-supplementary Material-

# Seed-to-Seed: Unpaired Image Translation in Diffusion Seed Space

## A Deterministic DDIM

Early denoising diffusion and score-based generative models [7, 19, 21] sample seeds from white Gaussian noise and progressively map them to images using a stochastic sampling process. Denoising Diffusion Implicit Models (DDIM) [20] offer a generalization which enables deterministic sampling. In addition to reducing the required number of sampling steps, the DDIM process lends itself to inversion [5, 20], making it possible to map images back to the seed-space. Inversion is crucial for the ability to edit real images using pre-trained diffusion models [14, 16, 22, 24]. The *deterministic* DDIM sampling process that denoises the current sample $x_t$ to yield the next step $x_{t-1}$ can be formulated as:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \cdot \hat{x}_0 + \sqrt{1-\alpha_{t-1}} \cdot \varepsilon_\theta^t(x_t) \tag{1}$$

where $\hat{x}_0$ is a prediction of the final denoised sample $x_0$ from $x_t$, given by:

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\alpha_t} \cdot \varepsilon_\theta^t(x_t)}{\sqrt{\alpha_t}}. \tag{2}$$

Here $\alpha_{t-1}, \alpha_t$ are the per-timestep diffusion schedule hyperparameters, and $\varepsilon_\theta^t$ is the noise prediction U-net, parameterized by $\theta$.

The reverse process, referred to as DDIM inversion, is formulated as follows (at the limit of decreasing step size):

$$x_{t+1} = \sqrt{\alpha_{t+1}} \cdot \hat{x}_0 + \sqrt{1-\alpha_{t+1}} \cdot \varepsilon_\theta^t(x_t) \tag{3}$$

## B Implementation Details

In this section, we provide detailed implementation and training information regarding the different models that were trained during this work.

### B.1 Seed-Space Classifier

We use a uniform *ResNet18*-based classifier for all classification tasks presented in Table 1 in the main text. For the tasks applied within the seed-space, we adjust the first layer of the classifier to 4-channeled input to fit the dimensionality of Stable Diffusion's latent representation. We split the training data to 80% for training and 20% for validation. We trained all models to a maximum of 80 epochs over the training set, and chose the best accuracy over the validation. We use the Adam optimizer [10] with lr=0.001 for all tasks. Task-specific details are provided below:

| | |
|---|---|
| **Weather** | rainy, snowy, clear, overcast, undefined, partly cloudy, foggy |
| **Time-Of-Day** | daytime, night, dawn/dusk, undefined |

Table 1: Attributes and corresponding options provided in BDD100k metadata logs.

- **day\night:** We trained over *BDD100k* “daytime” and “night” splits. For the seed-space version, we first center-cropped each sample to 512 × 512, then inverted them to the seed space.

- **cat\dog:** We trained over the “cat” and “dog” splits of the *AFHQ* [3] dataset without additional preproccessing.

- **older\younger:** We used the provided metadata of the *FFHQ* [9] dataset and chose samples tagged as 55+ years old as the “older” split and those tagged in the range of 17-40 years old as the “younger” split. We used the 512 × 512 version of the dataset.

## B.2 Finetuning SD for Automotive Dataset

The pre-trained version of Stable Diffusion (SD) 2.1 performs poorly on realistic driving datasets. As a result, we fine-tune SD 2.1 using the *BDD100k* training set. We automatically generate the textual conditions using information provided in the dataset’s metadata logs regarding Weather and Time-Of-Day. The resulting prompts have the form:

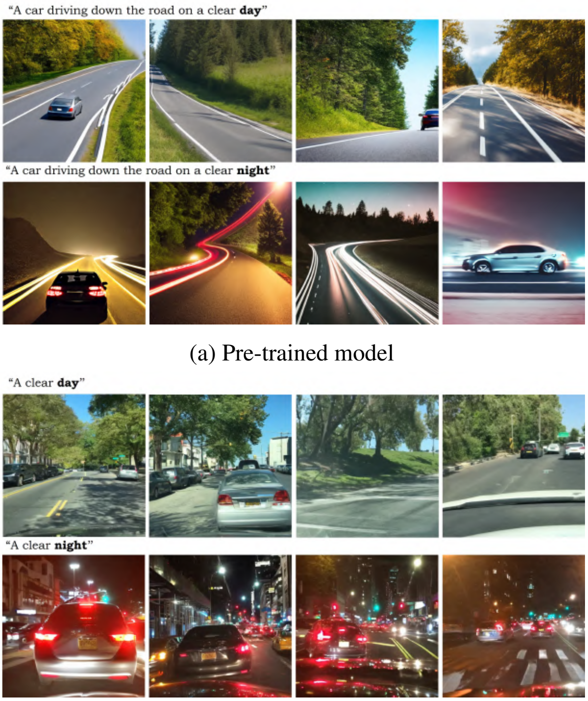
> “A **Weather** **Time-Of-Day**”

The various choices available in the metadata logs of *BDD100k* for individual attributes are delineated in Table 1. It should be noted that all images featuring an “undefined” label for any attribute have been excluded from the training set. “dawn/dusk” images were also excluded due to low amount of samples and unclear thresholds between “dawn/dusk” and “daytime/night”.

Some synthetic images before and after fine tuning are illustrated in Figure 1.

