Q_R^c(t) = \frac{1}{|R|} \sum_{i \in R} \frac{\sum_{j \in \mathcal{K}_R^c} A_{ij}^{(t)}}{\sum_{j \in \mathcal{K}^c} A_{ij}^{(t)}}. \tag{3}$$

where $\mathcal{K}^c$ the complete condition-key set and $\mathcal{K}_R^c$ denote those covering target R. On SegRGBHeat, we add background condition keys with the same condition reference progressively. 0, 445, and 890 background keys monotonically reduce Q_R^c from 53.78% to 42.06% and 34.97%, showing that background keys divert target queries from target-specific support.

3.4 Area-Dominated Training Supervision

Attention determines the evidence available to a target query, while the training objective determines its preservation incentive. Standard flow matching averages error over all N image tokens:

$$\mathcal{L}_{\mathrm{FM}} = \frac{1}{N} \sum_{i=1}^{N} \|v_\theta(z_t, t)_i - v_i\|^2 . \tag{4}$$

A target region R therefore receives only $|R|/N$ of the spatial weight, giving small targets no explicit preservation priority and motivating target-aware supervision.

4 TCWeatherGen

Given a clear source x_s, its annotation, and a weather prompt, TCWeatherGen transfers adverse weather while preserving the annotated targets (Figure 1).

4.1 SegRGBHeat: Sparse Target-Aware Condition

We construct SegRGBHeat from x_s and its Cityscapes annotations (Cordts et al., 2016) for traffic lights, traffic signs, people, riders, motorcycles, bicycles, cars, and poles. For their union mask $M \in \{0,1\}^{H \times W}$, we define

$$h = \max\big(M, \mathcal{G}_\sigma(M)\big), \qquad c = x_s \odot h, \tag{5}$$

where $\mathcal{G}_\sigma$ denotes Gaussian blur and $\odot$ broadcasts h across RGB channels. The maximum preserves target interiors while retaining a decaying context halo (Figure 4, left). After encoding c with the frozen image autoencoder, we append its tokens to the text, source, and noisy-image sequence. The asymmetric mask lets noisy-image queries access all four groups, restricts condition queries to condition keys, and blocks text and source queries from condition keys (Figure 4, right).

4.2 SC-LoRA and Target-Aware Supervision

Source and condition pathway adaptation. Section 3 shows that aligned source attention supports target preservation but weakens during denoising. SC-LoRA therefore assigns independent low-rank updates to the source and condition token spans for each QKV projection $p \in \{Q, K, V\}$:

$$W_p^s = W_p + \frac{\alpha}{r} B_p^s A_p^s, \qquad W_p^c = W_p + \frac{\alpha}{r} B_p^c A_p^c, \tag{6}$$



Figure 4: Left: a clear source and its SegRGBHeat condition. Right: the joint-attention mask; black crosses indicate blocked query–key interactions.

where the frozen W_p remains active for text and noisy-image tokens, while source and condition tokens use W_p^s and W_p^c. The adapters cover both attention-block types and all denoising steps, allowing the source branch to preserve aligned scene evidence and the condition branch to learn SegRGBHeat guidance. Compared with full fine-tuning, SC-LoRA trains adapters amounting to 0.26% of the base model's parameters.

