Diffusion-Based Image-to-Image Translation by Noise Correction via Prompt Interpolation

Supplementary Document

Junsung Lee[1] Minsoo Kang[1] Bohyung Han[1,2]

[1]ECE & [2]IPAI, Seoul National University
{leejs0525,kminsoo,bhhan}@snu.ac.kr

A Appendix

In the appendix, we compare the proposed method with Prompt Tuning Inversion [3], which also uses prompt embedding interpolation. Additionally, we compare our method with recent models, MasaCtrl [1] and Edit-Friendly DDIM Inversion [5]. Next, we illustrate the noise correction term and present additional qualitative results. Then, we present additional quantitative results of previous text-driven image-to-image translation methods and our algorithm by measuring relational distance (RD) [6]. We also demonstrate various details of our method, explaining the differences between PIC and classifier-free guidance, as well as the effectiveness of our interpolation strategy and hyperparameters (β, τ). Finally, we utilize another version of Stable Diffusion, Distilled Stable Diffusion, to verify the generalization of our method.

A.1 Comparison with Prompt Tuning Inversion [3]

We compare our method with DDIM [11], Prompt Tuning Inversion (PTI) [3] and a variant version of PTI using images sampled from the LAION-5B dataset [10]. As shown in Tab. 7, although the comparison methods exhibit slightly higher values in CS, the proposed method consistently outperforms DDIM, PTI, and the modified version of PTI in terms of BD and SD by large margins. Furthermore, Fig. 6 illustrates that our method successfully edits the region of interest while maintaining the other parts. In contrast, the other comparison approaches often fail to preserve the background or structure of the source images. Additionally, as depicted in Figs. 6 and 7, relying solely on the prompt interpolation is insufficient to preserve the original background of the source images. Note that the optimized embedding used for reconstruction is replaced with the source prompt embedding provided by BLIP [7] in the variant version of PTI for fair comparisons according to our experimental setting.

Although PTI is somewhat related to PIC in the sense that both methods utilize prompt interpolation, the role of the prompt interpolation is entirely different. The proposed method uses the prompt interpolation in order to estimate the noise correction term which is effective to edit the region of interest while preserving the background area. On the other hand, PTI generates target images

Table 7: Quantitative results to compare the proposed method with DDIM [11], Prompt Tuning Inversion (PTI) [3] and a modified version of PTI, referred to as PTI (modified), on images sampled from the LAION-5B dataset [10] using the pretrained Stable Diffusion [9]. Black and red bold-faced numbers represent the best and second-best performances for each metric in each row.

Task	DDIM			PTI			PTI (modified)			PIC (Ours)		
	CS (↑)	BD (↓)	SD (↓)	CS (↑)	BD (↓)	SD (↓)	CS (↑)	BD (↓)	SD (↓)	CS (↑)	BD (↓)	SD (↓)
dog → cat	0.289	**0.158**	**0.086**	**0.316**	0.193	0.089	0.289	0.177	0.093	**0.293**	**0.045**	**0.031**
cat → dog	0.283	**0.185**	0.089	**0.315**	0.203	**0.085**	0.280	0.202	0.096	**0.288**	**0.057**	**0.033**
horse → zebra	0.325	**0.287**	0.123	**0.346**	0.305	**0.105**	**0.326**	0.306	0.131	0.324	**0.085**	**0.037**
zebra → horse	**0.294**	**0.295**	0.104	**0.316**	0.335	**0.101**	**0.294**	0.322	0.112	0.292	**0.126**	**0.050**
tree → palm tree	0.304	0.234	0.088	**0.340**	**0.208**	**0.077**	0.299	0.255	0.095	**0.314**	**0.085**	**0.036**
dog → dog w/ glasses	**0.318**	**0.134**	**0.072**	**0.340**	0.182	0.080	0.309	0.151	0.077	0.312	**0.026**	**0.016**
Average	0.302	**0.216**	0.094	**0.329**	0.238	**0.090**	0.300	0.236	0.101	**0.304**	**0.071**	**0.034**

by using the standard DDIM-based translation with the prompt interpolation instead of the target prompt embedding. Unlike PTI, the proposed method does not require a backpropagation process through the denoising network, which significantly accelerates the inference time of the proposed method.

A.2 Comparison with Recent Methods

We also compared our algorithm (PIC) with recent models, MasaCtrl [1] and Edit-Friendly DDIM Inversion [5]. Table 8 demonstrates that the results from PIC exhibit superior quality across all metrics. Specifically, PIC achieves the best BD and SD scores, and the second-best CS score.

