Language-Grounded Multi-Domain Image Translation via Semantic Difference Guidance

Jongwon Ryu[1]*, Joonhyung Park[2]*, Jaeho Han[1], Yeong-Seok Kim[2],
Hye-rin Kim[2], Sunjae Yoon[3], Junyeong Kim[1]

[1]Department of Artificial Intelligence, Chung-Ang University,
[2]Hyundai Mobis, [3]Korea Advanced Institute of Science and Technology
Correspondence: Junyeongkim@cau.ac.kr

Abstract

Multi-domain image-to-image translation requires grounding semantic differences expressed in natural language prompts into corresponding visual transformations, while preserving unrelated structural and semantic content. Existing methods struggle to maintain structural integrity and provide fine-grained, attribute-specific control, especially when multiple domains are involved. We propose LACE (Language-grounded Attribute-Controllable Translation), built on two components: (1) a GLIP-Adapter that fuses global semantics with local structural features to preserve consistency, and (2) a Multi-Domain Control Guidance mechanism that explicitly grounds the semantic delta between source and target prompts into per-attribute translation vectors, aligning linguistic semantics with domain-level visual changes. Together, these modules enable compositional multi-domain control with independent strength modulation for each attribute. Experiments on CelebA(Dialog) and BDD100K demonstrate that LACE achieves high visual fidelity, structural preservation, and interpretable domain-specific control, surpassing prior baselines. This positions LACE as a cross-modal content generation framework bridging language semantics and controllable visual translation.

1 Introduction

Image-to-image (I2I) translation is a fundamental task in computer vision that aims to alter specific visual attributes of an image while preserving its structural and semantic integrity (Huang et al., 2018; Lee et al., 2018). It has a wide range of applications, including facial attribute editing (Choi et al., 2018), weather or time simulation in driving scenes (Sun et al., 2022), and artistic style transfer (Zhang et al., 2023b, 2024, 2025). In many
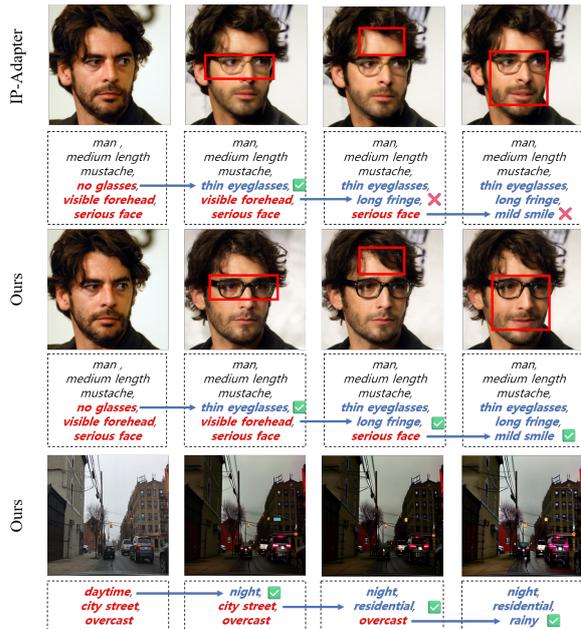

Figure 1: Multi-domain translation results with progressive attribute modifications. Our method enables compositional editing of multiple attributes while preserving non-target regions and structural consistency.

real-world settings, images are influenced by multiple entangled domain attributes (e.g., "snowy night city"), which calls for models capable of modifying several attributes simultaneously without disrupting non-target content, and crucially, grounding semantic differences expressed in natural language prompts into corresponding visual transformations.

Compared to traditional GAN-based methods (Goodfellow et al., 2014), diffusion-based generative models provide more stable training and higher-quality results in image-to-image translation (Saharia et al., 2022; Rombach et al., 2022). Nevertheless, current diffusion frameworks still struggle with controllable and structure-preserving multi-attribute translation. Most are designed for single-attribute editing and rely heavily on text prompts or class labels (Nichol et al., 2021; Kim

---

*Equal contribution

et al., 2022), which limits their flexibility in multi-domain scenarios. Some recent approaches attempt to preserve overall image content by inverting inputs into noise and leveraging attention-based mechanisms for domain-specific attribute editing (Mokady et al., 2023; Hertz et al., 2022), but they often fail to retain fine-grained structures such as small yet semantically critical elements in complex scenes (Gómez et al., 2023). In addition, they lack mechanisms for prompt-driven per-attribute control and fine-grained adjustment of translation strength (Tumanyan et al., 2023; Cao et al., 2023). Their reliance on segmentation masks (Choi et al., 2020) or spatial annotations (Chen, 2017; Ren et al., 2016) further limits applicability in unstructured real-world settings. Consequently, existing models often fail to achieve precise and coherent multi-domain translation. As shown in Figure 1, they yield entangled or distorted outputs when editing multiple attributes simultaneously, underscoring the need for a framework that preserves structure while enabling prompt-level controllability.

To overcome these limitations, we propose LACE (Language-grounded Attribute-Controllable Translation), a diffusion-based framework for multi-domain I2I translation that enables attribute-wise control, including selective editing and per-domain strength modulation. LACE performs multi-attribute translation without relying on region masks or supervision, while preserving the structural integrity of the input image. The LACE introduces two key components: (1) a Global-Local Image Prompt Adapter (GLIP-Adapter) that extracts visual cues by combining global semantics and local structures from a source image, and (2) a Multi-Domain Control Guidance (MCG) module that explicitly grounds the semantic delta between source and target prompts into noise-space translation vectors, thereby aligning linguistic semantics with domain-level visual changes. This design allows our model to support targeted editing, compositional prompt control, and domain-specific strength adjustment, making it suitable for complex real-world scenarios involving entangled visual factors.

We validate our approach on CelebA(Dialog) (Jiang et al., 2021), which involves fine-grained facial attribute editing (e.g., age, gender), and BDD100K (Yu et al., 2020), which requires broader domain-level translation across scene factors such as weather and time.

