XI'AN JIAOTONG UNIVERSITY

Microsoft

Diffusion Model with Cross Attention as an Inductive Bias for Disentanglement

Speaker: Tao Yang

CONTENTS

1. Background

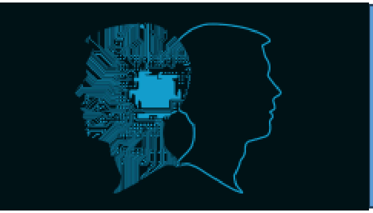
2. Method

3. Experiment

1. Background

(4) Disentangled Models & Visual Concept Learning

Objective：Empower the existing AI models the ability of human-like concept induction, and develop a unified foundation model to achieve artificial general intelligence.

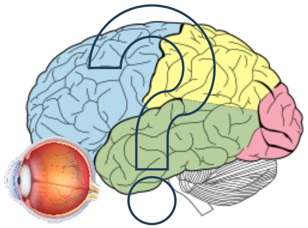


An AI must fundamentally understand the world around us, and we argue that this can only be achieved if it can learn to identify and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data ——Yshua Bengio "Representation Learning: A

Concepts can be represented by distributed patterns of activaty in networks of neuron-like units. One advantage is automatic generalization. ——Geoffre Hinton "Distributed Representation

(5) Challenges

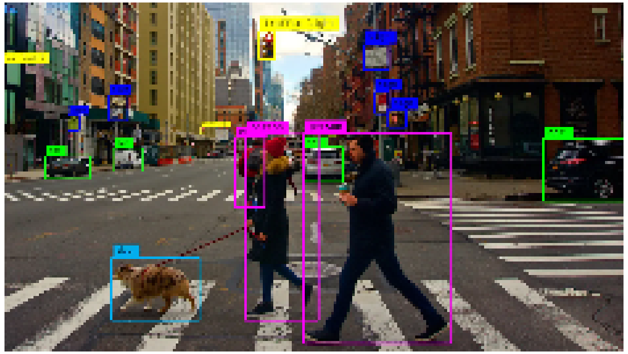
□ Unknown Mechanism of Induction



Color: black, brown, …
Shape: square, round, ….
Material: wooden, …
Azimuth: left, right, …

□ Complexity of Image data

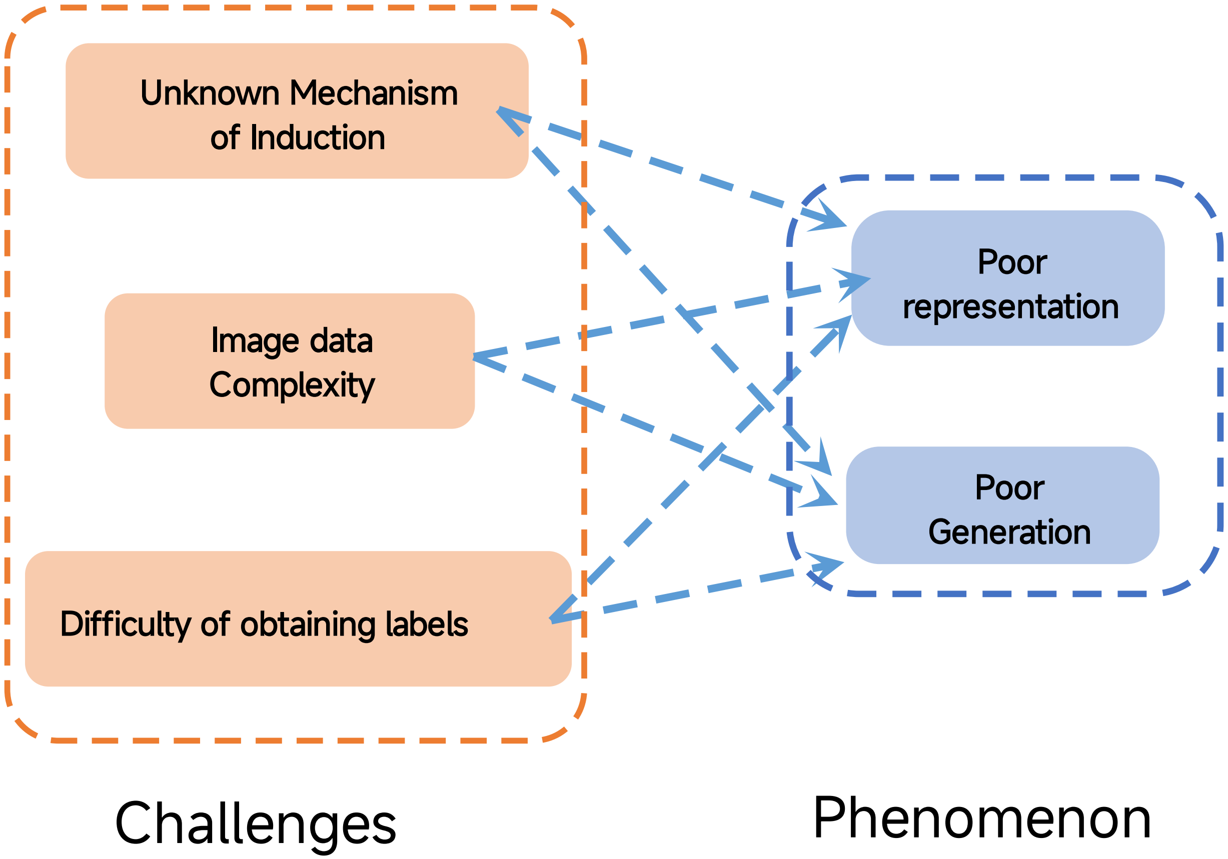
- Occlusion
- Complex Concept
- Complex Scene