We train a ControlNet over our fine-tuned SD using the same dataset. We utilize a Canny-like spatial control, derived by applying a Canny edge detector over a segmentation mask obtained using the publicly available version of the Segment-Anything Model [11] (SAM). This approach ensures that only the boundaries of each object and sub-object are considered. Through experimentation, we found this spatial control to be superior to using Canny directly with different thresholds or a direct SAM mask. Some controlled synthetic images before and after fine tuning are illustrated in Figure 2.

Figure 3 shows a qualitative comparison between our method and additional diffusion-based baselines for the day-to-night translation task. Specifically, it extends the comparison shown in Figure 7b in the main text by including two additional methods, T2I-Adapters [15] and InstructPix2Pix [2]. These methods involve a training process specific to them, and could not utilize our fine-tuned U-net.


"A car driving down the road on a clear **day**"
"A car driving down the road on a clear **night**"
"A clear **day**"
"A clear **night**"

(a) Pre-trained model

(b) Fine-Tuned model

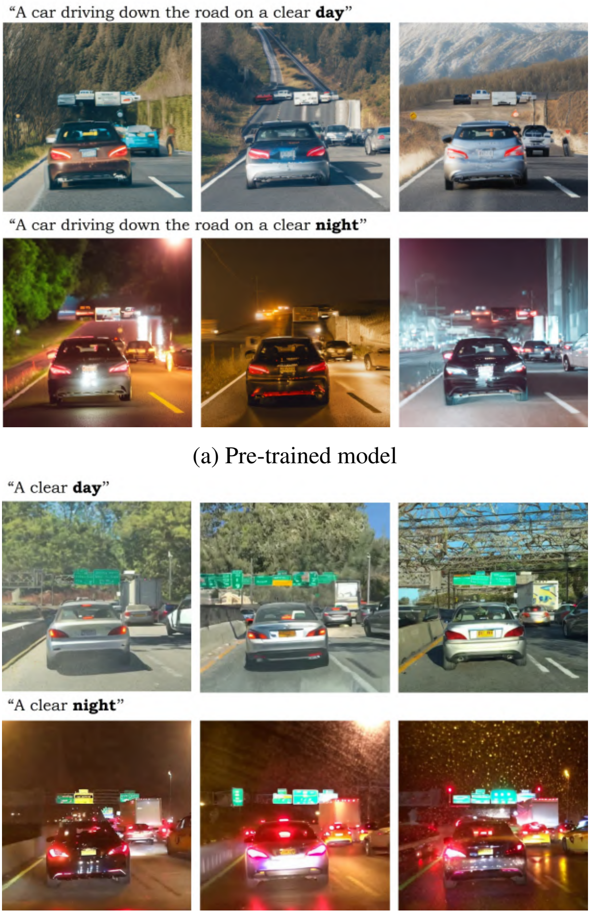
Figure 1: Fine-tuning SD 2.1 for automotive images using the BDD100k dataset: (a) before, and (b) after.

| CFG-scale | FID ↓ | MMD ↓ | KID ↓ | SSIM ↑ |
|---|---|---|---|---|
| 1.0 | 25.955 | 67.404 | 5.364 | 0.549 |
| 3.0 | 17.454 | 45.540 | 3.988 | 0.526 |
| 5.0 | 16.384 | 41.344 | 3.718 | 0.505 |

Table 2: **CFG Ablation study.** Balance between content preservation and target domain appearance via CFG-scale.

| Method | FID ↓ | MMD ↓ | KID ↓ | SSIM ↑ |
|---|---|---|---|---|
| ControlNet+RS | 35.091 (+114%) | 95.171 (+130%) | 6.340 (+70%) | 0.493 (+2%) |
| ControlNet+Inv | 49.572 (+202%) | 411.060 (+894%) | 14.981 (+296%) | **0.756 (-49%)** |
| ControlNet+RS+ST | 21.316 (+30%) | 67.650 (+63%) | 5.5456 (+49%) | 0.450 (+11%) |
| ControlNet+Inv+ST (*StS*) | **16.384** | **41.344** | **3.718** | 0.505 |

Table 3: **Ablation study - Model Components.** Day-to-Night translation over *BDD100k*.


“A car driving down the road on a clear **day**”
“A car driving down the road on a clear **night**”
“A clear **day**”
“A clear **night**”

(a) Pre-trained model

(b) Fine-Tuned model

Figure 2: Trained ControlNet over fine-tuned SD 2.1 for automotive, vs. pretrained ControlNet from [25].