This choice of datasets allows us to demonstrate the versatility of our method across both local attribute manipulation and global background/style translation. Our experiments involve up to three simultaneous attribute translations per image, and the results show that our method surpasses prior work in translation fidelity, structural consistency, and interpretability of domain-level control, establishing a strong benchmark for controllable diffusion-based multi-domain image translation.

2 Related work

2.1 Conditional Diffusion for I2I Translation

Diffusion models have recently emerged as a powerful alternative to GANs for image-to-image (I2I) translation, offering improved training stability and high-quality outputs (Ho et al., 2020; Saharia et al., 2022). Conditional diffusion methods such as DiffEdit (Meng et al., 2021), Pix2PixDiff (Tumanyan et al., 2023), and ControlNet (Zhang et al., 2023a) enable task-guided or paired I2I generation by injecting structural inputs like edges or pose maps. However, these approaches are primarily designed for one-to-one or task-specific translation, and often lack scalability to more general multi-domain or compositional scenarios. While models like DiT-Image2Image incorporate CLIP guidance to align with visual semantics, they do not explicitly model domain-aware or attribute-wise transformation, limiting their controllability in multi-attribute editing tasks.

2.2 Prompt-Based Guidance and Domain Control

Classifier-Free Guidance (CFG) (Ho and Salimans, 2022) introduced a simple yet effective method for prompt-based conditioning in diffusion models, interpolating between unconditional and conditional predictions. While widely adopted, CFG is inherently limited to single-attribute control and lacks compositionality. Follow-up methods such as Blended CFG (Hertz et al., 2022) and FlexiDiffusion (Cao et al., 2023) extend this idea by mixing multiple prompt embeddings, but without explicitly disentangling or controlling individual domain attributes.

Image-conditioned adapters such as IP-Adapter (Ye et al., 2023) and T2I-Adapter (Mou et al., 2024) use CLIP-based embeddings from a reference image to preserve identity or style during generation. However, these approaches
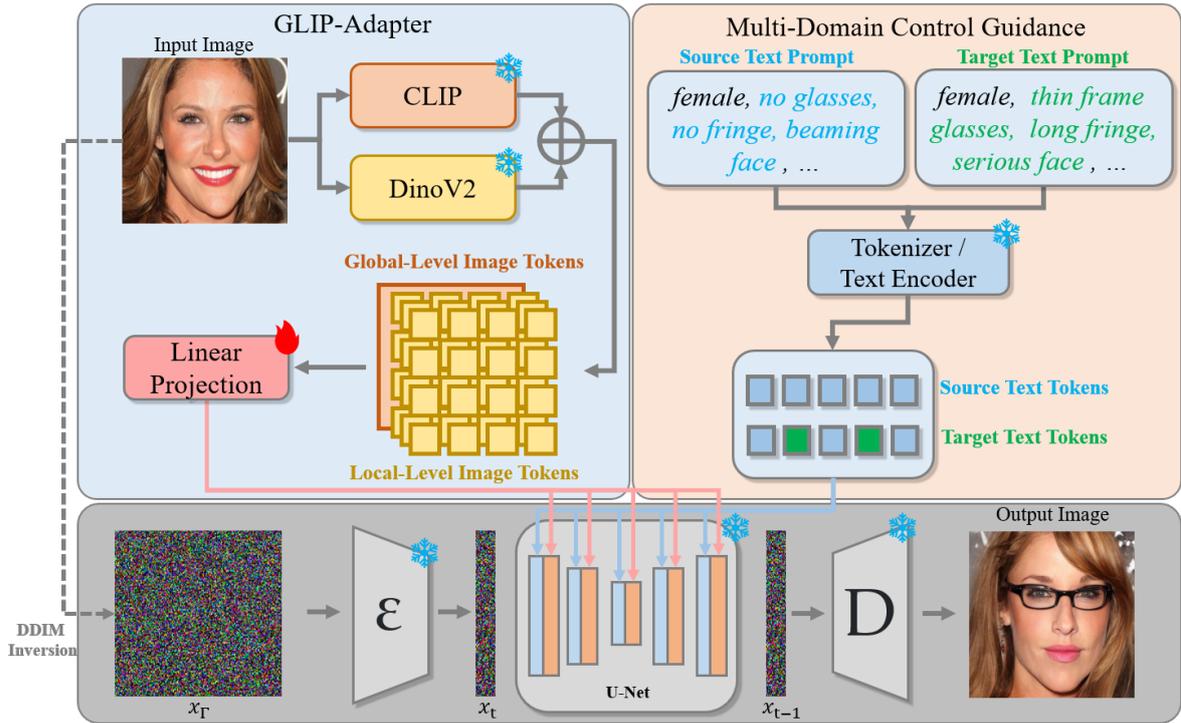
Figure 2: Overview of our proposed multi-domain I2I translation framework. The GLIP-Adapter projects global (CLIP) and local (DINOv2) features into a linear projection, which together with the source text prompt is used to train the U-Net's cross-attention layers. At inference, the MCG module leverages source–target prompt differences to guide controllable multi-attribute translation.

offer limited control over which attributes to modify and lack explicit mechanisms for modeling domain-wise transformation strength or direction.

3 Method

3.1 Preliminaries

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Rombach et al., 2022) are generative models that learn to reverse a stochastic noising process by predicting the noise added to clean data. Given an image x_0, a forward process progressively adds Gaussian noise to obtain a noisy image x_t at time step t. The reverse process, parameterized by ϵ_θ, is trained to predict the added noise ϵ given the noisy input x_t, a conditioning signal c, and timestep t. The training objective is the denoising score matching loss:

$$\mathcal{L} = \mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(0,I), c, t} \left\| \epsilon - \epsilon_\theta(x_t, c, t) \right\|^2, \quad (1)$$

where c can be interpreted as a linguistic condition, where semantic differences across prompts define the attribute shifts to be grounded in the visual domain.