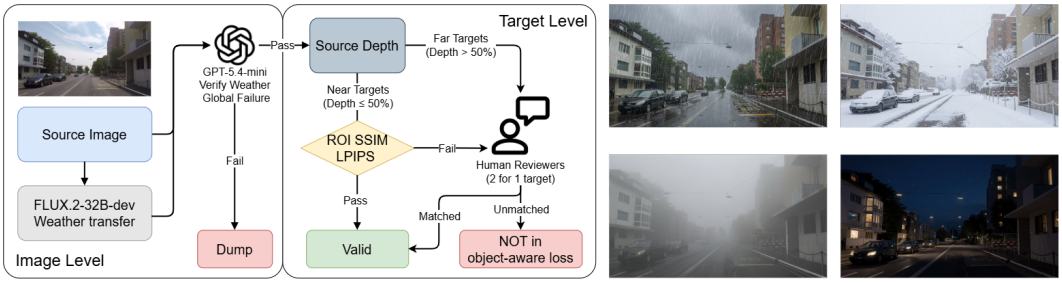

Figure 5: Two-level quality control for teacher-generated pairs and example outputs. GPT-5.4-mini rejects pairs that fail the requested weather globally. For admitted pairs, Depth Anything (Yang et al., 2024) divides annotated targets by source relative depth into near ($\leq 50\%$) and far ($> 50\%$). Near targets pass automatically if $\mathrm{SSIM}_i \geq \tau_{\mathrm{S}}$ ($\tau_{\mathrm{S}} = 0.3$) and $\mathrm{LPIPS}_i \leq \tau_{\mathrm{P}}$ ($\tau_{\mathrm{P}} = 0.7$); all others undergo paired-crop review by two reviewers.

Two-level pseudo-target quality control. A frozen FLUX.2-32B-dev teacher generates pseudo adverse-weather frames y for each annotated source x_s (Figure 5). Two-level screening checks both global weather fidelity and per-target annotation validity, with depth-aware routing allowing plausible visibility loss for distant targets. We set $q_i = 1$ if target i passes automatic or human review and 0 otherwise; only passing targets contribute to target-aware supervision.

Target-aware supervision. We dilate each image-space annotation mask with a 7×7 max filter and bilinearly project it onto the N-token latent grid, yielding $m_i \in [0,1]^N$ and the reliable-target union $m_q = \max_i(q_i m_i)$. For pseudo-target latent z_y, noise ϵ, and $t \in [0,1]$, the student predicts velocity $\hat{v}$ from $z_t = (1-t)z_y + t\epsilon$. Its target-aware flow-matching loss is

$$\mathcal{L}_{\mathrm{obj}} = \frac{\sum_j m_{q,j} \,\mathrm{mean}_c \,|\hat{v}_{j,c} - (\epsilon_{j,c} - z_{y,j,c})|}{\sum_j m_{q,j} + \varepsilon}, \tag{7}$$

where j and c index tokens and channels, and $\varepsilon = 10^{-6}$. The global reconstruction term $\mathcal{L}_{x_0}$ is the L1 error between $\hat{z}_0 = \epsilon - \hat{v}$ and z_y; $\mathcal{L}_{\mathrm{bg}}$ is the mean squared flow-matching error of the background. We optimize

$$\mathcal{L} = \lambda_{\mathrm{obj}} w(m_q) \mathcal{L}_{\mathrm{obj}} + \lambda_{x_0} \eta(t) \mathcal{L}_{x_0} + \lambda_{\mathrm{bg}} \mathcal{L}_{\mathrm{bg}}, \tag{8}$$

with $\lambda_{\mathrm{obj}} = 1.5$, $\lambda_{x_0} = 7.5$, and $\lambda_{\mathrm{bg}} = 1$. The area factor $w(m_q) = \mathrm{clip}(0.08/\rho_q, 1, 4)$ upweights sparse reliable targets, where $\rho_q = N^{-1} \sum_j m_{q,j} > 0$; if $\rho_q = 0$, we set $w = 1$ and $\mathcal{L}_{\mathrm{obj}} = 0$. We reduce the reconstruction weight at higher noise levels because the clean latent is less reliably recoverable from heavily corrupted inputs:

$$\eta(t) = (1-t) \begin{cases} 1, & 0 \leq t \leq 0.2, \\ 0.6, & 0.2 < t \leq 0.5, \\ 0.3, & 0.5 < t \leq 0.8, \\ 0, & 0.8 < t \leq 1. \end{cases} \tag{9}$$

During training, $t = \text{sigmoid}(\xi)$ with $\xi \sim \mathcal{N}(0, 1)$.

4.3 CATP: Content-Aware Condition Token Pruning

OminiControl2 shows that non-informative tokens can be removed from sparse conditions while retaining the spatial indices of the selected tokens (Tan et al., 2026).