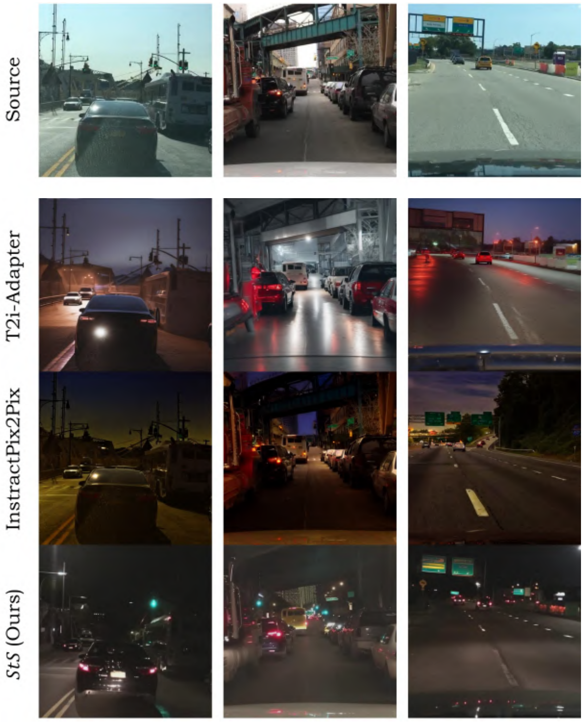


Figure 3: A qualitative comparison between our method and existing diffusion-based methods for day-to-night translation. This figure is an extension of Figure 7b in the main text, that includes two additional diffusion-based methods (T2I-Adapters [15] and InstructPix2Pix [2]), each of which involves a specific training process and could not utilize our fine-tuned U-net.

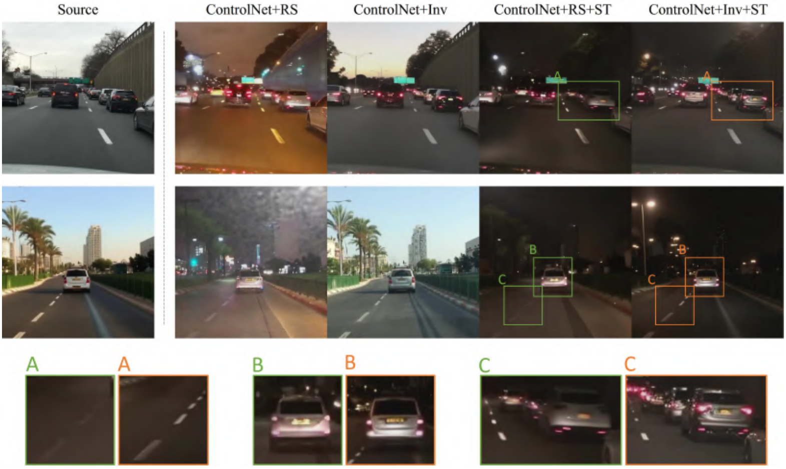


Figure 4: **Qualitative Ablation Study.** Day-to-Night translation over *BDD100k*. All samples were generated using the prompt “A clear night”.

# C Ablation Study

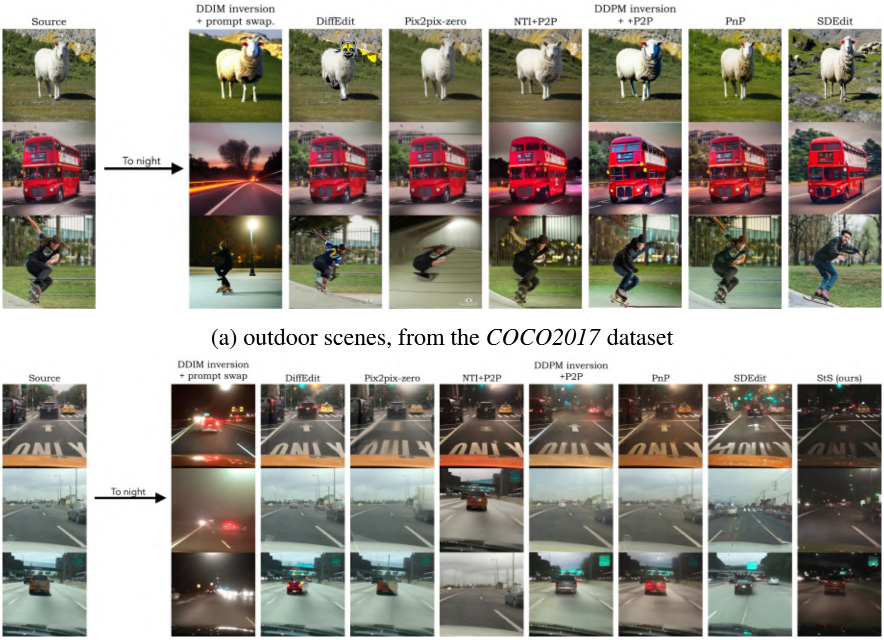
We examine the effectiveness of the different components of our approach by incorporating them one at a time (see Table 3). On top of a pure off-the-shelf ControlNet initiated with a random seed (RS) $z_T \sim \mathcal{N}(0, I)$ (*ControlNet+RS*), we add two components that make up the complete Seed Translation (ST) block: (1) initiation with a meaningful seed obtained by DDIM inversion (*ControlNet+Inv*) and (2) using the *sts-GAN* in the seed space for the seed translation (*ControlNet+Inv+ST*). We also demonstrate the contribution of the *sts-GAN* block applied over a random seed, rather than an inverted one (*ControlNet+RS+ST*). The contribution of each component is reported in Table 3 and qualitatively illustrated in Figure 4. In all configurations we used the whole *BDD100k* test set with 20 DDIM steps and CFG-scale $\omega = 5.0$.

Notably, initiating the ControlNet sampling process with the inverted seed (*ControlNet+Inv*) imposes an overly rigid constraint, significantly reducing editability during sampling. As proposed in the introduction, the semantic information encoded within the inverted seed includes daytime-related attributes, which conflict with the desired appearance indicated by the textual guidance, resulting in low editability. This challenge is reflected in the combination of a very high SSIM score alongside poor appearance metrics.