In multi-domain image-to-image (I2I) translation, the goal is to selectively modify specific domain attributes (e.g., weather, expression, or lighting) of an input image while preserving its structural layout and semantic content. This task involves three key challenges: (1) isolating and editing only the intended target attributes, (2) maintaining the integrity of non-target regions, and (3) enabling compositional and fine-grained control across multiple domain axes.

Previous diffusion-based translation methods typically rely on single text prompts or global visual features, limiting their applicability in multi-attribute translation. Approaches such as ControlNet (Zhang et al., 2023a) or IP-Adapter (Ye et al., 2023) introduce strong spatial or identity conditioning but lack explicit mechanisms for prompt-driven, compositional control across domains. To overcome these limitations, we propose a LACE (Language-grounded Attribute-Controllable Translation). An overview of our proposed architecture is shown in Figure 2.

3.2 Global-Local Image Prompt Adapter (GLIP-Adapter)

To preserve structural and semantic details during translation, we introduce the Global-Local Image Prompt Adapter (GLIP-Adapter), which injects visual prompts from a source image into the diffusion model alongside text conditions. Inspired by prior adapter-based methods (Mou et al., 2024; Ye et al., 2023), GLIP-Adapter enables image-to-image translation guided by example-based visual conditioning.

We design GLIP-Adapter to leverage two complementary sources of information: global semantic context and local structural detail. Specifically, we extract:

- **Global tokens** from CLIP (Radford et al., 2021), which capture high-level semantic content (e.g., scene category, weather type) based on joint image-text alignment.

- **Local tokens** from DINOv2 (Oquab et al., 2023), which encode fine-grained spatial features such as object shape, boundary, and arrangement via self-supervised learning.

These global and local tokens are concatenated and projected via a lightweight adapter network, which is trained while keeping the image encoders and diffusion backbone frozen. The resulting prompt embedding c_i is injected into the denoising model through cross-attention layers, alongside the text prompt c_t. The modified loss function becomes:

$$\mathcal{L} = \mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(0,I), c_t, c_i, t} \left\| \epsilon - \epsilon_\theta(x_t, c_t, c_i, t) \right\|^2, \quad (2)$$

By combining CLIP's global semantic awareness with DINOv2's fine-grained spatial representation, GLIP-Adapter enables the model to maintain the layout, structure, and visual identity of the input image throughout the translation process. This design improves upon prior prompt-based methods (e.g., IP-Adapter) that rely solely on global-level image embeddings and often fail to preserve object-level consistency in multi-domain scenarios. To support downstream cross-attention operations such as those in our Multi-Domain Control Guidance (MCG), we apply a linear projection to align the prompt embeddings to a unified feature dimension.

3.3 Multi-Domain Control Guidance (MCG)

To enable flexible and interpretable control over domain-specific attribute translation, we propose Multi-Domain Control Guidance (MCG), a prompt-driven conditioning mechanism that operates on the difference between noise predictions guided by source and target prompts. Unlike traditional classifier-free guidance (CFG) (Ho and Salimans, 2022), which interpolates between unconditional and conditional predictions, MCG explicitly models the domain shift between attributes in the noise space.

Given a source domain prompt r (e.g., *cloudy*) and a target domain prompt $\hat{r}$ (e.g., *sunny*), the diffusion model computes two conditional noise predictions:

$$\epsilon_\theta(x_t, r, c_i, t) \quad \text{and} \quad \epsilon_\theta(x_t, \hat{r}, c_i, t),$$

where c_i is the image prompt from GLIP-Adapter. The translation direction is derived from the difference between these two predictions. The final guided noise is obtained by adding a scaled difference to the source-prompt prediction:

$$\hat{\epsilon}_\theta(x_t, \hat{r}, c_i, t) = \epsilon_\theta(x_t, r, c_i, t) + \\ s \cdot (\epsilon_\theta(x_t, \hat{r}, c_i, t) - \epsilon_\theta(x_t, r, c_i, t)), \quad (3)$$

where s is a translation scale parameter that controls the strength of attribute change. This formulation enables the model to retain source characteristics and apply only the directional change specified by the prompt difference.

To extend this mechanism to multi-domain translation, we introduce compositional control via linear combination. Let $\{(r_d, \hat{r}_d)\}_{d=1}^{D}$ be a set of D source-target domain prompt pairs (e.g., *cloudy $\rightarrow$ sunny*, *day $\rightarrow$ night*). We apply independent noise deltas for each attribute dimension and scale them with domain-specific strengths $\{s_d\}_{d=1}^{D}$:

$$\hat{\epsilon}_\theta(x_t, \hat{r}_d \, c_i, t) = \epsilon_\theta(x_t, r, c_i, t) + \\ \sum_{d=1}^{D} s_d \cdot (\epsilon_\theta(x_t, \hat{r}_d, c_i, t) - \epsilon_\theta(x_t, r, c_i, t)). \quad (4)$$

This design enables three core capabilities:

- **Selective translation:** Only attributes with differing source-target prompts are modified.