Encoding turns the low-intensity SegRGBHeat background into condition keys that divert target queries from target-specific support (Section 3.3). CATP removes these keys before joint attention (Figure 6) by downsampling the condition to the $H/16 \times W/16$ patch grid and retaining token j when its patch maximum $u_j > \tau$, with $\tau = 5$. Retained tokens preserve their features and positional IDs, while source tokens remain unchanged. CATP thereby concentrates condition attention on aligned target evidence and shortens the sequence, reducing attention memory and runtime (Section 5.4).

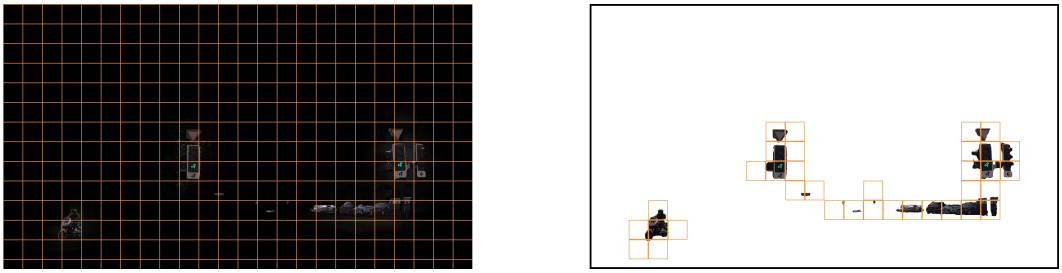

Figure 6: CATP prunes background tokens and retains positional IDs.

5 Experiments

5.1 Experimental Setup

Data and Results. We construct aligned training and evaluation pairs from ACDC clear scenes and Cityscapes-compatible annotations (Sakaridis et al., 2021; Cordts et al., 2016). FLUX.2-klein-base-4B is both the student model and baseline. Each weather uses 695 test images at 1536×864, with snow as the primary benchmark. All comparisons and ablations share teacher images, prompts, denoising schedules, and fixed seeds. For the resulting metrics, ROI SSIM is computed over annotated target regions; LPIPS, FID and KID are computed over the full image and are teacher-related.

Configurations. We fine-tune the student for 100 epochs using AdamW with a learning rate of 10^{-4} and weight decay of 10^{-3}. SC-LoRA uses rank 16 and scale 32. All training and inference are conducted on one NVIDIA H100 GPU.

5.2 Method-Mechanism Validation

SC-LoRA. Figure 7 (left) shows that, unlike the baseline and SegRGBHeat alone, which diffuse source attention across the whole scene, SC-LoRA concentrates responses around annotated small targets throughout denoising. At step 39, it raises S_R to 0.09871 and P_R to 0.06213, both over $20\times$ the baseline. (Figure 3)

Activating each branch from the same checkpoint shows that the source-only branch reduces condition attention to 0.0127 (-73%), while the condition-only branch reduces source attention to 0.0248 (-82%). Their dominant right-singular vectors yield an average orthogonality $\mathcal{O} = 1 - |\langle v_{\text{src}}, v_{\text{cond}} \rangle|$ of 0.90 across 25 DiT blocks, indicating that the SC-LoRA adapts each pathway independently: the source branch primarily maintains source routing, while the condition branch controls condition utilization.

CATP Suppresses Condition-Attention Leakage. Figure 7 (right) shows that SC-LoRA strengthens target responses and CATP further suppresses background responses by removing their

condition keys. The query-side target concentration D_R^c rises from 66% to 77% at step 9 and from 68% to 86% at step 49. While D_R^c identifies where the condition route originates, the key-side Q_R^c in Eq. equation 3 measures whether target queries read matching or background keys.

