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

Youngjae Choi*
yj951118@soongsil.ac.kr
Hyunseo Koh*
deepvelop@soongsil.ac.kr
Hojae Jeong*
98eddiechung@soongsil.ac.kr
Byungkwan Chae*
silencechae75@soongsil.ac.kr
Sungyong Park
ejqdl@soongsil.ac.kr
Heewon Kim†
hwkim@ssu.ac.kr

Soongsil University
Seoul, Republic of Korea

Abstract

Diffusion models achieve impressive image synthesis, yet unsupervised methods for latent space exploration remain limited in fine-grained class translation. Existing approaches struggle with fine-grained class translation, often producing low-diversity outputs within parent classes or inconsistent child-class mappings across images. We propose UDT (**U**nsupervised **D**iscovery of **T**ransformations), a framework that incorporates hierarchical structure into unsupervised direction discovery. UDT leverages parent-class prompts to decompose predicted noise into class-general and class-specific components, ensuring translations remain within the parent domain while enabling disentangled child-class transformations. A hierarchy-aware contrastive loss further enforces consistency, with each direction corresponding to a distinct child class. Experiments on dogs, cats, birds, and flowers show that UDT outperforms state-of-the-art methods both qualitatively and quantitatively. Moreover, UDT supports controllable interpolation, allowing for the smooth generation of intermediate classes (*e.g.*, mixed breeds). These results demonstrate UDT as a general and effective solution for fine-grained image translation. Our project website is available at: https://ssu-reality-lab.github.io/UDT.

1 Introduction

Diffusion models [9, 24] have become central to modern generative modeling, offering high-fidelity image synthesis in a wide range of applications. Among them, text-to-image generation [5, 18, 22, 28, 29] has been particularly impactful, allowing the translation of natural

* Equal contribution. † Corresponding author.

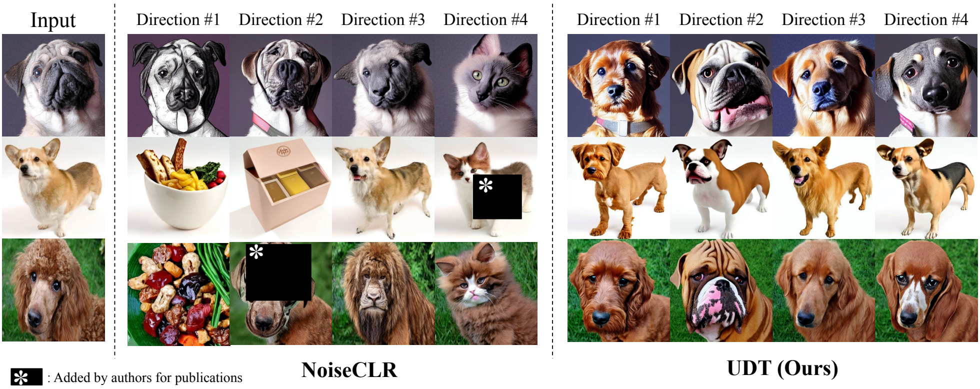

Figure 1: **Comparison of unsupervised translation methods**. Input images are translated along four directions that have been discovered. NoiseCLR [4] often produces limited breed diversity in the first row of Direction #1∼#3 or inconsistent changes across input images in each direction. In contrast, `UDT` achieves diverse and coherent fine-grained class transformations, consistently altering breed-defining traits while preserving pose and parent-class attributes. The parent class is dog, and the child classes are breeds in this example.

language prompts into compelling visual content. Although the limited availability of labeled data restricts supervised approaches, unsupervised methods have gained increasing attention for their potential to unlock greater diversity and generalization in generation tasks.

Recent unsupervised approaches for diffusion models can be categorized into two major groups: exploring intermediate features of the U-Net [6, 14, 21] and operating in the predicted noise space [4, 15]. Both approaches attempt to discover semantic directions (*i.e.*, condition vectors that replace textual guidance) to modify attributes of generated images. These methods have shown success in altering high-level semantic regions (*e.g.*, facial expressions, color, texture), validating the potential of unsupervised latent space exploration.

However, despite these advances, unsupervised methods face notable shortcomings when applied to fine-grained class translation. As illustrated in Figure 1, existing approaches struggle with two fundamental issues. First, the diversity of generated outputs is insufficient: discovered directions often tend to generate only low-diversity variations within a parent class (*e.g.*, different types of dog) or drift into unrelated classes (*e.g.*, cat, food). Second, discovered directions lack consistency across images: the same direction may correspond to different child classes depending on the input, requiring users to search for a desirable transformation manually. These limitations limit their applicability in scenarios that require reliable, fine-grained control, such as breed-to-breed transformations.