Number of translations	Methods	FID↓	FID$_{clip}$↓	Structure Distance↓	Background Preservation				CLIP Similarity↑	Human↑
					PSNR↑	LPIPS↓	MSE↓	SSIM↑		
1 domain	SDEdit (Meng et al., 2021)	53.09	15.59	0.1374	13.73	0.49	0.0278	0.58	21.74	0.78
	DDIM (Song et al., 2020)+PnP (Tumanyan et al., 2023)	69.10	25.00	0.1455	15.01	0.40	0.0273	0.60	22.04	0.82
	DDIM (Song et al., 2020)+MasaCtrl (Cao et al., 2023)	164.74	53.96	0.1747	15.58	0.49	0.0962	0.60	15.83	0.65
	Direct (Ju et al., 2023)+PnP (Tumanyan et al., 2023)	60.52	15.59	0.1195	15.34	0.37	0.0438	0.77	20.31	0.81
	Direct (Ju et al., 2023)+MasaCtrl (Cao et al., 2023)	49.31	17.29	0.0854	10.75	0.36	0.0336	0.68	17.29	0.79
	IP-Adapter (Ye et al., 2023)	47.53	11.38	0.0629	15.34	0.32	0.0257	0.56	17.73	0.88
	LACE (Ours)	**45.50**	**10.61**	**0.0622**	**16.24**	**0.30**	**0.0252**	0.73	**23.12**	**0.91**
2 domains	SDEdit (Meng et al., 2021)	56.97	18.31	0.1465	14.12	0.52	0.0359	0.55	22.35	0.61
	DDIM (Song et al., 2020)+PnP (Tumanyan et al., 2023)	79.53	28.92	0.1533	15.97	0.41	0.0342	0.57	22.70	0.68
	DDIM (Song et al., 2020)+MasaCtrl (Cao et al., 2023)	159.98	54.01	0.1820	15.40	0.49	0.0983	0.59	16.56	0.48
	Direct (Ju et al., 2023)+PnP (Tumanyan et al., 2023)	70.03	19.38	0.1175	11.89	0.41	0.0473	0.73	20.90	0.66
	Direct (Ju et al., 2023)+MasaCtrl (Cao et al., 2023)	56.12	19.88	0.0981	10.55	0.39	0.0381	0.64	16.09	0.63
	IP-Adapter (Ye et al., 2023)	46.57	13.45	0.0783	16.35	0.38	0.0315	0.57	17.87	0.70
	LACE (Ours)	**45.77**	**11.98**	**0.0761**	**16.90**	**0.33**	**0.0340**	0.69	**23.38**	**0.92**
3 domains	SDEdit (Meng et al., 2021)	58.97	19.74	0.1485	13.85	0.53	0.0398	0.54	22.27	0.45
	DDIM (Song et al., 2020)+PnP (Tumanyan et al., 2023)	76.28	28.65	0.1553	15.75	0.41	0.0397	0.55	22.60	0.54
	DDIM (Song et al., 2020)+MasaCtrl (Cao et al., 2023)	158.40	53.86	0.1864	15.26	0.50	0.1009	0.58	16.83	0.32
	Direct (Ju et al., 2023)+PnP (Tumanyan et al., 2023)	73.93	21.71	0.2219	10.58	0.43	0.0495	0.70	21.22	0.51
	Direct (Ju et al., 2023)+MasaCtrl (Cao et al., 2023)	54.94	22.49	0.1168	10.86	0.42	0.0451	0.60	14.01	0.48
	IP-Adapter (Ye et al., 2023)	47.56	15.84	0.0957	16.28	0.47	0.0392	0.55	17.62	0.52
	LACE (Ours)	**46.17**	**13.32**	**0.0833**	**16.31**	**0.34**	**0.0388**	0.68	**23.24**	**0.92**

Table 1: **Quantitative evaluation for multi-domain image-to-image translation methods on CelebA**. The best results are highlighted in bold, the second best results are marked with an underline.

Number of translations	Methods	FID↓	FID$_{clip}$↓	Structure Distance↓	Background Preservation				CLIP Similarity↑	Human↑
					PSNR↑	LPIPS↓	MSE↓	SSIM↑		
1 domain	SDEdit (Meng et al., 2021)	42.95	11.50	**0.0389**	20.87	0.23	0.0157	0.66	17.99	0.75
	DDIM (Song et al., 2020)+PnP (Tumanyan et al., 2023)	48.36	8.49	0.0737	14.80	0.34	0.0392	0.53	18.68	0.81
	DDIM (Song et al., 2020)+MasaCtrl (Cao et al., 2023)	112.53	27.52	0.1196	12.22	0.43	0.0664	0.43	15.37	0.62
	Direct (Ju et al., 2023)+PnP (Tumanyan et al., 2023)	40.56	**6.11**	0.0719	15.13	0.31	0.0372	0.54	18.29	0.83
	Direct (Ju et al., 2023)+MasaCtrl (Cao et al., 2023)	73.44	15.75	0.0959	13.59	0.38	0.0531	0.49	16.00	0.79
	IP-Adapter (Ye et al., 2023)	54.05	12.59	0.065	17.34	0.32	0.022	0.56	18.71	0.88
	LACE (Ours)	**40.15**	7.53	0.0453	**21.96**	**0.21**	**0.0092**	**0.73**	18.82	**0.91**
2 domains	SDEdit (Meng et al., 2021)	44.53	11.92	**0.0389**	20.87	0.23	0.0099	0.66	16.04	0.58
	DDIM (Song et al., 2020)+PnP (Tumanyan et al., 2023)	49.62	9.15	0.0752	14.93	0.34	0.0378	0.54	17.29	0.64
	DDIM (Song et al., 2020)+MasaCtrl (Cao et al., 2023)	113.84	27.54	0.1213	12.25	0.43	0.0663	0.43	14.85	0.45
	Direct (Ju et al., 2023)+PnP (Tumanyan et al., 2023)	43.83	7.07	0.0736	15.21	0.31	0.0367	0.55	16.88	0.67
	Direct (Ju et al., 2023)+MasaCtrl (Cao et al., 2023)	63.53	13.61	0.0932	13.61	0.37	0.0499	0.49	13.86	0.61
	IP-Adapter (Ye et al., 2023)	54.72	12.68	0.0652	17.35	0.32	0.022	0.56	16.55	0.70
	LACE (Ours)	**40.53**	**7.64**	0.0466	**21.98**	**0.21**	**0.0091**	0.72	**18.80**	**0.90**
3 domains	SDEdit (Meng et al., 2021)	45.61	11.87	**0.039**	20.87	0.23	0.0132	0.66	12.99	0.44
	DDIM (Song et al., 2020)+PnP (Tumanyan et al., 2023)	51.00	9.71	0.0759	15.01	0.34	0.0372	0.53	15.09	0.54
	DDIM (Song et al., 2020)+MasaCtrl (Cao et al., 2023)	115.11	27.97	0.1234	12.18	0.44	0.0673	0.43	12.39	0.31
	Direct (Ju et al., 2023)+PnP (Tumanyan et al., 2023)	46.61	7.83	0.075	15.25	0.31	0.0363	0.55	14.37	0.53
	Direct (Ju et al., 2023)+MasaCtrl (Cao et al., 2023)	62.71	13.16	0.0905	13.69	0.37	0.0492	0.49	10.61	0.48
	IP-Adapter (Ye et al., 2023)	55.09	12.88	0.0654	17.35	0.32	0.0219	0.56	13.15	0.52
	LACE (Ours)	**42.53**	**7.76**	0.0494	**21.95**	**0.22**	**0.0106**	0.69	**17.75**	**0.90**

Table 2: **Quantitative evaluation for multi-domain image-to-image translation methods on BDD100K**. The best results are highlighted in bold, the second best results are marked with an underline.