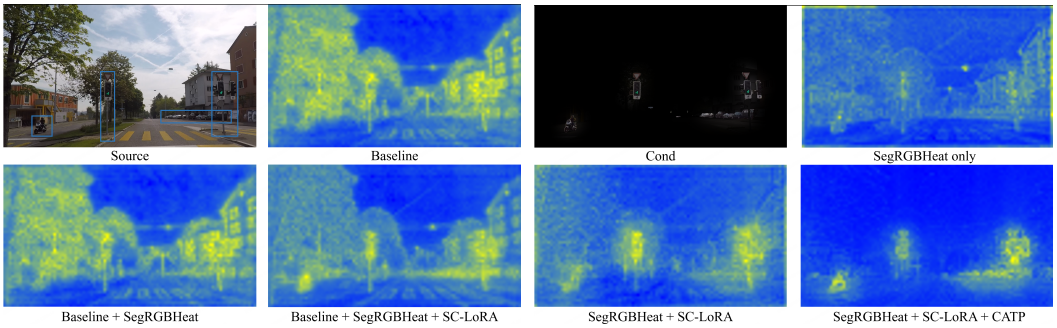

Figure 7: Query-side attention in the last single-stream block. On the 54×96 noisy-image token grid, each location sums attention to all source keys (left) or condition keys (right); blue and yellow indicate low and high values after per-map normalization. Averaging the raw source map over R gives S_R; SC-LoRA concentrates source routing around targets, and CATP concentrates condition routing there.



Figure 8: Qualitative comparison across fog, rain, snow, and night. Columns show the clear source, the base 4B editor, FLUX.1-kontext-dev, IntrinsicWeather, Qwen-Image-Edit-2509, and TCWeatherGen. Blue boxes mark task-critical targets, with insets enlarged.

5.3 Target Consistency

Table 1 compares TCWeatherGen with external image-editing models under the same teacher-relative snow protocol. The full system therefore improves both target-region structure and image-level agreement with the 32B teacher. Across 695 frames, TCWeatherGen raises teacher-relative target-object ROI SSIM from 0.560 to 0.727 and reduces FID from 83.6 to 57.1 relative to the baseline. LPIPS improves from 0.516 to 0.413. Appendix **??** reports more quantitative results on rain, fog, and night, where TCWeatherGen still shows consistent improvements in target consistency.

Figure 8 shows comparisons among the base editor, three external editors, and TCWeatherGen. All weather examples show the same trend: TCWeatherGen preserves the identity and local geometry of task-critical targets while applying the requested weather appearance. The difference is especially

Table 1: External comparison on 695 ACDC frames under the unified teacher-relative protocol. **Bold** marks the best result.

Method	Para.	ROI SSIM ↑	LPIPS ↓	FID ↓	KID ↓
Baseline	4B	0.560	0.516	83.6	0.0415
+ Vanilla LoRA	4B+0.13%	0.660	0.473	66.1	0.0339
FLUX.1-kontext-dev	12B	0.267	0.643	77.97	0.021
Qwen-Image-Edit-2509	20B	0.644	0.478	72.1	0.0211
IntrinsicWeather	2.5B	0.487	0.572	81.5	0.016
TCWeatherGen	4B+0.26%	**0.727**	**0.413**	**57.1**	**0.0042**

visible in small and highly informative targets: our result keeps the left-prohibit sign, green-arrow signals, and remote pedestrians recognizable, whereas competing editors blur or distort these structures.

5.4 Ablation Studies and Efficiency

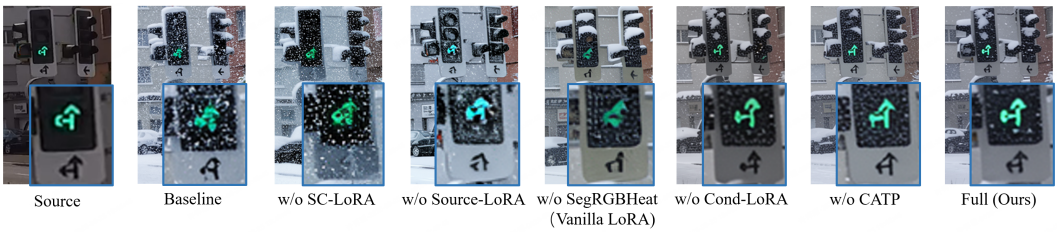

Figure 9: Snow ablation on one scene. Insets enlarge a traffic light example.

Table 2: Snow ablation on 695 images. *Italics* indicate results without this module. **Bold** marks the best results.