To overcome these challenges, we propose a novel method called `UDT` (**U**nsupervised **D**iscovery of **T**ransformations), that introduces a hierarchical structure into the fine-grained class translation process. Specifically, `UDT` employs parent-class information to decompose predicted noise divergences into two components: a parent-class vector that encodes general attributes (*e.g.*, 'dog') and a child-class vector that captures fine-grained variations (*e.g.*, 'Poodle'). By applying hierarchy-aware contrastive learning, we ensure that each discovered direction consistently maps to a distinct child class while preventing drift outside the parent domain.

Our contributions are summarized as follows:

- **Hierarchical latent structuring:** We introduce a decomposition of predicted noise guided by parent-class information, enabling fine-grained class translations.
- **Hierarchy-aware contrastive learning:** We design a loss function that enforces each direction to a distinct child class, improving interpretability and diversity.
- **Extensive validation:** Experiments across multiple fine-grained domains demonstrate that `UDT` significantly outperforms prior methods in qualitative and quantitative comparisons, producing coherent, diverse, and controllable transformations.

By explicitly incorporating hierarchical knowledge into unsupervised latent exploration, `UDT` establishes a new paradigm for fine-grained class translation in diffusion models, bridging the gap between semantic attribute editing and consistent class-level transformations.

2 Related Work

Latent Space Exploration of Diffusion Models. Diffusion models encode rich semantics in their latent spaces, enabling controllable image manipulation by identifying edit directions [3, 17]. One major line of work for discovering these directions has focused on analyzing the intermediate features within the U-Net bottleneck. This category includes both guided approaches, which rely on supervision from vision-language models like CLIP [13] or pre-trained classifiers [6], and unsupervised approaches that analyze the features' internal structure through methods like PCA [6], Jacobian analysis [21], or self-supervised learning [14]. However, a limitation of these methods is their difficulty with class transformations, as they often only capture attribute-level translations. In contrast, another line of work explores the predicted noise space. This includes methods such as Concept Discovery [15], which discovers compositional concepts, and NoiseCLR [4], which uses contrastive learning to identify interpretable directions. While these approaches [4, 15] have advanced the discovery of semantic concepts, they remain insufficient for fine-grained class translation, where consistent and diverse transformations are required. In contrast, our hierarchical framework decomposes predicted noise, allowing for the structured exploration of transformations.

Contrastive Learning in Generative Models. The principle of contrastive learning, introduced by Hadsell et al. [7] to learn invariant features by pulling similar samples closer and pushing dissimilar ones apart, has been widely adopted in areas like data augmentation [2, 20], and diverse scene generation [26]. This principle has been adapted to find meaningful semantic directions in generative models, such as GANs [30] and diffusion models [4]. We reformulate contrastive learning for fine-grained class translation, enforcing diffusion models to produce clearly distinguishable images across fine-grained classes.

3 Preliminary

3.1 Denoising Probabilistic Diffusion Models

Diffusion models [9, 25] generate data through an iterative denoising process, often referred to as the reverse process. The denoising network ε_θ is trained to predict the noise ε from a noised latent code x_t. Here, x_t is obtained by corrupting the original latent x_0 with Gaussian noise $\varepsilon \sim \mathcal{N}(0,1)$ at timestep t, following a predefined noise schedule β_t. We define $\alpha_t = 1-\beta_t$, and the cumulative product as $\bar{\alpha}t = \prod i = 1^t \alpha_i$. The training objective becomes:

$$\mathcal{L}_{\mathrm{DM}} = \mathbb{E}_{x_0,\varepsilon\sim\mathcal{N}(0,1),t}\left[\|\varepsilon - \varepsilon_\theta(x_t,t)\|_2^2\right]. \tag{1}$$

In the deterministic DDIM[25] reverse process, the model $\varepsilon_\theta(x_t,t,c)$ first predicts the noise from the current latent x_t. This allows us to estimate the clean latent $\hat{x}_0$. The less noisy latent x_{t-1} is computed by combining the estimated clean latent and the predicted noise:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}} \cdot \varepsilon_\theta(x_t,t,c). \tag{2}$$

This iterative step ensures a consistent generation process guided by the text condition c.