In the second configuration (*ControlNet+RS+ST*) we randomly sampled an initial seed, then translated it into a target-related one using our *sts-GAN*. The translated seed is subsequently sampled using ControlNet. Here, the seed that initiates the sampling process unconditionally possesses attributes of the target domain, hence results with more accurate appearance of the output image in the target domain, which is reflected in better appearance measures compared to *ControlNet+RS*. Yet, the generated images tend to over-blurriness, reflected by a lower SSIM score.

Combining the ST block with an initially meaningful inverted seed (*ControlNet+Inv+ST* (*StS*)) yields the desired combination of best appearance and high fidelity to the structure. The details are sharper and more accurate, compared to the *ControlNet+RS+ST* configuration (as illustrated by crops *A, B* and *C* in Figure 4), which lead to an improvement of 11% in the SSIM score. Nonetheless, the images generated by our *StS* model commonly exhibit more accurate and semantically meaningful *local* target-related effects. When a random seed is being translated to a target-domain-related seed using *sts-GAN*, the translated seed encodes information about the global appearance of the target domain but lacks details about the local semantics of the source image. Consequently, local, semantics-related effects are better generated using the translated-inverted seed than a random-translated one. This phenomenon is evident in features such as head/tail lights, street lights, reflections, etc., and is quantitatively demonstrated by superior performance in target-appearance metrics. This ablation provides insight into both the contribution of the *sts-GAN* and what it has actually learned.

As mentioned, the spatial control stabilizes the loss of details caused by the CFG mechanism. While contributing to the target-domain appearance of the output (see Figure 3 in the main text), it somewhat reduces the structural preservation, as quantitatively evaluated in Table 2.



(a) outdoor scenes, from the *COCO2017* dataset

(b) automotive scenes, from the *BDD100K* dataset

Figure 5: Comparison of different DM-based methods on the global translation of day-to-night

# D Additional DM-based baselines

As detailed in the related work in the main text, we chose to evaluate our method against DM-based methods that impose global constraints on the source image and therefore allow global edits while maintaining adherence to the source image. In this section, we demonstrate the limitations of locally constrained and non-constrained methods in such edits.

Locally constrained methods, such as Prompt-to-Prompt (P2P) [6], Pix2Pix-Zero [16], and DiffEdit [4], use local spatial constraints in the form of attention maps or dynamically generated masks to focus edits on specific regions of the image while leaving the rest mostly unchanged. Figure 5a demonstrates global edits (day-to-night) on outdoor samples from the *COCO2017* dataset [12] using non-constrained DDIM-inversion with prompt swapping, locally constrained methods including P2P [6] with NTI [14] and DDPM-Inversion [8], pix2pix-zero [16]and DiffEdit [4] as well as globally-constrained methods: PnP [22] and SDEdit [13]. All methods use SD2.1 with the default hyperparameters provided by their respective authors. Notably, non-constrained methods (DDIM inversion + prompt swap) fail to preserve the details of the source image. Both Locally and globally constrained methods struggle to achieve global edits while maintaining source image details. In most cases, the target domain (night) is almost not reflected in the translated images.

Figure 5b demonstrated the performance of the aforementioned methods on automotive scenes from the *BDD100k* dataset using the fine-tuned Unet (see Section B.2.)

# E Additional Examples

Figure 6 presents more examples of different weather translation performed over the *BDD100k* dataset.

# F Non-Automotive Applications

Though focused on automotive-related translations, our model is also suitable for additional diverse translation tasks on other datasets. In this section we provide some examples of non-automotive translations using our proposed *sts-GAN*, followed by a short discussion of possible limitations.