- **Compositional control:** Multiple attributes (e.g., weather, time, style) can be changed simultaneously via prompt composition.

- **Per-attribute scaling:** Each domain axis can be independently modulated using s_d, allowing fine-grained control over translation intensity.

Unlike prior approaches that rely on binary class labels or segmentation masks to isolate attributes, MCG enables soft, interpretable guidance using domain-aware prompt differences. Combined with GLIP-Adapter, this module provides a unified mechanism for structure-preserving, multi-domain translation with precise attribute-level control.

4 Experiments

We evaluated our model on two real-world datasets, CelebA(Dialog) and BDD100K, each featuring multiple domain attributes such as facial expression, identity, weather, time, and scene type. Our experiments are designed to assess the model's ability to (1) perform multi-attribute translations while preserving structural integrity, (2) enable compositional control across multiple domains, and (3) support fine-grained per-domain strength adjustment.

4.1 Implementation Details

Our method is implemented using the Hugging-Face Diffusers library (von Platen et al., 2022), with Stable Diffusion v2.1 as the base model. During training, all images are resized to 512×512 and encoded into a latent space using the pretrained
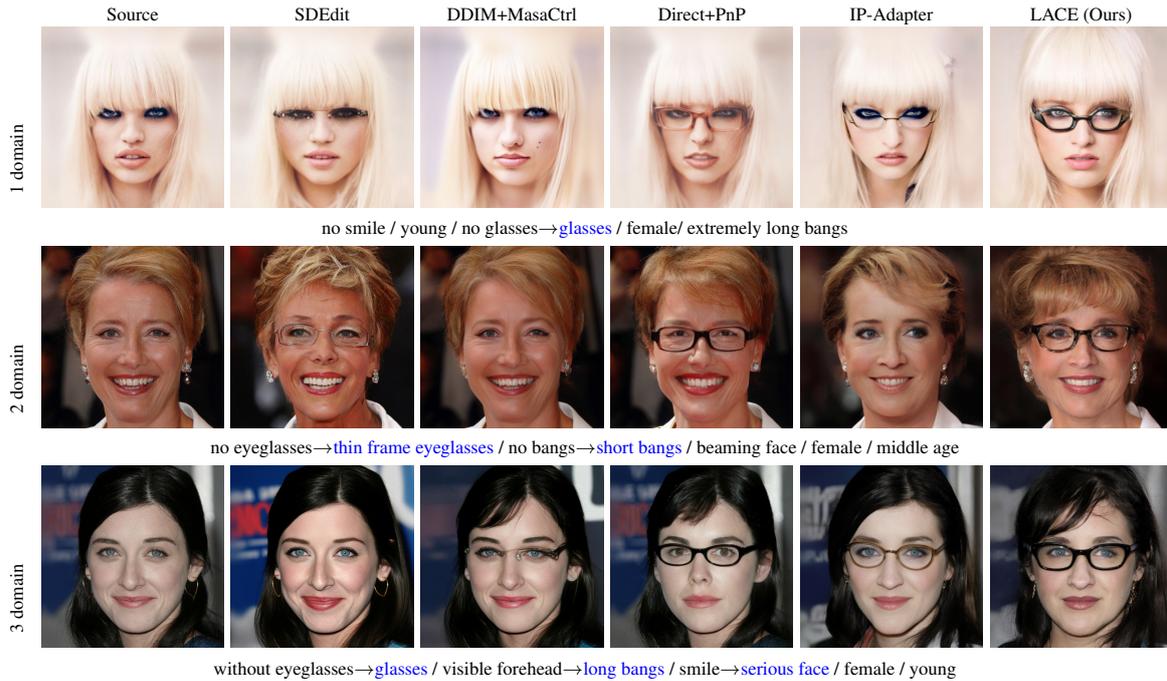
Source   **SDEdit**   **DDIM+MasaCtrl**   **Direct+PnP**   **IP-Adapter**   **LACE (Ours)**

1 domain: no smile / young / no glasses→glasses / female/ extremely long bangs

2 domain: no eyeglasses→thin frame eyeglasses / no bangs→short bangs / beaming face / female / middle age

3 domain: without eyeglasses→glasses / visible forehead→long bangs / smile→serious face / female / young

Figure 3: **Qualitative evaluation for multi-domain image-to-image translation methods on CelebA.**

VAE. The model is fine-tuned using two NVIDIA A6000 GPUs for 200,000 steps with a batch size of 24 per GPU and a fixed learning rate of 1×10^{-5}. The adapters were trained using two A6000 GPUs for 200,000 steps, with a batch size of 24 per GPU, using CLIP-ViT-H/14 (Ilharco et al., 2021) and DINOv2-Large (Oquab et al., 2023) as image encoders. For compatibility with baseline models, we utilized the HuggingFace Diffusers library (von Platen et al., 2022) for all diffusion-based experiments. We also conducted additional experiment on animal face dataset to compare with existing text-guided translation models, which are described in the supplementary material.