3.2 Classifier-free Guidance

Classifier-free guidance [8] allows conditioned sampling by modifying the noise prediction $\varepsilon_\theta(x_t)$ to incorporate conditioning c. The guided noise estimate is computed as:

$$\tilde{\varepsilon}_\theta(x_t,c) = \varepsilon_\theta(x_t,\phi) + \lambda_g(\varepsilon_\theta(x_t,c) - \varepsilon_\theta(x_t,\phi)), \tag{3}$$

where ϕ is the null condition and λ_g is the guidance scale. Leveraging the adaptability of this framework, our image translation pipeline substitutes the text condition c with a learned semantic direction c_k to translate the image.

4 Proposed Method

In this section, we describe how our method discovers interpretable directions for breed translation in an unsupervised setting using a dataset $X = \{x^1,\ldots,x^N\}$, where N represents the number of images. The number of discovered directions is controlled by a hyperparameter K, and the set of directions is denoted as $C = \{c_1,\ldots,c_K\}$. We assume that the dataset X possesses a hierarchical structure, consisting of a parent class (*e.g.* dog, cat, and bird) and an unknown number of child classes.

4.1 Hierarchy-Aware Feature Divergence

The k-th interpretable direction c_k is a learned weight vector that replaces the condition feature c in Eq. (3). To learn these directions, NoiseCLR [4] employs a divergence in the predicted noise, conditioned by a null-text prompt ϕ:

$$\Delta\varepsilon_k^n = \varepsilon_\theta(x_t^n,c_k) - \varepsilon_\theta(x_t^n,\phi). \tag{4}$$

This feature divergence, $\Delta\varepsilon_k^n$, struggles with breed transformations because its discovery scope is too broad, often leading to the trivial translations or nonsensical failures shown for NoiseCLR in Figure 1. Since it measures the total deviation from the null condition ϕ, its scope often extends beyond the parent-class distribution of $\varepsilon_\theta(x_t^n,p)$. To address this limitation, we decompose this divergence based on a parent class prompt p as follows:

$$\Delta\varepsilon_k^n = \underbrace{\varepsilon_\theta(x_t^n,c_k) - \varepsilon_\theta(x_t^n,p)}_{\Delta\mathcal{T}_k^n} + \underbrace{\varepsilon_\theta(x_t^n,p) - \varepsilon_\theta(x_t^n,\phi)}_{\Delta\mathcal{P}_k^n}. \tag{5}$$

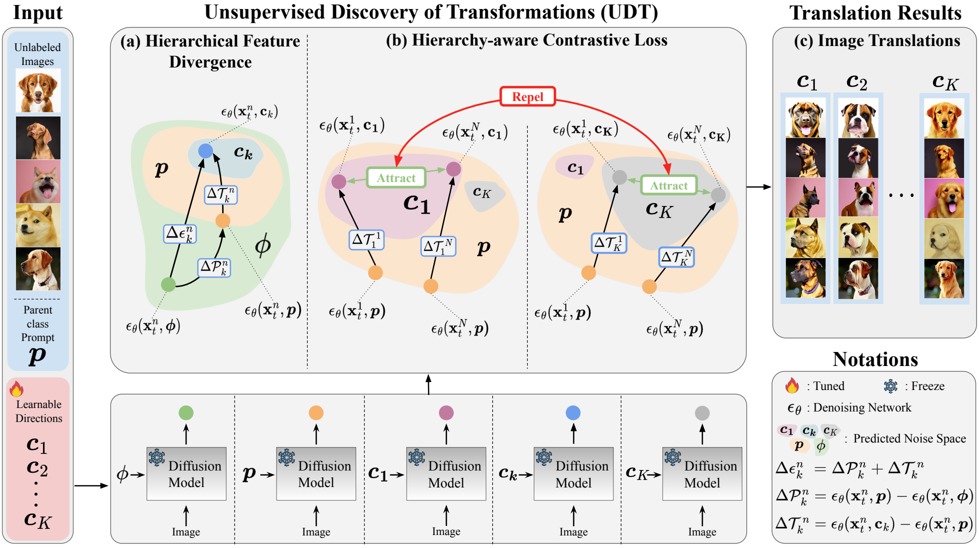
Input
Unlabeled Images
Parent class Prompt
p
Learnable Directions
c_1
c_2
c_K
Unsupervised Discovery of Transformations (UDT)
(a) Hierarchical Feature Divergence
$\epsilon_\theta(\mathbf{x}_t^n, \mathbf{c}_k)$
p
c_k
$\Delta\mathcal{T}_k^n$
$\Delta\epsilon_k^n$
$\Delta\mathcal{P}_k^n$
ϕ
$\epsilon_\theta(\mathbf{x}_t^n, \phi)$
$\epsilon_\theta(\mathbf{x}_t^n, \boldsymbol{p})$
(b) Hierarchy-aware Contrastive Loss
Repel
Attract
$\epsilon_\theta(\mathbf{x}_t^1, \mathbf{c}_1)$
$\epsilon_\theta(\mathbf{x}_t^N, \mathbf{c}_1)$
c_1
c_K
p
$\Delta\mathcal{T}_1^1$
$\Delta\mathcal{T}_1^N$
$\epsilon_\theta(\mathbf{x}_t^1, \boldsymbol{p})$
$\epsilon_\theta(\mathbf{x}_t^N, \boldsymbol{p})$
$\epsilon_\theta(\mathbf{x}_t^1, \mathbf{c}_K)$
$\epsilon_\theta(\mathbf{x}_t^N, \mathbf{c}_K)$
$\Delta\mathcal{T}_K^1$
$\Delta\mathcal{T}_K^N$
Translation Results
(c) Image Translations
c_1 c_2 ... c_K
ϕ → Diffusion Model
p → Diffusion Model
c_1 → Diffusion Model
c_k → Diffusion Model
c_K → Diffusion Model
Image
Notations
: Tuned
: Freeze
ϵ_θ : Denoising Network
c_1 c_k c_K p ϕ : Predicted Noise Space
$\Delta\epsilon_k^n = \Delta\mathcal{P}_k^n + \Delta\mathcal{T}_k^n$
$\Delta\mathcal{P}_k^n = \epsilon_\theta(\mathbf{x}_t^n, \boldsymbol{p}) - \epsilon_\theta(\mathbf{x}_t^n, \phi)$
$\Delta\mathcal{T}_k^n = \epsilon_\theta(\mathbf{x}_t^n, \mathbf{c}_k) - \epsilon_\theta(\mathbf{x}_t^n, \boldsymbol{p})$