CS can be trivially improved by translating images to be coherent with the target prompt without considering the similarity with source images. Therefore, there exists an inherent trade-off between CS and {BD, SD} and increasing CS often overfits target prompts only; evaluating algorithms based on CS alone is not reasonable. Although PIC sometimes exhibits lower CS than Edit-Friendly DDIM Inversion [5], it improves BD and SD more significantly.

A.3 Additional Visualization of Noise Correction

Besides the visualization of the noise correction term, $\Delta\epsilon_\theta(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{y}_t)$, in Fig. 2 in main paper, we provide additional results in Fig. 8. The figure validates our intuition that the noise correction term progressively focuses on revising the region of interest while setting the background area to negligible values. Therefore, our method is reasonable and the comprehensive experimental results verify that the proposed method is effective for text-driven image editing tasks.

A.4 Additional Qualitative Results

To validate the generalizability of the proposed method, we present qualitative results of the proposed method on the Animal FacesHQ (AFHQ) dataset [2] in Fig. 9. The figure also supports that the proposed method also generates target

images with high visual quality on the AFHQ dataset. Furthermore, we visualize additional qualitative results of the proposed method along with Prompt-to-Prompt [4], Plug-and-Play [12], and Pix2Pix-Zero [8] in Figs. 10 and 11 using the LAION-5B dataset [10]. These figures demonstrate that PIC achieves remarkable performance across various tasks, and outperforms the previous methods. In addition to Fig. 4 of the main paper, Figs. 12 to 14 illustrate that PIC improves the performance of the state-of-the-art methods when integrated into them.

A.5 Additional Quantitative Results

In addition to Tab. 1 in the main paper, we provide additional quantitative results using the relational distance (RD) [6] to compare the proposed method with existing state-of-the-art approaches [4, 8, 12], where the introduced metric aims to measure how well the relational information between source images is maintained between generated target ones. As shown in Tab. 9, PIC consistently achieves the lowest RD values across all tasks, which implies that PIC effectively preserves the pairwise relationships between images before and after translation.

A.6 Difference from Classifier-free Guidance

The noise prediction in classifier-free guidance is $\hat{\epsilon}_\theta^{\text{cls}}(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}}) \equiv \epsilon_\theta(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}}) + \gamma(\epsilon_\theta(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}}) - \epsilon_\theta(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{y}^{\text{neg}}))$, where $\mathbf{y}^{\text{neg}}$ is set to either the null prompt or the source prompt depending on algorithms. In contrast, our method employs $\epsilon_\theta(\mathbf{x}_t^{\text{src}}, t, \mathbf{y}^{\text{src}}) + \gamma(\epsilon_\theta(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{y}_t) - \epsilon_\theta(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{y}^{\text{src}}))$, which has many different terms from the classifier-free guidance. Note that PIC significantly outperforms the classifier-free guidance with DDIM as presented in Tab. 6 of the main paper. Moreover, the first term in our model aims to reconstruct must-be-preserved regions while the rest provides the flexibility to edit the regions relevant to the target prompt.

A.7 Effect of Adaptive Interpolation in Adding-phrase Cases

Our strategy for phrase-adding cases in Eq. (10) of the main paper simply matches the tokens between the source and target prompts, where the target tokens are used for the added phrase because it doesn't exist in the source prompt. Each token embedding is influenced by the previous token embeddings. Therefore, we consider this for the prompt interpolation strategy.

Fig. 15 shows a comparison between PIC and PIC with simple LERP. The latter fails to adequately preserve both the background and the structure of the object, whereas PIC effectively maintains both. So, we conclude that our interpolation method significantly improves the performance of our method.

A.8 Effect of Hyperparameters τ and β

We analyze the effect of τ and β in Tab. 10. A high β or low τ tends to increase fidelity to the target prompt, resulting in worse BD and SD scores. Conversely,

Table 8: Quantitative comparisons of PIC with MasaCtrl [1] and Edit-Friendly DDIM Inversion [5] on images sampled from the LAION-5B dataset [10] using the pretrained Stable Diffusion [9].