Metric	Baseline	*w/o SC-LoRA*	*w/o Source-LoRA*	*w/o SegRGBHeat*	*w/o Cond-LoRA*	*w/o CATP*	Full
FID ↓	83.6	78.8	82.9	70.6	66.1	59.1	**57.1**
ROI SSIM ↑	0.560	0.533	0.539	0.648	0.660	0.725	**0.727**
LPIPS ↓	0.516	0.479	0.482	0.446	0.421	0.416	**0.413**

SC-LoRA is the main contributor: removing it drops ROI SSIM from 0.727 to 0.533 and raises FID from 57.1 to 78.8 (Table 2). Without SegRGBHeat, source-only Vanilla LoRA improves over the baseline (0.648 vs. 0.560 ROI SSIM) but remains below the full model, showing that one adapted pathway is insufficient. In Figure 9, removing Source-LoRA causes the strongest distortion, whereas removing Cond-LoRA largely preserves the symbol topology; the former uses an EasyControl-style condition-only LoRA allocation (Zhang et al., 2025), here applied to SegRGBHeat, and their contrast supports separate source and condition adapters. CATP modestly improves ROI SSIM/FID from $0.725/59.1$ to $0.727/57.1$. Its larger measured effect is computational, reducing peak training memory from 75 to 65 GB, inference memory from 18 to 15 GB, and single-image inference time by 35.8%.

5.5 Downstream Augmentation

We compare clear-only training (A), baseline augmentation (B), and TCWeatherGen augmentation (C). The downstream models are a COCO-pretrained DeepLabV3-ResNet50 with a 19-class head and a COCO-pretrained Faster R-CNN-ResNet50-FPN with a four-class head. Both are trained for 30 epochs using AdamW and learning rates of 10^{-3} and 10^{-4}, respectively. We report mIoU over traffic light, traffic sign, person, and car; detection boxes are derived from aligned semantic masks. A is trained on 2,000 ACDC clear-weather images, with B and C on 1,305 clear-weather and 695 generated adverse-weather images, which are produced only for the training split. Both tasks are evaluated on 800 real ACDC adverse-weather frames, 200 per weather. GT in validation comes from the ACDC dataset.

Table 3: Downstream augmentation. Segmentation reports pixel-level IoU for four task-critical classes and their mean; detection reports bounding-box mAP over the same classes at IoU thresholds 0.50 and 0.75. **Bold** marks the best result.

	Segmentation (pixel IoU)					Detection (box AP)	
Training group	$\text{mIoU}_4 \uparrow$	Traffic light	Traffic sign	Person	Car	$\text{mAP}_{50} \uparrow$	$\text{mAP}_{75} \uparrow$
A (clear only)	0.350	0.366	0.403	0.256	0.375	0.235	0.059
B (+baseline)	0.484	0.516	0.488	0.383	0.551	0.239	0.056
C (+TCWeatherGen)	**0.598**	**0.610**	**0.522**	**0.489**	**0.772**	**0.339**	**0.120**

Table 3 shows that TCWeatherGen augmentation consistently improves both segmentation and detection on the real adverse-weather set, increasing four-class mIoU to 0.598 and mAP_{50} to 0.339. All task-critical segmentation classes also improve over both baselines, indicating gains in both pixel-level recognition and object localization.(Figure 10)

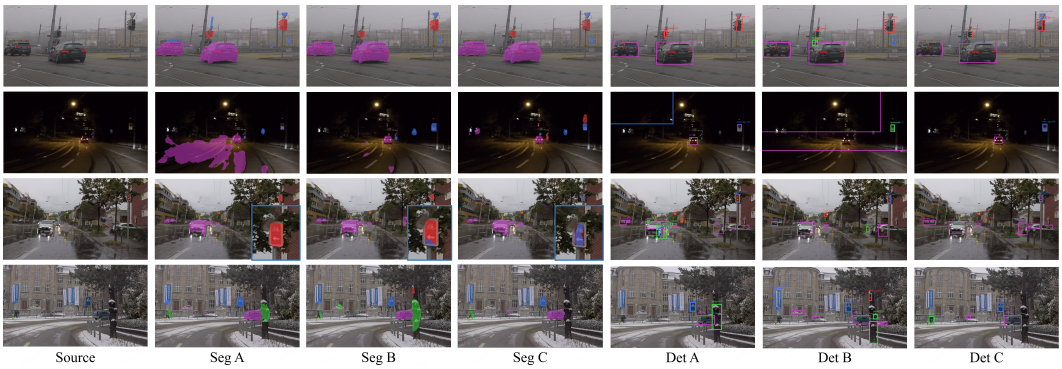

Figure 10: Downstream segmentation and detection. B and C models are trained on generated adverse-weather frames and evaluated on real ones.

