Supplementary Material for UDT:Unsupervised Discovery of Transformations between Fine-Grained Classes in Diffusion Models

A Details of Baselines

NoiseCLR [4]. To discover interpretable directions, NoiseCLR utilizes a contrastive learning approach within diffusion models by comparing feature representations. Nevertheless, the method is challenged when performing transformations between fine-grained classes, such as different animal breeds, and it provides limited command over specific visual details. This issue arises from its lack of a hierarchically structured latent space, a factor that impedes the discovery of nuanced directions. For training, NoiseCLR used the same image dataset as our method, and its hyperparameters were also identical; $K = 100$ directions were trained using $N = 100$ unique images from each domain (one image per direction). Furthermore, to ensure a fair comparison, the results were generated using the same pipeline and hyperparameters as our method.

Concept Discovery [15]. This method discovers generative directions in an entirely unsupervised manner by decomposing a collection of unlabeled images into a set of compositional concepts using a pretrained text-to-image model. However, its primary limitation is that the scope of the discovered directions is strictly confined to the concepts present within the training dataset. Consequently, unlike methods such as NoiseCLR or our proposed approach which can identify more generalized semantic directions, Concept Discovery is less effective at exploring concepts beyond its immediate training distribution. For our training, we used the Stanford dataset, from which we selected 5 images each for the following four breeds—Otterhound, Toy Poodle, Entle Bucher, and Silky Terrier—focusing on photos where the dog was the main subject. To ensure a fair comparison, the model was trained for 3000 iterations with an effective batch size of 16, and its results were generated using the same pipeline as our method.

Interpret Diffusion [14]. Interpret Diffusion proposes a self-supervised method to discover interpretable directions for a specific concept in the latent space of text-to-image diffusion models. The technique optimizes a concept vector by first generating an image with a

* Equal contribution. † Corresponding author.

prompt containing a specific concept (e.g., 'Toy Poodle'), and then reconstructing the original image using a general prompt from which the concept has been removed (e.g., 'dog'). Following this methodology, the experiment was configured to learn the unique direction for each dog breed by synthesizing 1,000 images for it. With the pre-trained diffusion model frozen, training only the concept vector to minimize reconstruction loss forces the vector to represent the information missing from the text—namely, the unique characteristics that distinguish specific breeds, such as an Otterhound or an Entlebucher. The vectors discovered this way can find any desired concept without external classifiers or labeled data, and in the original study, they were used for responsible image generation, including fairness and safety.

Null-Text Inversion [18] & LEDITS++ [1]. In order to compare our method with established editing baselines, we utilize LEDITS++ and Null-Text Inversion. Both are prominent training-free editing techniques. For our experiments, we first apply DDIM inversion to a source image and then guide the transformation using a specific editing prompt; for LEDITS++, we set the `edit guidance scale` to 7.5 and the `edit threshold` to 0.75. The set of prompts used for each dog breed is as follows:

- Otterhound: `"a photo of an Otterhound dog"`
- Toy Poodle: `"a photo of a Toy Poodle dog"`
- Entle Bucher: `"a photo of a Entle Bucher dog"`
- Silky Terrier: `"a photo of a Silky Terrier dog"`

The effectiveness of these methods is underscored by recent findings. According to [10], which proposes the LMM Score as a metric for evaluating image edits, a higher score indicates a more successful modification. In their evaluation, both LEDITS++ and Null-Text Inversion ranked within the top three, demonstrating their state-of-the-art performance.

B Additional Experiments

Prompt-Guided Image Transformation. We evaluated `UDT`'s capability to transform dog images based on textual prompts that specify various attributes such as poses and artistic styles. Figure S.1 visualizes that our model successfully transforms input images into diverse breed variations while accurately reflecting the semantics specified in the text prompts. For pose control, the prompt *"A running dog"* not only changes the input to different breeds but also renders all of them in dynamic running poses. Similarly, *"A sitting dog"* generates various breeds, all consistently shown in sitting positions. This demonstrates that our model can handle both breed-level and pose-level transformations jointly. Furthermore, `UDT` successfully combines breed transformation with artistic style transfer. The prompt *"A drawing of a dog"* produces different dog breeds rendered in pencil-sketch style, while *"A Van Gogh dog"* generates breed variations that all exhibit Van Gogh's characteristic brushstrokes and color palette. These results demonstrate that our model's disentangled representations enable simultaneous control over multiple attributes—changing breeds while applying consistent stylistic or pose modifications across all outputs.