Task	MasaCtrl			DDPM Inv.			PIC (Ours)		
	CS (↑)	BD (↓)	SD (↓)	CS (↑)	BD (↓)	SD (↓)	CS (↑)	BD (↓)	SD (↓)
dog → cat	0.286	**0.150**	**0.060**	**0.312**	0.245	0.100	**0.293**	**0.045**	**0.031**
cat → dog	0.280	**0.140**	**0.058**	**0.313**	0.186	0.076	**0.288**	**0.057**	**0.033**
horse → zebra	0.266	**0.201**	**0.062**	**0.343**	0.306	0.086	**0.324**	**0.085**	**0.037**
zebra → horse	0.288	**0.321**	**0.070**	**0.303**	0.351	0.091	**0.292**	**0.126**	**0.050**
dog → dog w/ glasses	0.295	**0.136**	**0.056**	**0.342**	0.237	0.092	**0.312**	**0.026**	**0.016**
tree → palm tree	0.290	**0.114**	**0.043**	**0.330**	0.239	0.068	**0.314**	**0.085**	**0.036**
Average	0.284	**0.177**	**0.058**	**0.324**	0.261	0.086	**0.304**	**0.071**	**0.034**

Table 9: Quantitative results to compare the proposed method with Prompt-to-Prompt [4], Plug-and-Play [12], and Pix2Pix-Zero [8] on images sampled from the LAION-5B dataset [10] using the pretrained Stable Diffusion [9].

Task	PtP	PnP	P2P	PIC (Ours)
	RD (↓)	RD (↓)	RD (↓)	RD (↓)
dog → cat	**0.350**	0.877	0.411	**0.106**
cat → dog	**0.382**	0.634	0.731	**0.061**
horse → zebra	**1.054**	2.111	1.270	**0.317**
zebra → horse	**0.715**	1.386	1.685	**0.403**
tree → palm tree	0.210	0.660	**0.104**	**0.057**
dog → dog w/ glasses	0.888	1.131	**0.548**	**0.338**
Average	**0.600**	1.133	0.792	**0.214**

Table 10: Results by varying (β, τ) on two tasks, horse → zebra and dog → dog w/ glasses, using data retrieved from the LAION-5B dataset [10].

(β, τ)	horse → zebra						dog → dog w/ glasses					
	(0.1, 25)	**(0.3, 25)**	(0.5, 25)	(0.3, 15)	**(0.3, 25)**	(0.3, 35)	(0.6, 25)	**(0.8, 25)**	(1.0, 25)	(0.8, 15)	**(0.8, 25)**	(0.8, 35)
CS	0.322	**0.324**	0.329	0.332	**0.324**	0.300	0.309	**0.312**	0.319	0.317	**0.312**	0.307
BD	0.077	**0.085**	0.097	0.183	**0.085**	0.035	0.025	**0.026**	0.035	0.068	**0.026**	0.017
SD	0.033	**0.037**	0.041	0.065	**0.037**	0.020	0.015	**0.016**	0.021	0.034	**0.016**	0.012

a low β or high τ enhance the preservation of background or structure in the source images, while decreasing the CS score. Throughout our experiments, we set $\tau = 25$ and $\beta = 0.3$ for word-swap while $\beta = 0.8$ for phrase-adding.

A.9 Other Backbone Models

In the main paper, Stable Diffusion v1.4 is employed for various experiments. Additionally, we utilize another version of the diffusion model, Distilled Stable Diffusion. Figure 16 presents examples demonstrating that our algorithm generalizes well to this alternative pretrained backbone model, Distilled Stable Diffusion.

References

1. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In: ICCV (2023)
2. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: StarGAN v2: Diverse Image Synthesis for Multiple Domains. In: CVPR (2020)
3. Dong, W., Xue, S., Duan, X., Han, S.: Prompt Tuning Inversion for Text-Driven Image Editing Using Diffusion Models. In: ICCV (2023)
4. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-Prompt Image Editing with Cross-Attention Control. In: ICLR (2023)
5. Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An Edit Friendly DDPM Noise Space: Inversion and Manipulations. In: CVPR (2024)
6. Lee, H., Kang, M., Han, B.: Conditional Score Guidance for Text-Driven Image-to-Image Translation. In: NeurIPS (2023)
7. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In: ICML (2022)
8. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-Shot Image-to-Image Translation. In: SIGGRAPH (2023)
9. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: CVPR (2022)
10. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS Datasets and Benchmarks Track (2022)
11. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. In: ICLR (2021)
12. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In: CVPR (2023)