Figure 2: **Overview of the UDT framework.** (a) UDT decomposes predicted noise divergence $\Delta\epsilon_k^n$ into a parent-class component $\Delta\mathcal{P}_k^n$ (general attributes) and a child-class component $\Delta\mathcal{T}_k^n$ (fine-grained traits). (b) Contrastive learning is then applied only to the child-class vectors, ensuring each discovered direction corresponds to a consistent child class. This hierarchical formulation enables UDT to discover interpretable directions for fine-grained class translation, later used in the image translation pipeline (c).

where p is a non-trainable text condition of the parent class (*e.g.*, dog). This decomposition allows our method to isolate the signal required for fine-grained control, as illustrated in Figure 2 (a). The first component, $\Delta\mathcal{T}_k^n$, is the hierarchy-aware feature divergence, which represents the child class vector that distinguishes a specific child class from its parent. In contrast, the second component, $\Delta\mathcal{P}_k^n$, is a parent-class vector that isolates the broad attributes of the parent class itself.

4.2 Loss Function with Hierarchical Information

Hierarchy-aware Contrastive Loss. Contrastive learning aims to discover interpretable directions by defining positive and negative divergences using $\Delta\mathcal{T}$ in Eq. (5). For a given target divergence $\Delta\mathcal{T}_j^a$, which corresponds to the a-th image and the j-th direction, positive divergences are those that share the same direction c_j but come from different images. In contrast, negative divergences share the same image x^a but correspond to different directions. Contrastive loss encourages similarity in positive divergences while pushing negative divergences to be dissimilar, a process visually summarized in Figure 2 (b). The loss function is defined as:

$$\mathcal{L}_\mathcal{T} = -\log \frac{\sum_{a=1}^{|X'|}\sum_{b=1}^{|X'|} \mathbf{1}[a \neq b] \exp\left(\mathrm{sim}(\Delta\mathcal{T}_j^a, \Delta\mathcal{T}_j^b)\right)}{\sum_{a=1}^{|X'|}\sum_{i=1}^{|C'|} \mathbf{1}[i \neq j] \exp\left(\mathrm{sim}(\Delta\mathcal{T}_j^a, \Delta\mathcal{T}_i^a)\right)}, \quad (6)$$

where $X' \subset X$ is a subset of the image dataset and $C' \subset C - \{c_j\}$ is a subset of the interpretable directions, excluding c_j. The similarity function $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity, given by:

$$\mathrm{sim}(\Delta\mathcal{T}_1, \Delta\mathcal{T}_2) = \frac{\Delta\mathcal{T}_1 \cdot \Delta\mathcal{T}_2}{\|\Delta\mathcal{T}_1\|\|\Delta\mathcal{T}_2\|}. \tag{7}$$