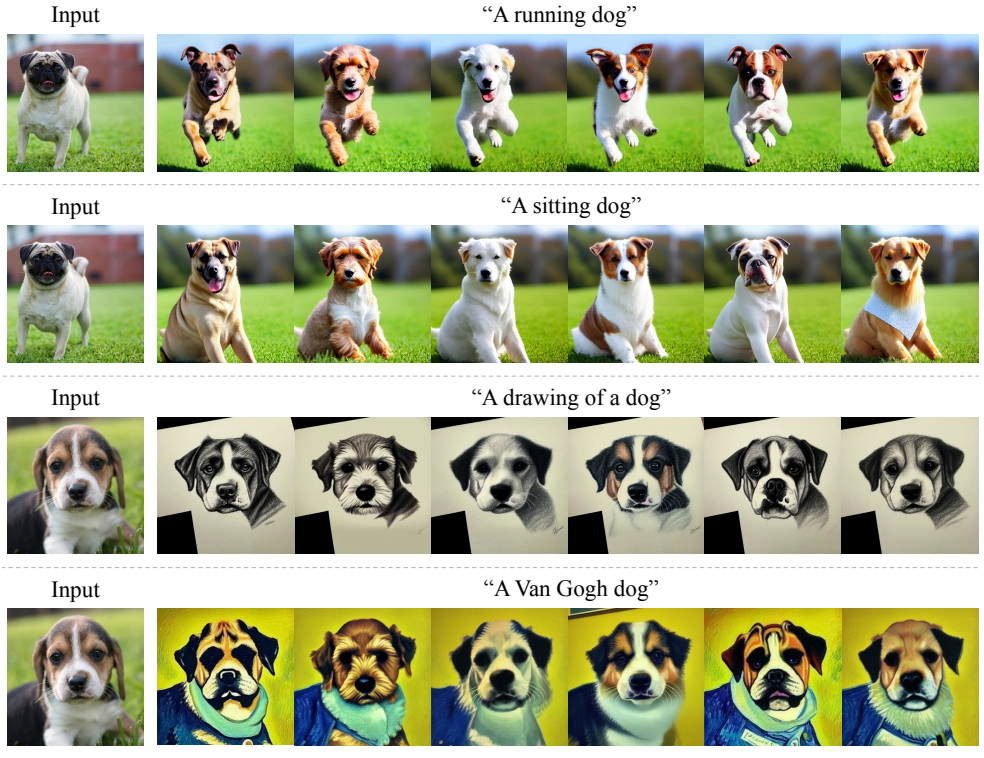

Figure S.1: **Qualitative results for prompt-based transformation.** Examples include transformations for poses (*e.g.*, running, sitting) and artistic styles (*e.g.*, "A Van Gogh dog").

Effect of Timesteps. The choice of timesteps significantly impacts the semantic disentanglement in diffusion models, as shown in previous studies [4, 28]. As illustrated in Figure S.2, early timesteps primarily govern structural attributes, whereas later timesteps refine fine-grained details. Consequently, for effective breed-to-breed transformation, including early timestep ranges like 5-15 or 10-20 is crucial. We found that the 10-20 interval optimally preserves the input's pose and spatial coherence. Conversely, later timesteps are more influential in enhancing identity or breed preservation by refining detailed features.

Face Transformation: Comparison with NoiseCLR. We compare UDT with NoiseCLR [4] and its variant, NoiseCLR*, with qualitative results shown in Figure S.3. For a fair comparison, NoiseCLR* is generated using the same editing timestep (10-30) as UDT. The figure is split into two parts: (a) a comparison of facial attribute transformation between NoiseCLR and NoiseCLR*, and (b) a comparison of identity transformation performance between UDT and NoiseCLR*. In Fig. S.3 (a), NoiseCLR often disrupts identity consistency when editing attributes (e.g., applying lipstick to a male face). Similarly, the "Red Nose" direction in NoiseCLR* reddens the entire face, indicating that the learned direction is entangled with a specific identity. In contrast, as shown in Fig. S.3 (b), UDT preserves class-extrinsic properties of the source (e.g., pose) while modifying core identity-defining factors such as hairstyle, eye shape, and facial structure, outperforming NoiseCLR*. This reflects stronger semantic