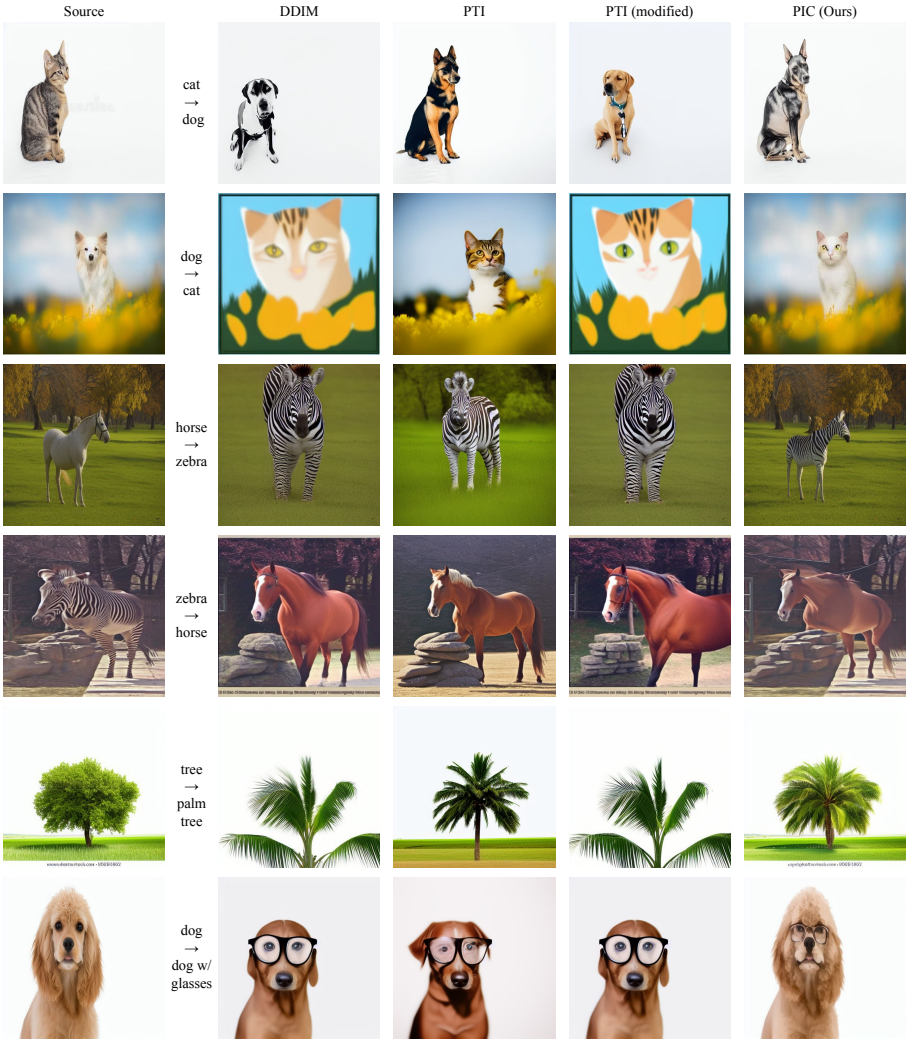

Fig. 6: Quantitative results to compare the proposed method with DDIM [11], PTI [3], and the variant of PTI on real images sampled from the LAION-5B dataset [10] using the pretrained Stable Diffusion [9].