6 Conclusion and Limitations

We study small-target drift during weather augmentation and propose TCWeatherGen. It addresses weakening source evidence, condition-token competition, and area-dominated supervision through SegRGBHeat, SC-LoRA, and CATP. It improves target consistency and downstream adverse-weather perception while maintaining global quality, demonstrating the value of explicit target preservation in diffusion augmentation. The current design requires task annotations to construct SegRGBHeat and target-aware supervision, limiting direct use with unlabeled source data. Moreover, training and most evaluation rely on FLUX.2-32B-dev pseudo-targets in ACDC/Cityscapes-style scenes, potentially inheriting teacher bias and leaving generalization to real extreme weather, other sensors, and broader driving domains unverified

References

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

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)*, pp. 18392–18402, 2023.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Harsh Goel, Sai Shankar Narasimhan, Oguzhan Akcin, and Sandeep Chinchali. Syndiff-ad: Improving semantic segmentation and end-to-end autonomous driving with synthetic data from latent diffusion models. *arXiv preprint arXiv:2411.16776*, 2024.

Thang-Anh-Quan Nguyen, Moussab Bennehar, Luis Guillermo Roldao Jimenez, Nathan Piasco, Dzmitry Tsishkou, Laurent Caraffa, Jean-Philippe Tarel, and Roland Brémond. Cyclone: Diffusion model for cycle-consistent weather editing from unpaired driving data. *arXiv preprint arXiv:2607.13927*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Fengyi Shen, Li Zhou, Kagan Kucukaytekin, Ziyuan Liu, He Wang, and Alois Knoll. Controluda: Controllable diffusion-assisted unsupervised domain adaptation for cross-weather semantic segmentation. *arXiv preprint arXiv:2402.06446*, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14940–14950, 2025.

Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. Ominicontrol2: Efficient conditioning for diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4256–4265, 2026.

Zhechao Wang, Yiming Zeng, Lufan Ma, Zeqing Fu, Chen Bai, Dongshuo Yin, Ziyao Lin, and Cheng Lu. Drivepts: A progressive learning framework with textual and structural enhancement for driving scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3699–3708, 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381, 2024.

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)*, pp. 3836–3847, 2023.

Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19513–19524, 2025.

Shilong Zhao, Dongdong Chen, Yen-Chun Chen, Jun Bao, Shaozhe Hao, Lu Yuan, and K. Y. K. Wong. Unicontrolnet: All-in-one control to text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 11127–11150, 2023.

Shen Zheng, Anurag Ghosh, Gaurav Parmar, and Srinivasa Narasimhan. Warpi2i: Image warping for image-to-image translation. *arXiv preprint arXiv:2606.31018*, 2026.

Yixin Zhu, Zuo-Liang Zhu, Jian Yang, Miloš Hašan, Jin Xie, and Beibei Wang. Intrinsicweather: Controllable weather editing in intrinsic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30772–30781, 2026.

AI USE STATEMENT

Adverse data generation. We used generative AI to construct and screen the pseudo-target image pairs employed by TCWeatherGen. FLUX.2-32B-dev generated the adverse-weather pseudo-targets from clear source images, and GPT-5.4-mini performed image-level screening for the requested weather condition and obvious global generation failures. These uses form part of the method described in Section 4.2.

Research and manuscript assistance. We used OpenAI Codex to refine research hypotheses, express implemented losses and attention diagnostics mathematically, assist with quantitative result interpretation, search and summarize literature, translate and edit manuscript text, format tables, and arrange existing experimental figures. Codex also assisted with auxiliary code and scripts for analysis, metric aggregation, manuscript preparation, and consistency checks. The authors manually reviewed the AI-assisted text, code, analyses, citations, figures, and reported values against the implementation and experimental records. The authors made the final research and writing decisions and take responsibility for the complete content of this work.