Regularization Loss. The contrastive loss in Eq. (6) fluctuates significantly during training, as the hierarchy-aware feature divergence $\Delta\mathcal{T}$ alters the overall appearance of images. To enhance training stability, we introduce a regularization loss using $\Delta\varepsilon$ from Eq. (4), encouraging positive divergences to be closer as follows:

$$\mathcal{L}_{\mathrm{Reg}} = -\log \sum_{a=1}^{|X'|} \sum_{b=1}^{|X'|} \mathbf{1}[a \neq b] \exp\left(\mathrm{sim}(\Delta\varepsilon_j^a, \Delta\varepsilon_j^b)\right). \tag{8}$$

Total Loss Function. The overall loss function to learn interpretable directions C becomes:

$$\mathcal{L} = \mathcal{L}_{\mathcal{T}} + \lambda_{\mathrm{Reg}} \mathcal{L}_{\mathrm{Reg}}, \tag{9}$$

where $\mathcal{L}_{\mathcal{T}}$ is the contrastive loss in Eq. (6), $\mathcal{L}_{\mathrm{Reg}}$ is the regularization loss in Eq. (8), and λ_{Reg} is a hyperparameter that balances two terms.

4.3 Image Translation Pipeline

To modify an image I, the initial latent x_T is obtained through DDIM inversion [18], which anchors the subsequent denoising trajectory to the input content. At each timestep t, the diffusion process is then guided along a learned direction c_k by the following estimate:

$$\tilde{\varepsilon}_\theta(x_t, c_k) = \varepsilon_\theta(x_t, \phi) + \lambda_e(\varepsilon_\theta(x_t, c_k) - \varepsilon_\theta(x_t, \phi)), \tag{10}$$

where λ_e controls the editing strength. The guided estimate $\tilde{\varepsilon}_\theta$ is applied to the denoising process to transform the image to the target child class, leading to the results shown in Figure 2 (c).

5 Experiment

5.1 Experimental Setup

Datasets. We used standard fine-grained benchmarks: Flowers102 [19] (102 species), Stanford Dogs [12] (120 breeds, 20,580 images), CUB-200-2011 [27] (200 bird species, 11,788 images), and FFHQ [11] (70,000 faces). Additionally, we utilized a custom cat dataset comprising 100 breeds, each represented by a single internet-sourced image.

Implementation Details. We used Stable Diffusion v1.5 as the backbone for all experiments. The model is trained with hyperparameters $K = 100$, $N = 100$, and $\lambda_{\mathrm{Reg}} = 1$. Following NoiseCLR [4], AdamW [16] is used with a learning rate of 10^{-3} and a batch size of 6. Training each category takes 10–13 minutes on a single NVIDIA A6000 GPU (48GB), while inference generates transformed images in under 10 seconds across domains.

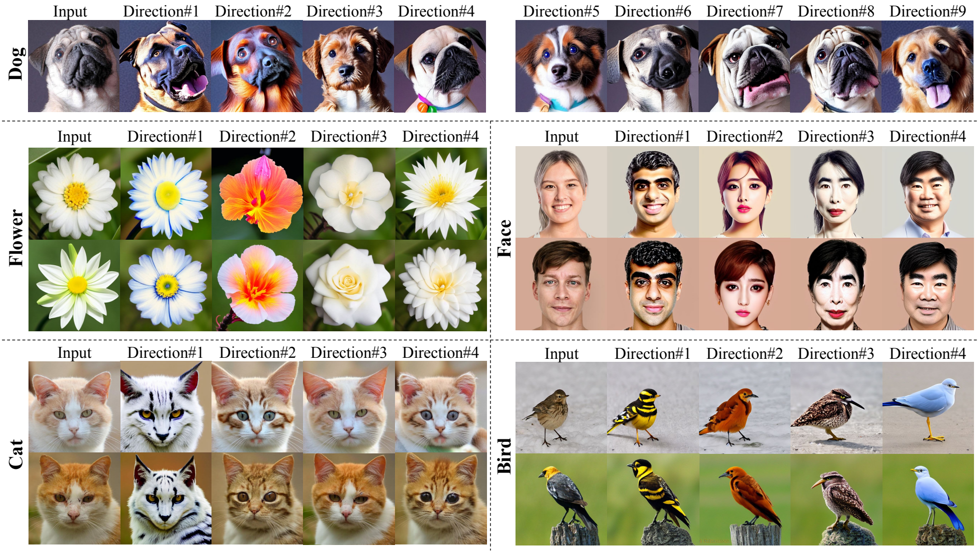

Figure 3: **Visualizing discovered transformations by `UDT`.** `UDT` discovers a diverse set of interpretable transformations within a single category, and generalizes this ability across various domains such as dogs, flowers, faces, cats, and birds.