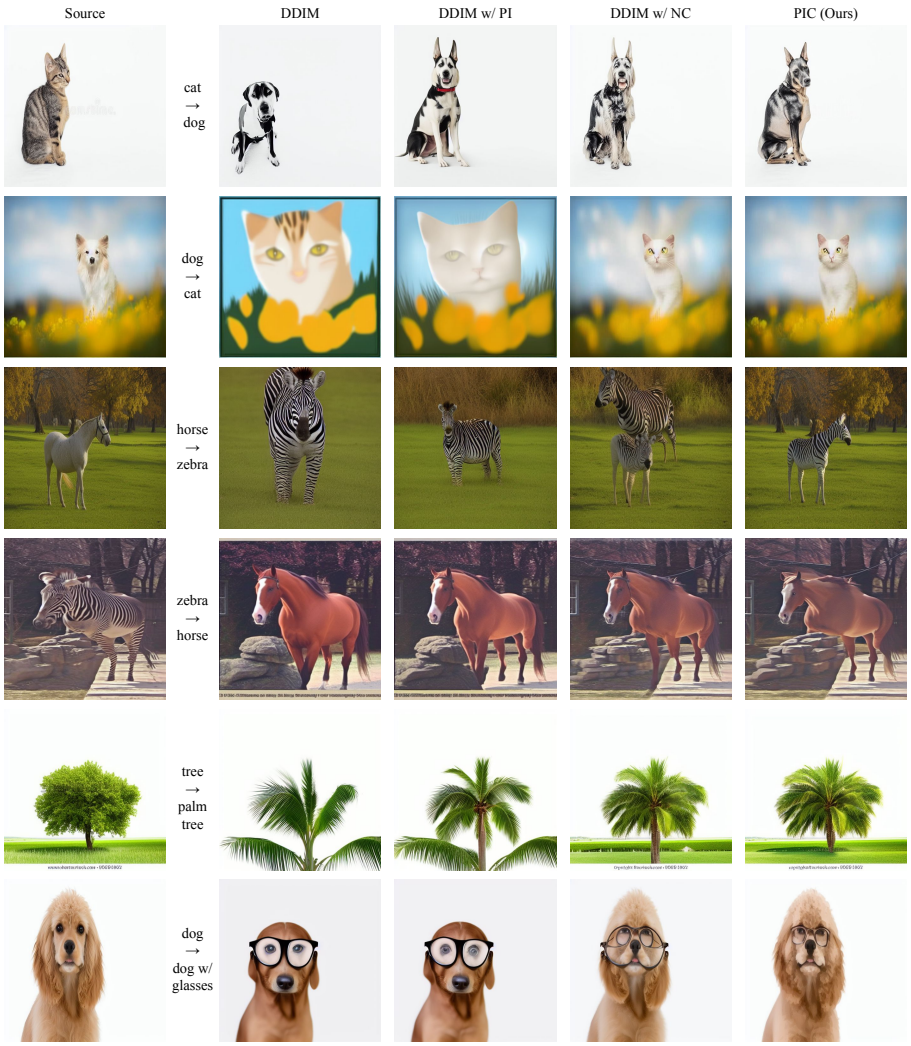

Fig. 7: Qualitative results of our contribution components on real images sampled from the LAION-5B dataset [10] using the pretrained Stable Diffusion [9].

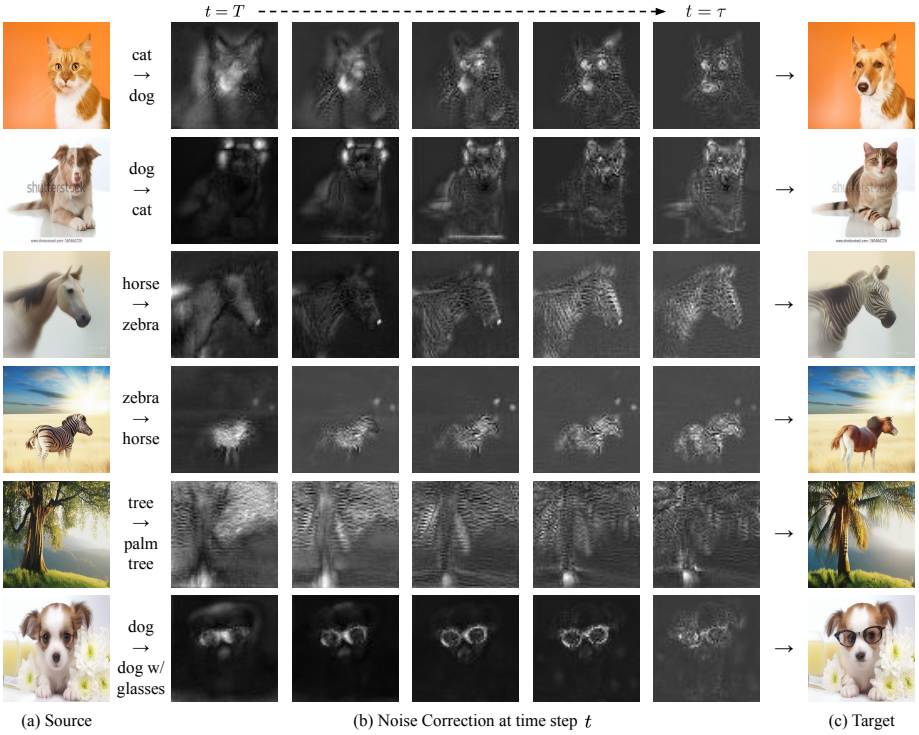

Fig. 8: Additional visualization of the source image, the noise correction $\Delta\epsilon_\theta(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{y}_t)$, and the target image.