## F.1 Non-automotive examples

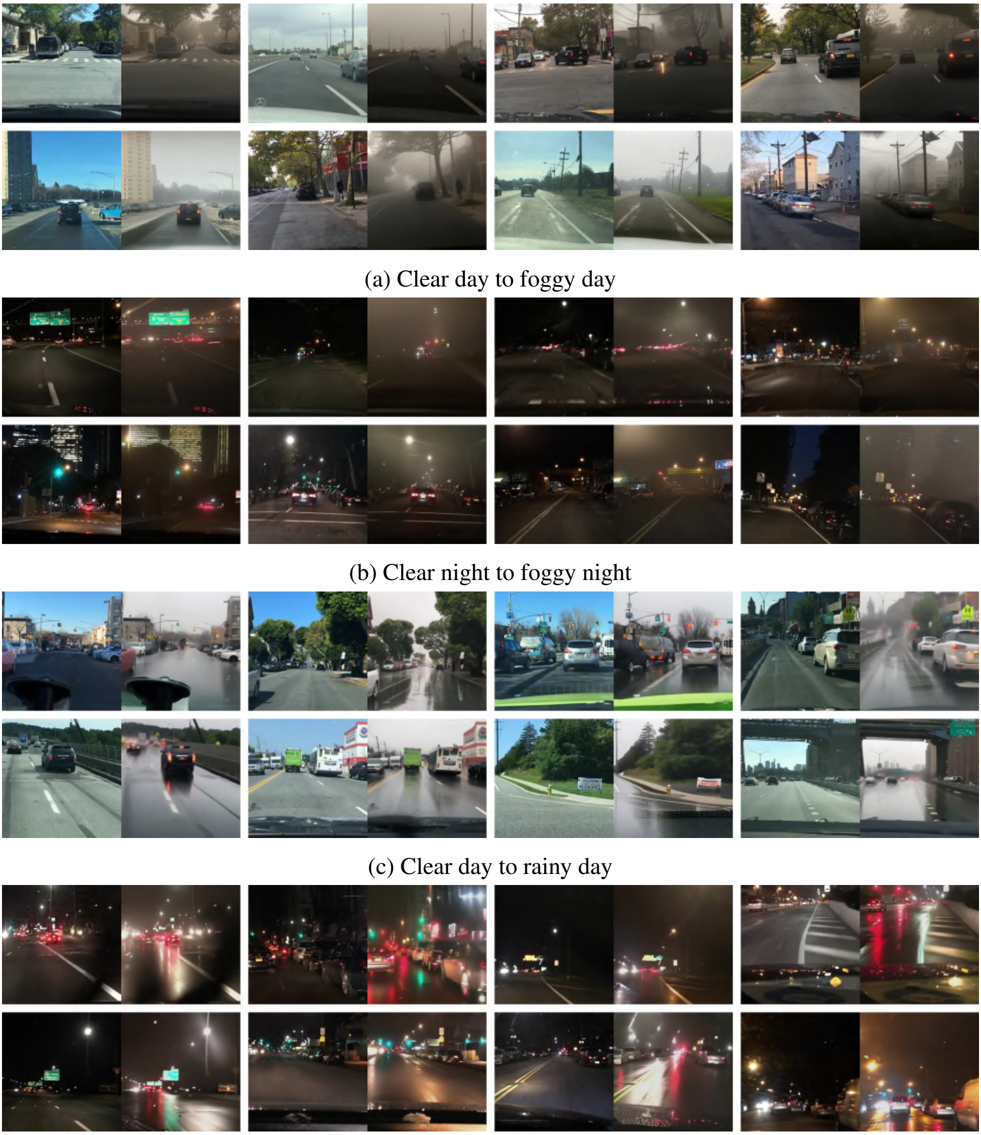
Figure 7, Figure 8 and Figure 9 demonstrate our model's performance in the more common fields of face editing (gender swap and aging, respectively) and object-swap (cat↔dog). For the faces and cat/dog tasks we used the *FFHQ* [9] and *AFHQ* [3] datasets, respectively. For both datasets we used the publicly available pretrained versions of SD2.1-base [18] with the pretrained version on ControlNet provided by [25] without any additional finetuning. *Sts-GAN* was trained using the same procedure as in the automotive case.

We emphasize that we claim no advantage for our model over existing diffusion-based editing techniques in cases where the source image is relatively simple (e.g., centered objects) or when the edits do not require close adherence to the source image. In such cases, existing diffusion-based methods produce high-quality results, and sometimes operate in a zero-shot manner (unlike our model, which requires task-specific optimization). Nevertheless, our model is versatile enough to achieve qualitatively competitive translations in these scenarios as well, compared to existing methods. Figure 10 illustrates the performance of our *StS* approach in gender-swap, compared to StyleGAN2-Distilation [23], and in age translation, compared to SAM [1], respectively, both over the *FFHQ* dataset. In both gender-swap and age translation tasks, our model demonstrates competitive capabilities compared to the task-oriented baselines. As expected, our model adheres to the facial structure, pose, and expression of the source image, resulting in outputs that resemble a transformation of the input individual rather than depicting a different person from the target domain. This adherence is notably superior compared to the baselines in both tasks.

## F.2 I2IT with Flexible Adherence to Source Images

It should be noted that since we use the Canny map as a spatial condition, our model is committed to preserving the object outlines present in the source image. Therefore, for example, *StS* will not shorten the hair of a female input image when translating it into a male. This attribute limits our model to constrained translations, where the nature of the constraint is determined by the spatial condition provided to the ControlNet. This limitation is discussed in detail below.

While diffusion-based image editing and translation are common, as detailed in the literature review of the main paper, our model excels in cases that require strict adherence to the source image. Unlike the automotive and facial translation tasks discussed — where preserving the structure and semantics of the original image is crucial even in edited areas