4.2 Multi-Domain Image to Image Translation

We conducted comparative experiments with existing multi-domain I2I translation models to evaluate the effectiveness of our method. For baseline models, we included SDEdit (Meng et al., 2021), DDIM, Direct Inversion (Song et al., 2020; Ju et al., 2023) with editing methods MasaCtrl (Cao et al., 2023), Plug-and-Play (Tumanyan et al., 2023), and IP-Adapter (Ye et al., 2023) to compare performance across a diverse range of models. Leveraging the multiple domain characteristics of our dataset, we conducted model performance evaluations by adjusting the number of translated domains. From 500 validation scenarios, we randomly selected one image and applied translations across a randomly chosen set of 1, 2 or 3 domains. Table 1, Table 2 show the results, indicating that our method achieved sota in most metrics, or second-best performance across most evaluation metrics. This suggests that our model not only preserves the structural and contextual content of the source image, but also performs accurate and effective multi-domain translation.

Furthermore, we observed that standard metrics such as FID and CLIP similarity may not fully capture the increasing difficulty of multi-attribute editing, as they often reward conservative, under-edited results that maintain high visual stability but fail to accurately reflect semantic changes. To address this limitation and better assess the true effectiveness of multi-domain translation, we conducted a human evaluation focusing on attribute correctness and visual naturalness. Five human evaluators assessed 100 samples from CelebA, BDD100K dataset. The evaluators rated the degree of semantic alignment between the visual outputs and the text prompts on a scale of 0 to 10, which were subsequently normalized to a range of $[0, 1]$. As shown in the Human column of Table 1 and Table 2, while the performance of most baselines degrades significantly as the number of edited domains increases. For instance, IP-Adapter's score drops from 0.88 to 0.52, LACE maintains a nearly constant high level of correctness (from 0.92 to 0.90). These results confirm that LACE is robust in accurately applying all requested attributes, a characteristic that human

Figure 4: **Qualitative evaluation for multi-domain image-to-image translation methods on BDD100K.**

Figure 5: **The ablation study of multi-domain image-to-image translation methods**

evaluation highlights even when standard metrics remain relatively stable.

Qualitative evaluation was performed on four popular criteria: image quality (FID (Heusel et al., 2017), FID$_{clip}$(Kynkäänniemi et al., 2022)), structure distance(Tumanyan et al., 2022), background preservation (PSNR, LPIPS (Zhang et al., 2018), MSE, SSIM (Wang et al., 2004)), translation quality (CLIP Similarity (Hessel et al., 2021)) to measure how well the visual outputs align with the linguistic semantics of the target prompt. Qualitative results in Figure 3,Figure 4 further demonstrate that our method produces visually coherent and attribute-consistent outputs across diverse translation scenarios.

4.3 Ablation Study

We conduct a series of ablation studies to evaluate the contribution of each component in our frame-work. The baseline is IP-Adapter (Ye et al., 2023), which uses CLIP-based global visual prompts. We progressively introduce modifications and measure performance changes across 3 domain translation settings.

- **CLIP-Full:** We first enhance the baseline by incorporating patch-wise local tokens extracted from the CLIP encoder, in addition to the original global token. This modification improves both FID and FID-CLIP scores, indicating better image quality and structural preservation.

- **DINOv2-Full:** Next, we replace CLIP with DINOv2 as the visual encoder, which provides stronger spatial features. This leads to further improvements in structure-sensitive metrics and better preservation of Local-level details such as object color and shape visualized in fig. 5.

- **GLIP-Adapter:** GLIP-Adapter builds on DINOv2-Full by replacing the global token with that of CLIP while retaining the local tokens from DINOv2.

- **MCG:** We compare our Multi-Domain Control Guidance (MCG) with traditional Classifier-Free Guidance (CFG) (Ho and Salimans, 2022). While CFG interpolates between unconditional and conditional denois-

CelebA

| Method | FID↓ | FID$_{clip}$↓ | Structure Distance↓ | Background Preservation | | | | |
				PSNR↑	LPIPS↓	MES↓	SSIM↑	ClIP Sim↑
Baseline	47.22	13.55	0.0789	15.99	0.39	0.033	0.56	17.74
+CLIP-Full △	-1.35	-1.36	-0.0032	+0.22	-0.0099	-0.0072	+0.05	+3.52
+DinoV2-Full △	-0.02	+0.09	-0.0002	+0.09	-0.0132	+0.0021	+0.01	+0.73
+GLIP-Adapter △	-0.02	-0.06	-0.0007	0.00	0.0000	-0.0017	0.00	+0.15
+MCG △	-0.57	-0.24	-0.0013	+0.18	-0.0169	-0.002	+0.06	+0.91
+DDIM △	- 0.08	- 0.58	- 0.0002	+ 0.10	- 0.0008	- 0.0001	+ 0.14	+0.19
Ours	45.18	11.97	0.0738	16.48	0.32	0.0318	0.7	23.24

BDD100K

| Method | FID↓ | FID$_{clip}$↓ | Structure Distance↓ | Background Preservation | | | | |
				PSNR↑	LPIPS↓	MES↓	SSIM↑	ClIP Sim↑
Baseline	54.62	12.71	0.0652	17.34	0.32	0.0219	0.56	16.13
+CLIP-Full △	-10.14	-3.16	-0.0045	+2.51	-0.03	+0.0010	-0.03	+1.37
+DinoV2-Full △	-0.19	+0.20	-0.0022	+0.36	-0.04	-0.0026	+0.01	+0.67
+GLIP-Adapter △	-0.26	-0.14	-0.0021	+0.02	-0.02	-0.0015	0.01	+0.17
+MCG △	-7.03	-0.57	-0.0040	+1.88	-0.09	-0.0032	+0.06	+0.27
+DDIM △	-0.98	-0.08	-0.0020	+0.30	-0.08	-0.0031	+0.14	+0.37
Ours	41.07	7.64	0.0471	21.96	0.21	0.0096	0.71	23.24

Table 3: **Ablation study results showing performance metric variations in the multi-domain i2i translation methods.** △ denotes the performance difference relative to the previous experiment.

ing, MCG explicitly models attribute shift using the difference between source and target prompt predictions. Our experiments show substantial performance gains in FID and CLIP similarity when using MCG.