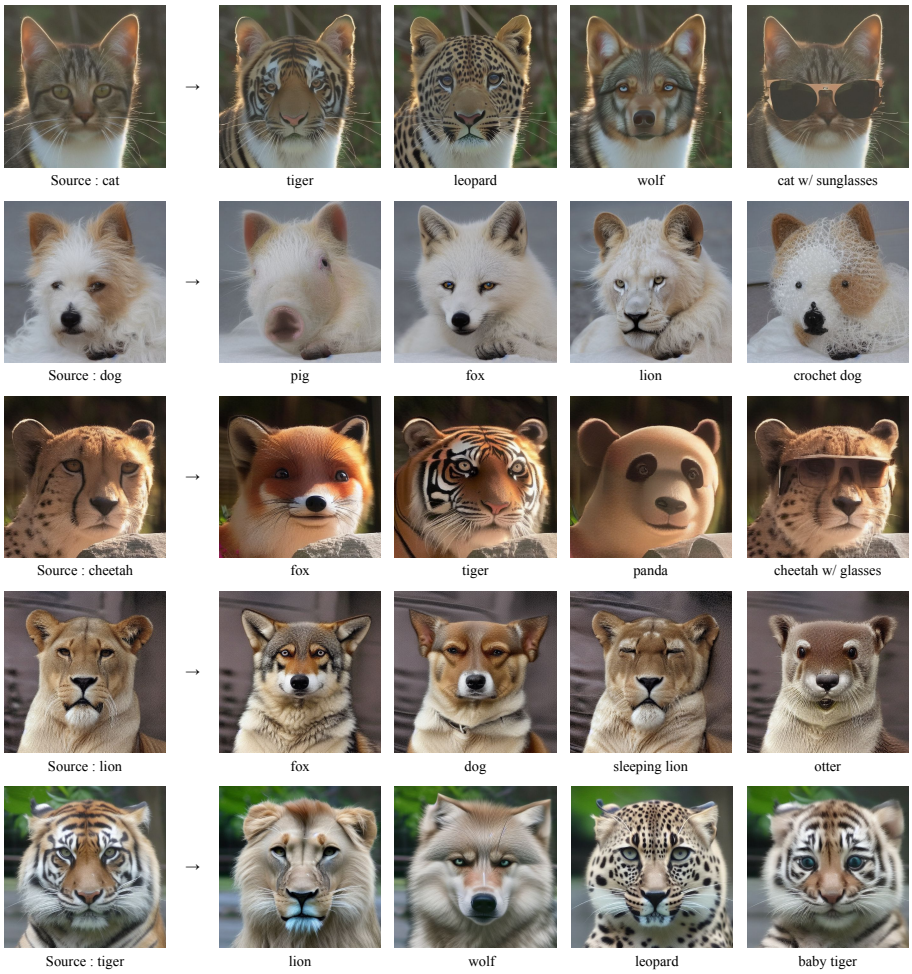

Fig. 9: Qualitative results of our proposed algorithm on data sampled from the AFHQ dataset [2] using the pretrained Stable Diffusion [9].

Fig. 10: Additional qualitative comparisons between PIC and previous methods [4, 8, 12] on real images sampled from the LAION-5B dataset [10] using the pretrained Stable Diffusion [9].