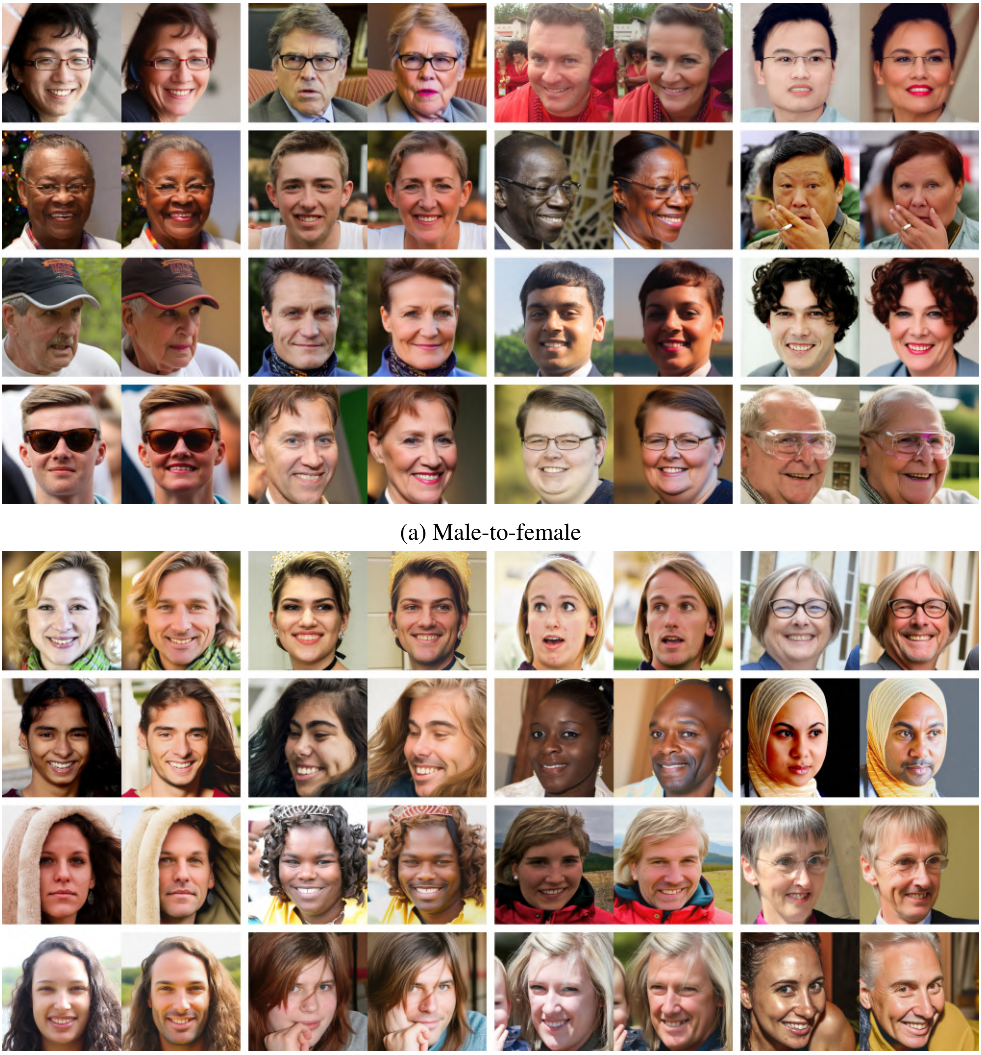
(a) Clear day to foggy day

(b) Clear night to foggy night

(c) Clear day to rainy day

(d) Clear night to rainy night

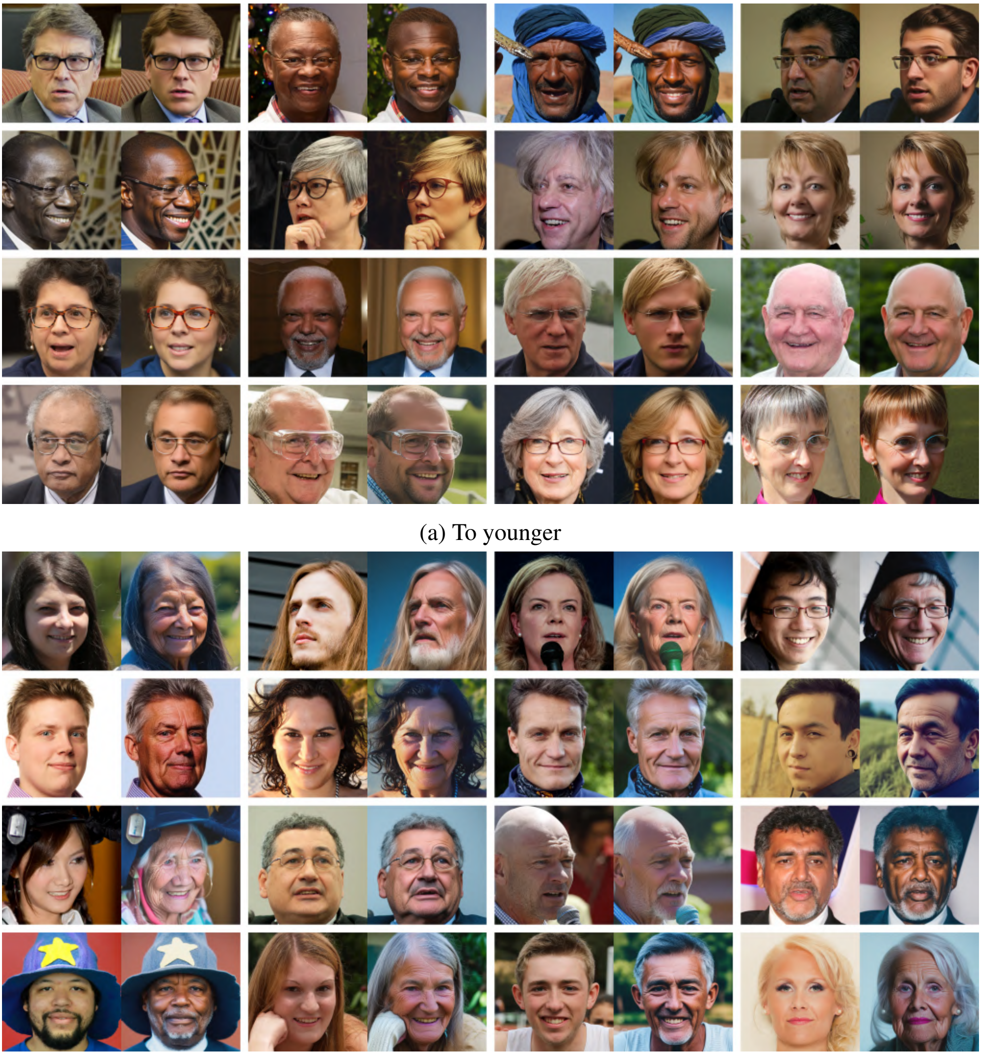
Figure 6: Weather Translation over the *BDD100k* Dataset. In each pair, the left image is the source and the right is the translated.