5.2 Experimental Results

Qualitative Results with `UDT`. Figure 3 demonstrates that `UDT` discovers semantically distinct transformation directions across diverse domains—from animals to flowers and human faces. Each discovered direction captures unique and interpretable attributes, making it clearly distinguishable from other directions within the same domain. For instance, in the dog domain, `UDT` successfully discovers numerous distinct breed transformations. "Direction #4" produces *Bulldog* features with characteristic wrinkled faces and robust builds, while "Direction #9" generates *Golden Retriever* traits, including flowing golden coats. Similar capabilities are observed in other animal categories; in cats, `UDT` accurately alters breed-defining traits such as fur color ("Direction #1") and ear length ("Direction #2"). Beyond animals, `UDT` captures diverse semantic variations across different domains. For human faces, "Direction #4" produces *Asian* facial features, whereas "Direction #1" yields a more *Western* appearance. These results demonstrate that `UDT` discovers semantically meaningful features without any explicit labels during training, highlighting its capability to identify distinctive visual variations within each domain.

Qualitative Comparison with SotA Methods. Figure 4 presents qualitative comparisons between `UDT` and state-of-the-art methods on unsupervised learning (NoiseCLR [4] and Concept Discovery [15]), self-supervised learning (Interpret Diffusion [14]), and image editing (LEDITS++ [1] and Null-Text [18]). To evaluate the performance of unsupervised methods [4, 15], a CLIP classifier [23] was used to verify whether the translated images were correctly predicted as the intended target classes. NoiseCLR and Concept Discovery struggle to find the transformation directions that match the target classes of Silky Terrier

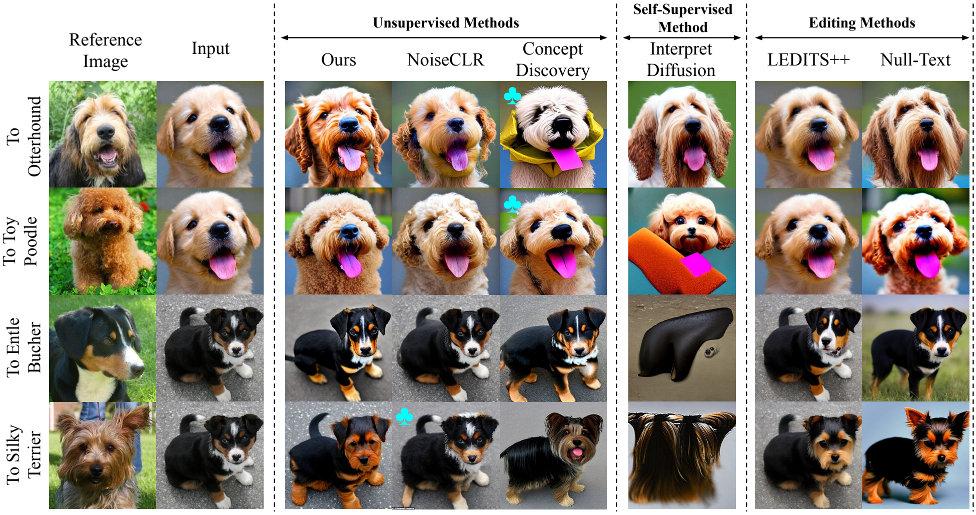

Figure 4: **Qualitative comparisons**. `UDT` successfully finds target breed transformations where other unsupervised methods often fail. Additionally, `UDT` outperforms editing methods by more accurately representing fine details, such as the Toy Poodle's curly texture, and preserving the original pose. ♣ indicates that the model failed to discover a direction for the target breed; in these cases, the images are visualized with the highest CLIP scores, regardless of class labels.

(NoiseCLR) and Otterhound and Toy Poodle (Concept Discovery). Additionally, the self-supervised method, Interpret Diffusion, does not provide adequate directions for dog transformations. In comparison, our method not only successfully transforms images to the target breed but also outperforms existing editing methods in specific aspects; `UDT` better represents the curly texture of the Toy Poodle than LEDITS++, and unlike Null-Text, it excels at maintaining the original pose during edits.