A Additional Weather-Transfer Results

A.1 Results Across Weather Conditions

Table A.1 reports the final model across snow, rain, fog, and night. Teacher-relative ROI SSIM measures agreement with the intended adverse-weather output, while source-relative ROI SSIM measures retention of the clear-scene target structure. The two references capture the appearance-change and label-preservation requirements of weather augmentation.

Table A.1: Target consistency of TCWeatherGen across weather conditions.

Weather	Frames	ROI SSIM vs. teacher ↑	FID vs. teacher ↓	LPIPS vs. teacher ↓
Snow	695	0.727	57.1	0.421
Rain	695	0.639	65.5	0.465
Fog	695	0.837	73.0	0.320
Night	695	0.392	78.6	0.433

A.2 Depth-Dependent Fog Plausibility

We compute the depth-related fog diagnostics using the depth map estimated from each clear source and the corresponding appearance measurements from its generated fog image. This protocol tests whether contrast, sharpness, saturation, and haze vary with source-scene depth in the expected direction; it does not estimate depth from the generated fog image. TCWeatherGen improves the fitted depth trends over the base model across the reported diagnostics (Table A.2). This is an additional empirical finding; the present experiments do not isolate its cause.

Table A.2: Fog physical-plausibility diagnostics on clear-source test frames on 695 fog transfers. Metrics use annotated task-critical regions; arrows indicate the preferred direction.

Metric	Baseline	Ours
Scattering-fit R^2 ↑	0.445	**0.480**
Depth–contrast R^2 ↑, slope ↓	0.761, -3.563	**0.847, -5.036**
Depth–contrast R^2 ↑, monotonicity ↑	0.761, 0.784	**0.847, 0.831**
Depth–sharpness Spearman ↓, slope ↓, R^2 ↑	-0.339, -0.014, 0.473	**-0.544, -0.057, 0.536**
Depth–saturation R^2 ↑, slope ↓, monotonicity ↑	0.603, -0.132, 0.681	**0.627, -0.155, 0.711**
Haze–depth Spearman ↑	0.532	**0.566**

B Pseudo-Target Construction and Training Details

B.1 Two-Level Pseudo-Target Quality Control

The image-level verifier receives the full teacher-generated frame together with the requested weather condition. GPT-5.4-mini accepts generations that exhibit the requested weather and rejects pairs with an incorrect condition or an obvious global failure. This stage operates on the complete image and does not assess pixel-level agreement inside small target regions. A rejected pair is removed before training.

For each admitted pair, Depth Anything (Yang et al., 2024) estimates relative depth from the clear source. Annotated targets with normalized source depth at or below 50% enter the near group, while the remaining targets enter the far group. The grouping uses relative depth and does not require metric reconstruction. For a near target, we compare the source and pseudo-target bounding-box crops using SSIM and LPIPS. The target passes automatically when

$$\mathrm{SSIM}_i \geq \tau_{\mathrm{S}} \quad \text{and} \quad \mathrm{LPIPS}_i \leq \tau_{\mathrm{P}}. \tag{B.1}$$

Near targets that fail or remain ambiguous, together with all far targets, undergo paired-crop review by two reviewers. The review asks whether the matched target remains structurally valid so that its

Table C.1: Query-side target concentration D_R^c: the fraction of all condition-route attention whose noisy-image query lies in an annotated target region. Steps 10 and 50 use one-based indexing and correspond to steps 9 and 49 in the main text.

Configuration	Step 10		Step 50	
	No CATP	+CATP	No CATP	+CATP
+Cond+SC-LoRA	66%	**77%**	68%	**86%**

original annotation is usable. This depth-aware routing prevents expected visibility loss in distant fog or night regions from being classified as structural corruption.

The resulting reliability variable is target-specific:

$$q_i = \begin{cases} 1, & \text{target } i \text{ passes automatic screening or human review,} \\ 0, & \text{otherwise.} \end{cases} \tag{B.2}$$

A target with $q_i = 0$ is excluded from $\mathcal{L}_{\text{obj}}$, while other reliable targets in the same frame remain active. The image pair still contributes to the global reconstruction and background objectives. Thus, image-level verification controls weather validity, and target-level verification controls which annotations supervise the object-aware term.