(a) Male-to-female

(b) Female-to-male

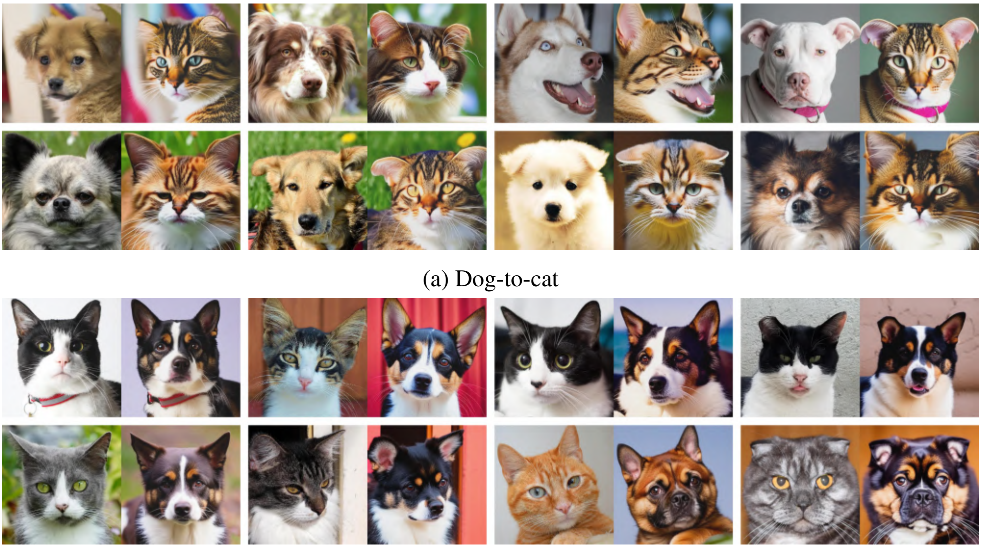
Figure 7: *Male↔Female* translation over the *FFHQ* dataset. In each pair, the left image is the source and the right is the translated.



(a) To younger

(b) To older

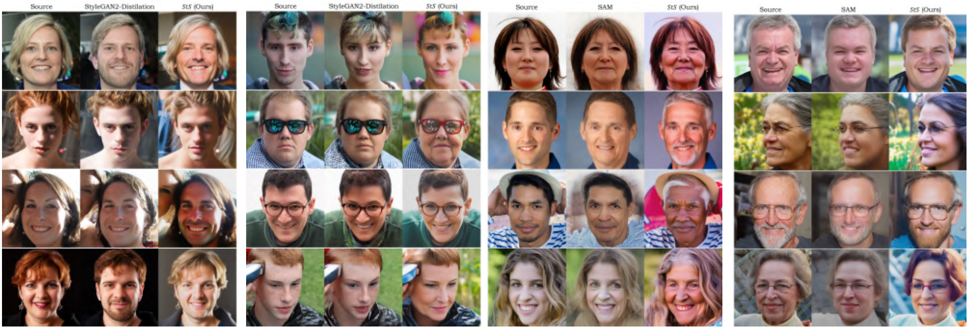
Figure 8: *Younger*↔*Older* translation over the *FFHQ* dataset. In each pair, the left image is the source and the right is the translated.



(a) Dog-to-cat

(b) Cat-to-Dog

Figure 9: *Cat*↔*Dog* translation over the *AFHQ* dataset. In each pair, the left image is the source and the right is the translated.



(a) Female-to-Male (b) Male-to-Female (c) To older (d) To younger

Figure 10: Additional applications: (a,b) Gender Swap, and (c,d) Age Translation over the *FFHQ* dataset.