- **DDIM:** Finally, we incorporate DDIM Inversion (Song et al., 2020) to accurately reconstruct the initial noise from the input image. This further enhances the preservation of structural and contextual details in complex scenes.

Quantitative results are reported in Table 3, and visual comparisons are shown in fig. 5, both demonstrating the effectiveness of each proposed component in improving translation fidelity, structural consistency, and controllability.

4.4 Translation Scale Control

We demonstrate the model's capability to adjust the strength of domain translation via the scaling factor s on the CelebA dataset. As shown in fig. 7, increasing s amplifies the degree of stylistic transfer from the target domain while maintaining the structural integrity of the source image. Lower values of s result in subtle modifications, preserving more of the original domain's appearance, whereas higher values push the model to emphasize features more closely aligned with the target domain. This behavior validates the model's ability to perform controllable and progressive translation, which is particularly important for applications requiring user-interactive manipulation or gradual domain adaptation. The use of a continuous scalar s introduces a simple yet effective mechanism to traverse the interpolation space between source and target domains without retraining the model.

4.5 Per-Domain Scaling

While global scaling via a single s value is effective for coarse control, many real-world applications demand finer, attribute-specific adjustments, especially in complex multi-domain translation settings. To address this, we introduce a differential translation scaling mechanism by assigning separate scaling factors $s_1, s_2, \ldots, s_D$ for each of the D domain attributes. On the BDD100K dataset, we apply distinct scaling coefficients to individual domain factors such as weather and time-of-day. As illustrated in fig. 8, this allows precise, independent modulation of each attribute's transformation strength, enabling complex yet interpretable compositions, for instance, increasing snowfall while keeping the scene at dusk. Such capability facilitates tailored image generation that adapts to specific user intents or context-aware conditions.

Unlike traditional methods that use a single unified scaling parameter, our approach disentangles the contribution of each domain axis, allowing diverse combinations of partial translations within a single forward pass. This fine-grained translation control significantly expands the model's expressiveness.

5 Conclusion

We presented LACE, a diffusion-based framework for controllable multi-domain image-to-image translation, addressing both structural preservation and fine-grained attribute control. Our approach introduces two key components:(1) the Global-Local Image Prompt Adapter (GLIP-Adapter), which fuses semantic and spatial cues from the input image to guide structure-aware translation, and(2) the Multi-Domain Control Guidance (MCG), which enables targeted and compositional editing via prompt-driven attribute steering. Unlike prior methods that rely on spatial masks or support only single-attribute editing, our framework enables flexible and interpretable translation across multiple domain axes without compromising scene consistency. Experiments on CelebA and BDD100K demonstrate superior performance in visual fidelity, structural integrity, and multi-attribute controllability. Further, ablation and scale control studies confirming that controllable image translation can be framed as a language grounding problem, where semantic differences between prompts drive structured, domain-specific transformations.

Limitations

While LACE demonstrates strong controllability and structural preservation across both fine-grained attribute editing (CelebA) and domain-level translation (BDD100K), several limitations remain.

First, the framework introduces additional computational overhead: the GLIP-Adapter and multi-domain guidance require multiple noise predictions per domain, and inference cost grows as more attributes are edited simultaneously, limiting real-time applicability.

Second, our evaluation is restricted to three datasets (CelebA, BDD100K, Animal Faces). Although these cover both local attributes and global style/background domains, broader validation on more diverse domains (e.g., medical images, artwork, video) is left for future work.

Third, attribute interference may arise when editing multiple factors simultaneously: While per-domain scaling alleviates this issue, semantic conflicts between attributes (e.g., gender and hair length) remain a challenge.

Fourth, the framework depends on source–target prompt differences, making it sensitive to linguistic ambiguity; future work could incorporate advances in natural language understanding or prompt paraphrasing to improve robustness; better prompt robustness or automatic prompt refinement could further improve stability.

Ethical Considerations

As a controllable image translation framework, LACE could be misused for generating or altering visual content in deceptive ways. While our work is intended for research and data augmentation under ethical use, future extensions should include watermarking or misuse detection mechanisms to mitigate potential risks.

Acknowledgements

References

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570.

Liang-Chieh Chen. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797.

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

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Jose L. Gómez, Manuel Silva, Antonio Seoane, Agnès Borrás, Mario Noriega, Germán Ros, Jose A. Iglesias-Guitian, and Antonio M. López. 2023. All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes. *arXiv preprint arXiv:2312.12176*.

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

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

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

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. 2021. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. 2023. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2022. The role of imagenet classes in fréchet inception distance. *arXiv preprint arXiv:2203.06026*.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51.

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

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. 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.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

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

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10.

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

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

Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. 2022. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382.

Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2022. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. 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.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman,

Yiyi Xu, Steven Liu, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023b. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156.

Zhanjie Zhang, Ao Ma, Ke Cao, Jing Wang, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, and Yuhui Yin. 2025. U-stydit: Ultra-high quality artistic style transfer using diffusion transformers. *arXiv preprint arXiv:2503.08157*.