Classification Accuracy on Class Translation. To quantitatively evaluate the effectiveness of `UDT` in capturing breed-specific characteristics, we applied 100 learned translation directions to 100 Pug images. We then measured the shifts in classification probability for these translated images using a CLIP classifier [23]. Since `UDT` discovers 100 transformation directions without labels, we had to select a representative direction for evaluating transformations towards specific target breeds (*e.g.*, Boxer, Beagle). For each target breed, we designated its representative direction as the one that generated images yielding the highest CLIP classification score [23]. Table 1 summarizes the resulting probability changes for these transformations, with target breeds showing the highest increases highlighted in bold. The diagonal entries in Table 1 indeed confirm `UDT`'s effectiveness, showing substantial CLIP confidence boosts for intended target breeds (*e.g.*, +43.24 for 'Golden Retriever'), which aligns with the ideal expectation that targeted semantic confidence should significantly increase while other semantics experience minimal alteration.

Shifts in classification prob.	**Boxer**	**Beagle**	**Golden Retriever**	**Siberian Husky**	**Miniature Poodle**
Boxer	**15.08**	8.72	-4.34	8.04	-3.86
Beagle	2.55	**20.75**	2.26	0.50	4.08
Golden Retriever	-5.13	5.32	**43.24**	12.42	8.17
Siberian Husky	-9.15	-7.63	1.85	**54.77**	-0.71
Miniature Poodle	-3.33	-6.22	21.57	7.56	**50.38**

Table 1: **Impact of breed-to-breed translation.** Shifts in classification probability measured on 100 Pug images after transformation with 100 directions. The diagonal entries, highlighted in bold, show substantial confidence boosts for the intended target breeds.

	# Predicted Breeds	
Breed	**NoiseCLR**	**Ours**
Appenzeller	17	**44**
Chihuahua	11	**46**
Eng. Foxhound	20	**58**
Fox Terrier	30	**59**
Golden Ret.	7	**42**
Labrador Ret.	6	**38**
Yorkshire Ter.	18	**66**
Avg.	15.57	**50.43**

Table 2: **Breed diversity comparison.** The number of distinct predicted breeds after applying 100 translation directions. High numbers for `UDT` highlight its superior capability in discovering a rich and diverse set of breed-specific transformation directions.

Class Diversity on Class Translation. To compare the diversity of breed transformations produced by `UDT` and NoiseCLR [4], images were generated using 100 distinct translation directions. Each resulting image was then classified by a CLIP classifier [23] to predict its breed. The predicted breed for each image was determined by selecting the class with the highest confidence score from the classifier's output. As summarized in Table 2, `UDT` achieved an average of 50.43 distinct predicted breeds, substantially outperforming NoiseCLR's average of 15.57. These results highlight `UDT`'s superior capability in discovering a richer and more diverse set of breed-specific transformation directions.

5.3 Ablation Studies

Impact of Loss Function. We evaluate the impact of each component of our loss function in Eq. (9). Figure 5 shows that different loss configurations significantly affect transformation quality when applied to the same discovered directions. The regularization loss ($\mathcal{L}_{\text{Reg}}$) preserves breed-specific characteristics; removing it causes the generated images to lose the distinguishing characteristics of the target breed. Similarly, without contrastive loss ($\mathcal{L}_{\mathcal{T}}$), the translated images exhibit only minor changes and fail to produce convincing fine-grained class transformations. In contrast, our full model, `UDT`, accurately achieves the desired transformations, significantly enhancing visual fidelity and overall image quality.

Controlling Transformation Intensity. Our method enables users to control the intensity of breed transformations by adjusting a scale parameter. Figure 6 shows that the translated images demonstrate a smooth progression to a breed, as indicated by an arrow. This progression enables a more precise analysis of breed-specific visual characteristics. For example, interpolating along Dog "Direction #3" gradually introduces wrinkles to the skin, whereas for Cat "Direction #3", the cat appears younger and its ears become larger. This ensures that interpolation along these directions remains disentangled, naturally capturing differences between breeds.

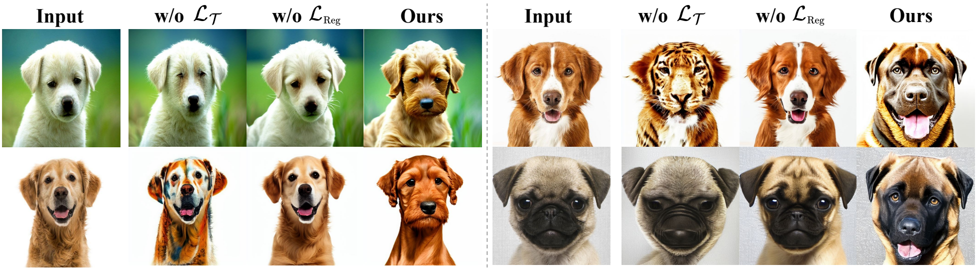
Input
w/o $\mathcal{L}_{\mathcal{T}}$
w/o $\mathcal{L}_{\text{Reg}}$
Ours
Input
w/o $\mathcal{L}_{\mathcal{T}}$
w/o $\mathcal{L}_{\text{Reg}}$
Ours