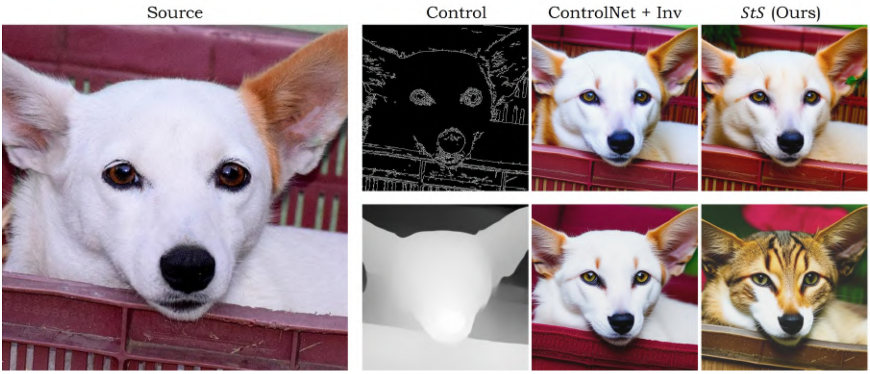


Figure 11: **Dog-to-Cat Translation.** A strict spatial constraint may hinder realism. Although a depth condition (bottom) is less restrictive than Canny (top), allowing the model to adjust the pattern of the fur to a more cat-like one, the boundaries still enforce a doglike structure (nose, eyes, ears, etc.)

— some other image translation tasks require only minimal adherence in the modified regions. For instance, in common dog↔cat or horse↔zebra translations, the primary focus is on replacing one object with another while maintaining only the position of the original object.

While diffusion-based image editing and translation are common, as detailed in the literature review of the main paper, our model excels in cases that require strict adherence to the source image. Unlike the automotive and facial translation tasks discussed — where preserving the structure and semantics of the original image is crucial even in edited areas — some other image translation tasks require only minimal adherence in the modified regions. For instance, in common dog↔cat or horse↔zebra translations, the primary focus is on replacing one object with another while maintaining only the position of the original object. Dogs and cats, for example, differ substantially from each other, so a dog translated from a cat is essentially just a dog posed in the same position as the source cat, without any further adherence to the cat's attributes. In such cases, the strict adherence of our model to the source image — expressed both in the seed space by the *sts-GAN* and along the sampling trajectory by the ControlNet — may limit its ability to perform a realistic translation.

As discussed in Section 5 of the main text, the spatial constraint may restrict translation performance when the provided spatial control only crudely reflects the source domain. For example, in the dog↔cat translation task, Figure 11 illustrates a scenario where substantial structural modifications are necessary to transform a dog into a cat, creating a conflict with the spatial control. In this example, the Canny control impeded the generation of the cat's distinctive fur pattern, as ControlNet prioritized replicating the smooth fur edges of the source dog. Replacing the Canny control with a depth map (obtained using MiDaS [17]) enabled the *sts-GAN* to produce a more cat-like pattern. However, the distinct dog-like boundaries continued to conflict with the desired structural transformation. In such scenarios, the model may struggle to satisfy both spatial and appearance constraints, leading to unedited or unrealistic outputs.

In cases where the spatial control does not crudely reflect the source domain — such as when the boundaries of the source dog can be considered "cat-like" — our model performs the translation accurately, as demonstrated in Figure 9 using a depth map as the spatial con-

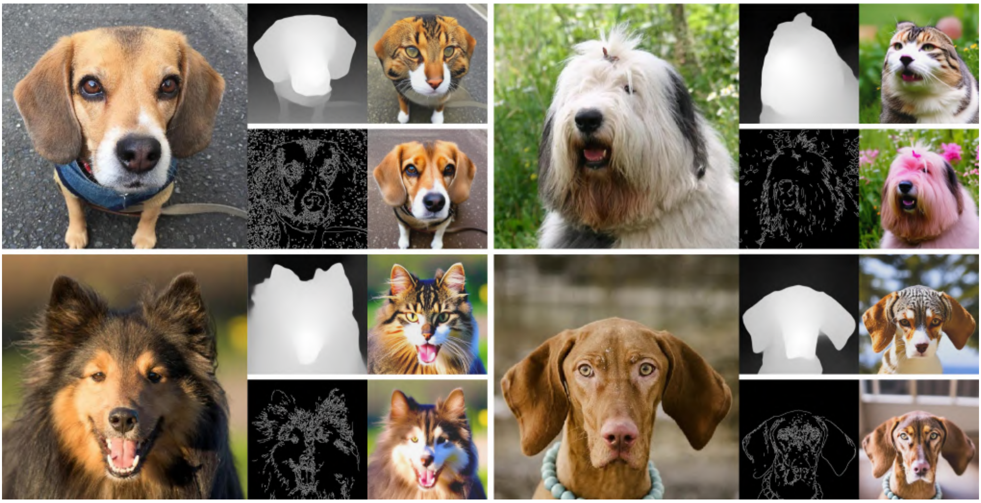


Figure 12: Example of failures in dog-to-cat translations. Each translation is shown for a ControlNet using depth (top) and Canny (bottom) conditions.

trol. It is important to note that while our model strives to closely adhere to attributes that can be preserved (e.g., expression), such strict adherence is not essential for these types of translations. Consequently, other image editing techniques may be more suitable for these kind of tasks.

# References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021.

[2] Tim Brooks, Aleksander Holynski, and Alexei Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proc. CVPR*, pages 18392–18402, 06 2023.

[3] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[4] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[8] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12469–12478, 2024.

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[13] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[14] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[15] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, Feb 2023.

[16] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[17] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[19] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[21] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019.

[22] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[23] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. Stylegan2 distillation for feed-forward image manipulation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 170–186. Springer, 2020.

[24] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023.

[25] Thibaud Zamora. controlnet-sd21-diffusers. https://huggingface.co/thibaud, 2023. Accessed: 2024-05-09.