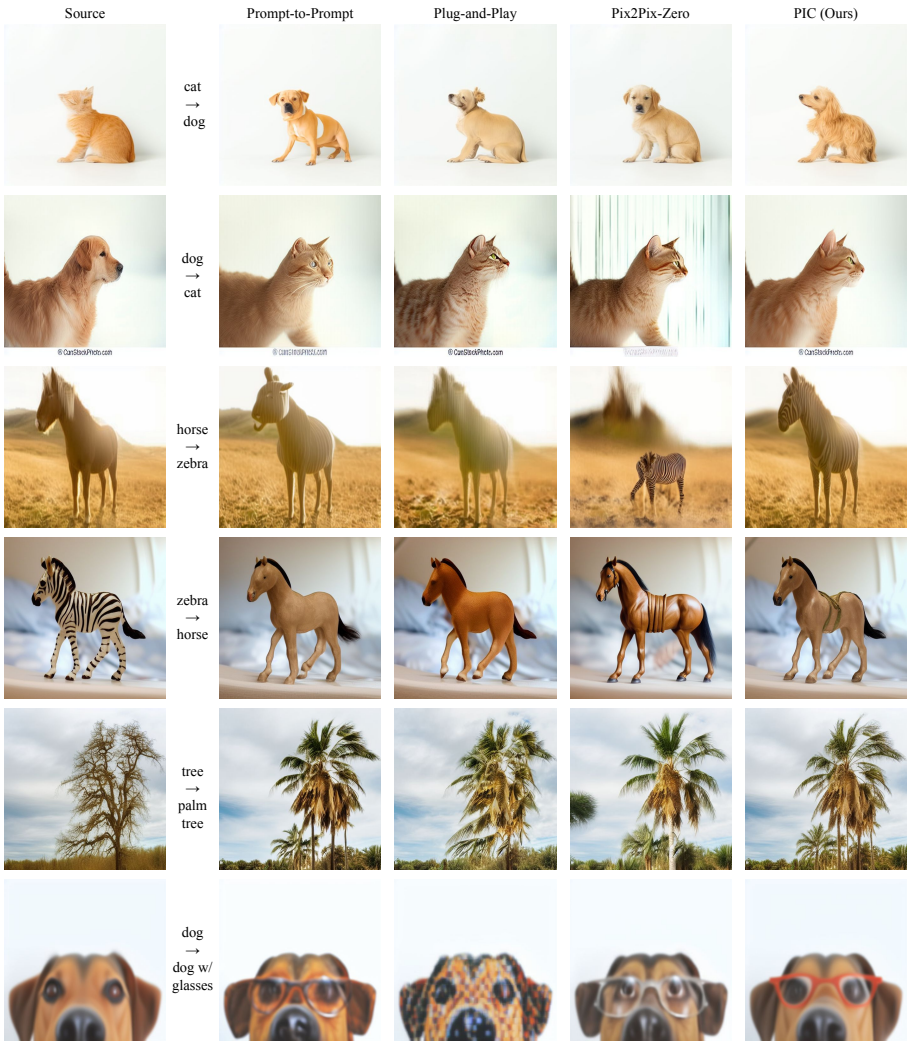

Fig. 11: Additional qualitative comparisons between PIC and previous methods [4, 8, 12] on real images sampled from the LAION-5B dataset [10] using the pretrained Stable Diffusion [9].

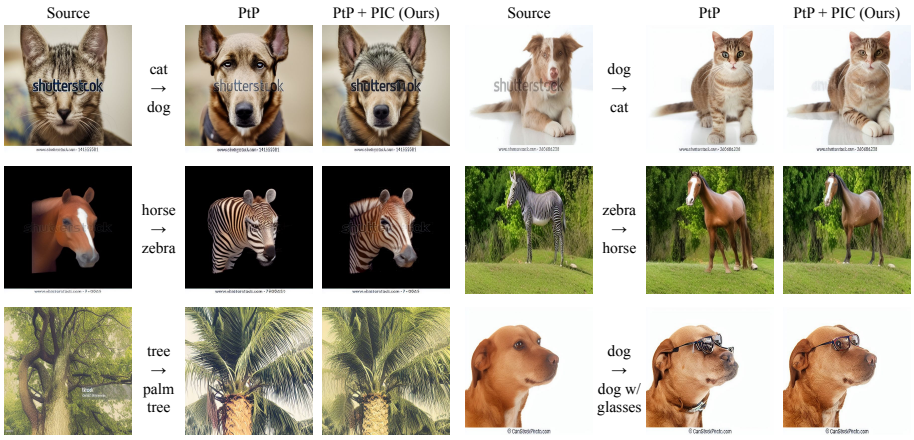

Fig. 12: Additional qualitative results of Prompt-to-Prompt [4] and its integration with the proposed method on data sampled from the LAION-5B dataset [10] using the pretrained Stable Diffusion [9].