Figure 5: **Ablation study of loss function.** We compare our full model (Ours) with variants that exclude the regularization loss (w/o $\mathcal{L}_{\text{Reg}}$) and the contrastive loss (w/o $\mathcal{L}_{\mathcal{T}}$).

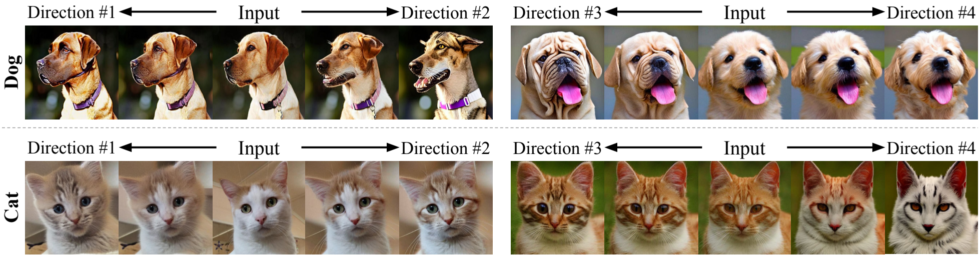

Figure 6: **Transformation results on scale parameter**. Adjusting the transformation scale enables smooth interpolation along semantic directions, offering fine-grained control over the modification intensity as shown for dog wrinkles and cat ear size.

6 Limitation

While UDT successfully performs meaningful transformations across various domains, it has a few limitations. First, UDT does not explicitly enforce the preservation of specific semantic details beyond class-level modifications, which can sometimes lead to slight changes in the background. Second, UDT's discovered transformation directions are unlabeled, requiring a post-hoc labeling process to map them to meaningful classes. Finally, the method also inherits limitations from the pre-trained diffusion model, and its performance can decline for underrepresented categories.

7 Conclusion

We introduce UDT (**U**nsupervised **D**iscovery of **T**ransformations), a novel framework for unsupervised discovery of fine-grained, class-level transformation directions within diffusion models. Additionally, our findings highlight that incorporating hierarchy information into contrastive learning improves the interpretability and control of fine-grained transformations. Unlike previous works, UDT discovers transformation directions that facilitate clear transitions between fine-grained categories—such as breeds and species—while preserving pose and object placement. Extensive qualitative and quantitative experiments demonstrate that UDT significantly outperforms existing methods, producing more distinct and visually meaningful transformations.

8 Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW (2024-0-00071), the Graduate School of Metaverse Convergence support program (IITP-2025-RS-2024-00430997), and the Convergence security core talent training business support program (IITP-2025-RS-2024-00426853) supervised by the IITP (Institute of Information & communications Technology Planning & evaluation). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2024-00456589, RS-2025-24803335). We thank Corca AI for sponsoring the compute used in this work.

References

[1] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *CVPR*, 2024.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[3] Jooyoung Choi, Dohyun Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Implicit latent diffusion models for neural radiance field editing. *arXiv preprint arXiv:2108.02938*, 2021.

[4] Yusuf Dalva and Pinar Yanardag. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. In *CVPR*, 2024.

[5] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *ICCV*, 2023.

[6] René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, Stella Graßhof, Sami S. Brandt, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. In *FG*, 2024.

[7] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

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

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[10] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE TPAMI*, 2025.

[11] Tero Karras, Samuli Laine, and Timo Aila. Flickr faces hq (ffhq) 70k from stylegan. *CoRR*, 2018.

[12] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPRW*, Colorado Springs, CO, 2011.

[13] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

[14] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *CVPR*, 2024.

[15] Nan Liu, Yilun Du, Shuang Li, Joshua B Tenenbaum, and Antonio Torralba. Unsupervised compositional concepts discovery with text-to-image generative models. In *ICCV*, 2023.

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[17] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[18] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023.

[19] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*. IEEE, 2008.

[20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[21] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *arXiv preprint arXiv:2307.12868*, 2023.

[22] 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.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

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

[26] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020.

[27] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[28] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *CVPR*, 2023.

[29] Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, et al. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *NeurIPS*, 2023.

[30] Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In *ICCV*, 2021.