Zhanjie Zhang, Quanwei Zhang, Huaizhong Lin, Wei Xing, Juncheng Mo, Shuaicheng Huang, Jinheng Xie, Guangyuan Li, Junsheng Luan, Lei Zhao, and 1 others. 2024. Towards highly realistic artistic style transfer via stable diffusion with step-aware and layer-aware prompt. *arXiv preprint arXiv:2404.11474*.
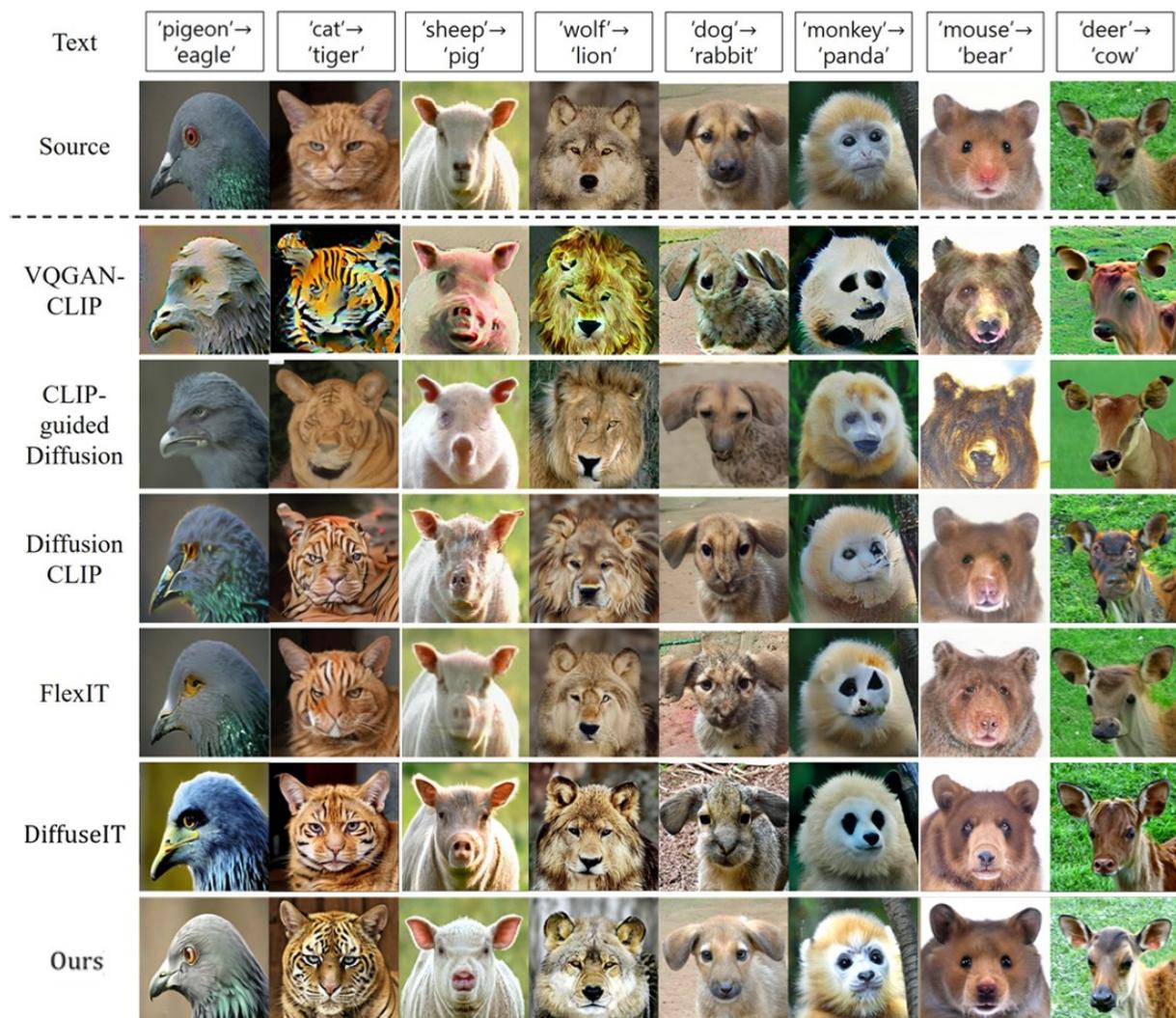
A Appendix

Figure 6: **Qualtitative comparison with text-guided translation models** on **Animal Faces** dataset, such as DiffuseIT. Given limitations in computational resources, we leveraged the experimental setup used in DiffuseIT, training our model under identical conditions to compare results with previously evaluated models. We selected Animals Faces dataset for comparison to showcase the effectiveness of our method on more diverse datasets. Our model effectively preserves the structural content from the source image while performing accurate translations.

Figure 7: **Variations in image translation results on CelebA across the translation scale** s. As the translation scale increases, the visual results show stronger translations in each domains guided by the target.
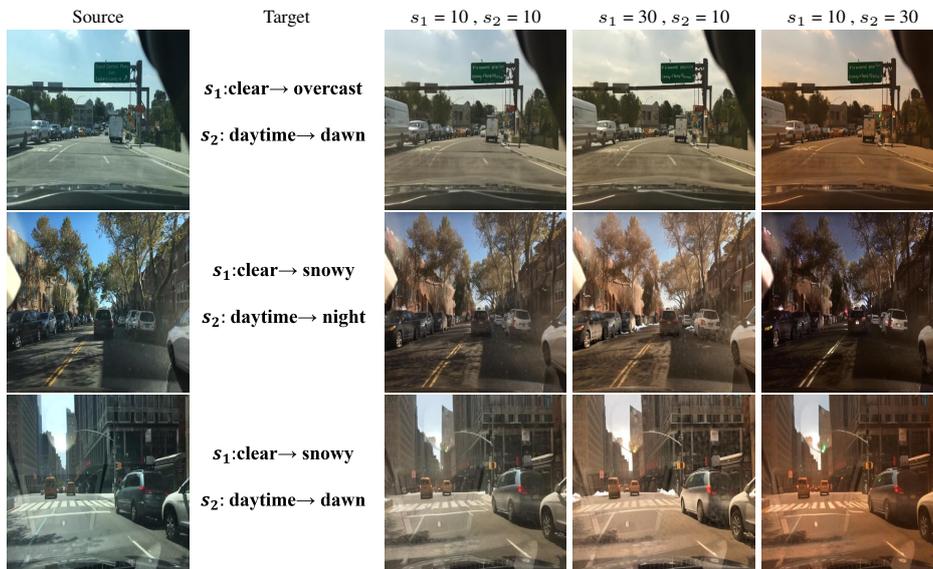

Figure 8: **Variations in image translation results on BDD100K across the differential translation scales** s_1, s_2. The two differential translation scales, s_1 and s_2, control the degree of translation for the weather and time of day domains, respectively.