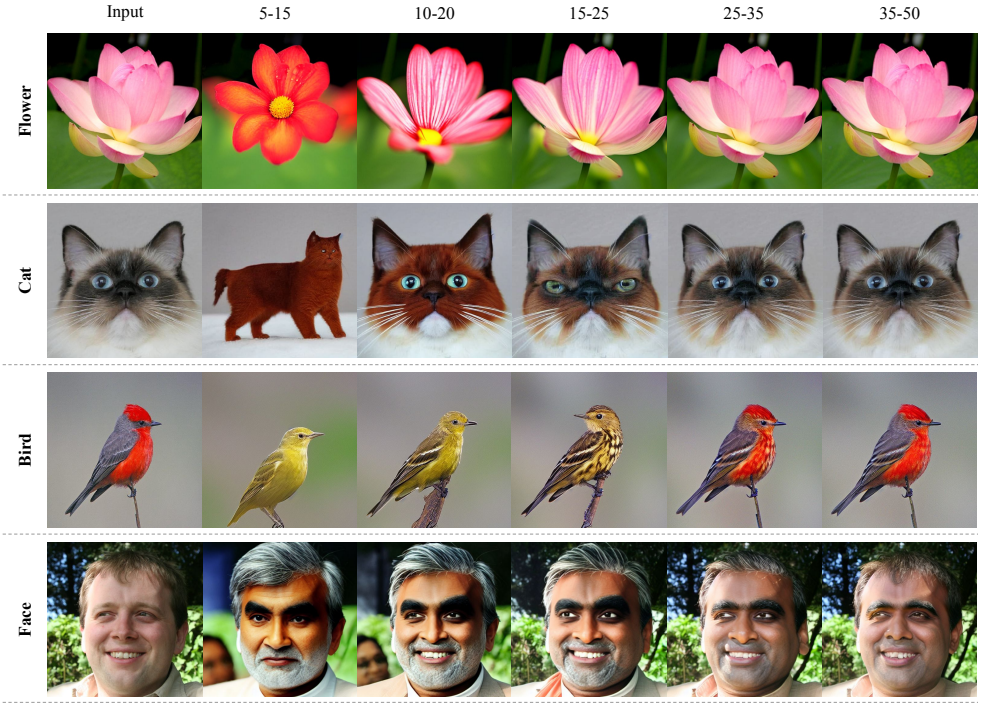

Figure S.2: **Impact of editing timesteps.** Early intervals are crucial for structural transformations, while later timesteps are better for preserving fine-grained details and identity.

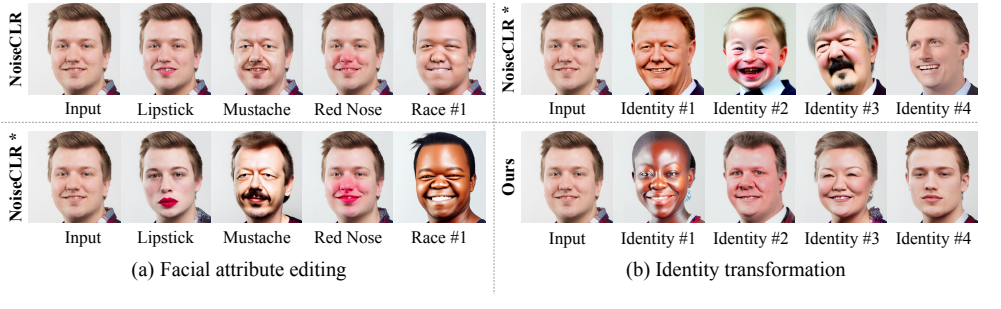

Figure S.3: **Qualitative comparison on face transformation.** We provide a qualitative comparison between our method, UDT, and the baseline NoiseCLR [4] on face manipulation tasks. (a) Facial attribute transformation: A comparison between NoiseCLR and NoiseCLR* using the same direction. NoiseCLR* is generated using the same editing timestep (t=10–30) UDT. (b) Identity transformation: A comparison between UDT and NoiseCLR*.

disentanglement across genders and ethnicities. Overall, UDT discovers more numerous translation directions and consistently yields semantically disentangled transformations.

C Additional Ablation Studies

Effect of Loss Function. To quantitatively evaluate the effectiveness of UDT in capturing breed-specific characteristics, we follow the experimental setting described in Section 5.2. Table S.1 summarizes these findings, with target breeds exhibiting the highest performance improvements highlighted in **bold**. The results clearly indicate the importance of each loss component. The variant without the contrastive loss ("w/o $\mathcal{L}_{\mathcal{T}}$") consistently achieves lower classification accuracy, implying that this component is critical for breed-specific transformations. Cases marked by †, which failed to identify meaningful semantic directions, further reinforce the essential role of $\mathcal{L}_{\mathcal{T}}$. Meanwhile, the variant without the regularization loss ("w/o $\mathcal{L}_{\text{Reg}}$") shows consistently positive yet lower performance, confirming the importance of $\mathcal{L}_{\text{Reg}}$ in maintaining attribute fidelity. In contrast, our complete UDT framework, which integrates both $\mathcal{L}_{\mathcal{T}}$ and $\mathcal{L}_{\text{Reg}}$, accurately achieves the desired transformations and delivers superior quantitative results. These findings validate that our CLIP-based [23] feature guidance effectively facilitates robust semantic transformation.