B.2 Optimization and Evaluation

The student model is initialized from FLUX.2-klein-base-4B, and the pseudo-target teacher is FLUX.2-32B-dev. We train for 100 epochs at 1536×864 resolution using AdamW with learning rate 10^{-4} and weight decay 10^{-3}. SC-LoRA uses rank 16 and scale 32, and CATP uses $\tau = 5$ on the 0–255 condition-intensity scale. Training uses one NVIDIA H100 GPU. The timestep distribution and all loss weights are specified in Section 4.2.

ROI SSIM measures teacher-relative similarity within annotated target regions; source-relative ROI SSIM separately measures structural retention from the clear input. LPIPS is computed over the full image with a VGG backbone. FID and KID use Inception-v3 pool3 features and compare the generated and matched teacher-image distributions. Methods within each comparison use the same source frames, teacher images, prompts, schedules, and random seeds.

C Additional Attention and Efficiency Diagnostics

C.1 Query-Side and Key-Side Condition Routing

The condition heatmaps in Figure 7 locate where the condition pathway is used. We downsample the noisy image to its 54×96 token grid and, for each noisy-image query, sum its attention over all condition keys. The resulting query-side map is projected back to image coordinates for visualization. Blue and yellow denote low and high values after per-map normalization; all reported numerical comparisons use the unnormalized attention weights.

For the union R of annotated target-query locations and noisy-image token set $\mathcal{I}$, the query-side concentration is

$$D_R^c(t) = \frac{\sum_{i \in R} \sum_{j \in \mathcal{K}^c} A_{ij}^{(t)}}{\sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{K}^c} A_{ij}^{(t)}}. \tag{C.1}$$

D_R^c measures the fraction of condition-route attention originating from target queries. The key-side quantity Q_R^c in Eq. equation 3 instead measures how much of a target query's condition attention reaches its matching target keys. These quantities describe complementary sides of condition-attention leakage.

The controlled key-addition experiment in Table C.2 keeps the target-support keys fixed and progressively restores background keys from the same condition. The monotonic decrease in Q_R^c isolates competition within the condition route.

Table C.2: Final-step target-instance condition share Q_R^c on 180 instances as background keys are added.

Added background keys	0	222	445	668	890
Target-instance condition share Q_R^c (%)	53.784	45.814	42.061	38.131	34.972

C.2 SC-LoRA Branch Specialization

We activate one SC-LoRA branch at a time from the same trained checkpoint. Disabling the condition branch reduces condition attention by 73%, while disabling the source branch reduces source attention by 82% (Table C.3). Across 25 DiT blocks, the dominant right-singular directions of the source and condition updates have mean orthogonality $1 - |\langle v_{\mathrm{src}}, v_{\mathrm{cond}} \rangle| = 0.90$. The two branches therefore learn distinct updates for source preservation and condition utilization.

Table C.3: Attention routing under single-branch activation (small-object region, average over 2 images $\times$ 4 steps).

Configuration	mean source attn	mean cond attn	Note
Both branches	0.1347	0.0464	base model
Source branch only	0.1314	0.0127	cond attn -73%
Cond branch only	0.0248	0.0386	source attn -82%

C.3 CATP Threshold and Token Retention

Table C.4 reports the threshold sweep. At the selected $\tau = 5$, CATP retains 14.8% of condition tokens and all task-critical tokens in this diagnostic, while retaining 10.4% of background tokens. Raising the threshold to 20 removes more background but begins to discard target support.

Table C.4: Threshold τ sweep.

τ	Retention ratio	Small-object retention	Background retention
2	19.6%	100.0%	15.6%
5	14.8%	100.0%	10.4%
10	10.8%	99.9%	6.3%
20	7.0%	91.4%	2.7%

CATP's primary effect is the condition-routing change quantified above. Shortening the condition sequence also reduces inference cost. Under a controlled 659-image snow profile with 50 denoising steps, CATP lowers peak memory of 18.35 to 15.32 GB, and runtime from 420.2 to 269.6 seconds per image.