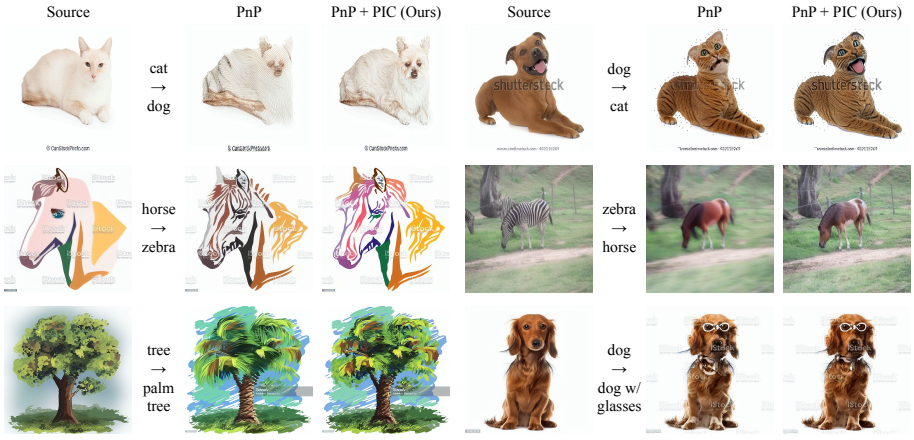

Fig. 13: Additional qualitative results of Plug-and-Play [12] and its integration with the proposed method on data sampled from the LAION-5B dataset [10] using the pretrained Stable Diffusion [9].

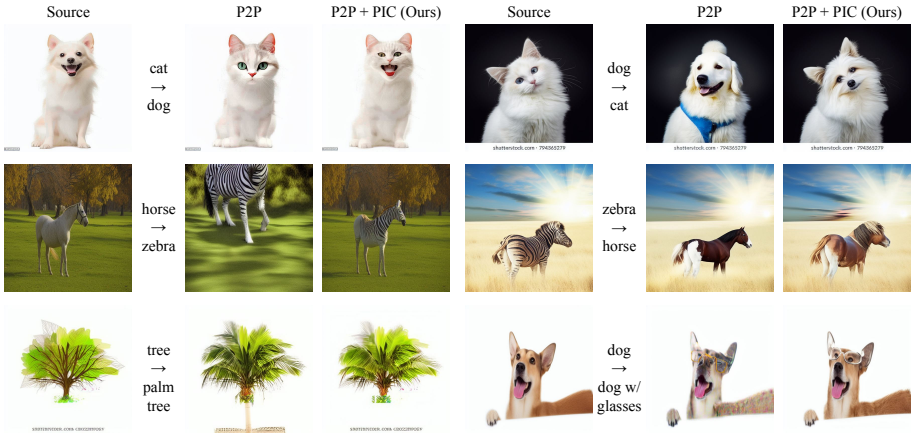

Fig. 14: Additional qualitative results of Pix2Pix-Zero [8] and its integration with the proposed method on data sampled from the LAION-5B dataset [10] using the pretrained Stable Diffusion [9].

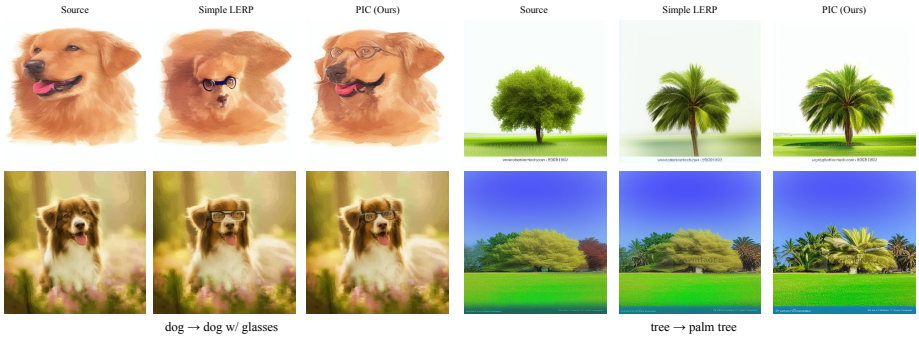

Fig. 15: Qualitative Comparison between PIC and PIC with simple LERP.

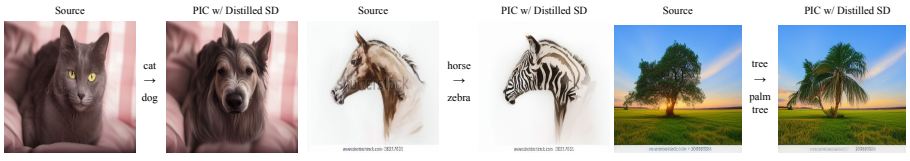

Fig. 16: Qualitative results of PIC using Distilled SD.