	Boxer			**Beagle**			**Golden Retriever**			**Siberian Husky**			**Miniature Poodle**		
	w/o $\mathcal{L}_{\mathcal{T}}$	w/o $\mathcal{L}_{\text{Reg}}$	Ours	w/o $\mathcal{L}_{\mathcal{T}}$	w/o $\mathcal{L}_{\text{Reg}}$	Ours	w/o $\mathcal{L}_{\mathcal{T}}$	w/o $\mathcal{L}_{\text{Reg}}$	Ours	w/o $\mathcal{L}_{\mathcal{T}}$	w/o $\mathcal{L}_{\text{Reg}}$	Ours	w/o $\mathcal{L}_{\mathcal{T}}$	w/o $\mathcal{L}_{\text{Reg}}$	Ours
Boxer	-24.74	3.15	**15.08**	-27.06	7.32	8.72	-27.06	0.77	-4.34	-22.07	9.82	8.04	-19.02	8.22	-3.86
Beagle	—†	10.54	2.55	—†	19.45	**20.75**	—†	10.24	2.26	—†	-6.83	0.50	—†	0.21	4.08
Golden Retriever	-26.10	7.49	-5.68	-22.08	12.29	5.32	-23.43	27.18	**43.24**	-20.38	-3.57	12.42	-15.12	17.80	8.17
Siberian Husky	-15.06	-8.77	-9.15	-9.79	-4.98	-7.63	-10.77	-6.59	1.85	-5.09	45.85	**54.77**	-6.69	-7.82	-0.71
Miniature Poodle	—†	-6.05	-3.33	—†	-4.04	-6.22	—†	8.81	21.57	—†	8.77	7.56	—†	39.45	**50.38**

Table S.1: **Ablation study on impact of breed-to-breed translation on classification.** We compare three configurations: without regularization loss (w/o $\mathcal{L}_{\text{Reg}}$), without contrastive loss (w/o $\mathcal{L}_{\mathcal{T}}$), and our full method (Ours). –†: Failed to identify translation directions.

Influence of Child-Class Prompts. To investigate the influence of child-class prompts, we compare three methods: (1) NoiseCLR [4]; (2) UDT; and (3) UDT w/ C. prompt, which augments our method with a target-breed text prompt. As shown in Table S.2, the UDT w/ C. prompt variant achieves the highest classification scores, confirming that explicit text prompts can effectively steer transformations. However, our core method, UDT, also demonstrates strong performance, lagging by a relatively modest margin. For instance, in the Golden Retriever and Siberian Husky transformations, the performance is highly competitive. This highlights that our unsupervised approach robustly captures core breed-level features without requiring explicit textual guidance, even though prompts can further sharpen breed specificity.

	Boxer			**Beagle**			**Golden Retriever**			**Siberian Husky**			**Miniature Poodle**		
	NoiseCLR	w/ C. prompt	Ours	NoiseCLR	w/ C. prompt	Ours	NoiseCLR	w/ C. prompt	Ours	NoiseCLR	w/ C. prompt	Ours	NoiseCLR	w/ C. prompt	Ours
Boxer	-13.90	**24.07**	11.20	-4.83	11.28	9.35	0.29	1.07	0.57	1.30	7.49	11.68	1.43	-3.35	-6.43
Beagle	-16.35	2.73	4.83	-6.94	**40.53**	16.52	-12.33	6.62	3.35	-3.16	3.78	8.91	-3.44	2.35	0.63
Golden Retriever	-8.12	-11.30	-6.24	-1.67	-0.56	2.98	-1.59	**40.64**	40.07	5.73	8.39	12.70	2.02	2.65	3.29
Siberian Husky	-14.10	-6.75	-11.63	-17.15	7.17	-7.58	-13.45	2.87	0.95	-8.56	**61.23**	51.27	-4.86	4.23	-0.60
Miniature Poodle	-20.45	-11.43	-8.86	-16.36	-7.82	-4.92	-17.38	7.74	17.79	-5.39	7.01	6.90	3.18	**67.28**	49.69

Table S.2: **Ablation study on the effect of child-class prompts.** Scores represent the percentage point shift in classification probability for transformations toward each target breed. We compare a NoiseCLR [4], Ours (UDT), and our method augmented with a text prompt(UDT w/ C. prompt).

Ablations on K. The hyperparameter K controls the number of learned translation directions in our framework. We conducted ablations with $K \in \{10, 50, 100, 120\}$, to analyze its effect on transformation quality, as illustrated in Figure S.4. We observe that smaller values of K provide insufficient capacity for diverse transformations. For example, at K=10, the model learns only broad, coarse-grained changes and often introduces unintended transformations. K=50 enabled some fine-grained transformation (e.g., breed changes), but inconsistency persisted. At K=100, directions were more clearly disentangled and behaved robustly across diverse images. Further increasing to K=120 provides no noticeable improvement in visual quality while increasing computational overhead.

Figure S.4: **Ablation study of K.** We show transformation results with varying K values (10, 50, 100, 120). The number of learned directions K directly impacts the quality of transformations.