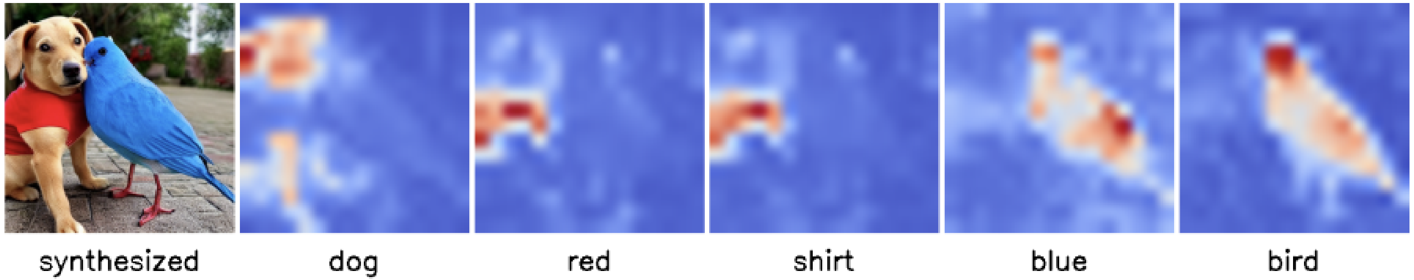
□ Difficulty of obtaining labels: there are many concepts and many object

1. Background

(5) Challenges



2. Visual Concept Learning by Disentanglement



☐ Motivation

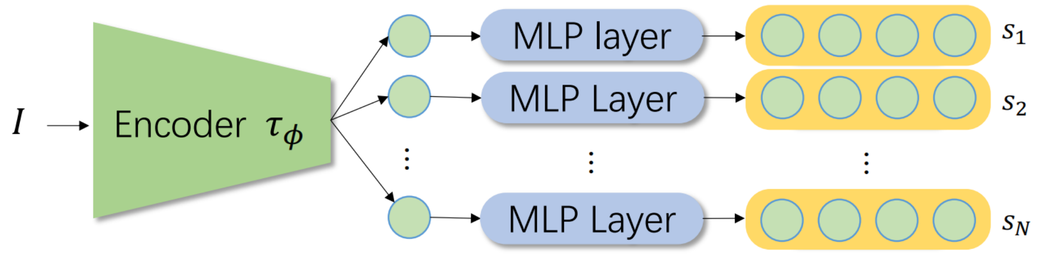
- VAEs and GANs based methods primarily rely on probability-based regularizations applied to the latent space.
- In text-to-image generation, a conditional diffusion model integrates the "disentangled" text tokens by cross attention, demonstrating the ability to generate semantically aligned images.
- Locatello et al. (2019) demonstrate that relying solely on regularizations is insufficient

2. Visual Concept Learning by Disentanglement

□ Network Design

- The image encoder τφ aims to provide a set of concept tokens $S = \{s1, \cdots, sN\}$, which act similarly to the word embeddings in the prompts for text-to-image generation in stable diffusion.

- We treat each dimension of the encoded feature vector as a disentangled factor and map each factor to a vector (i.e., concept token) by non-shared MLP layers



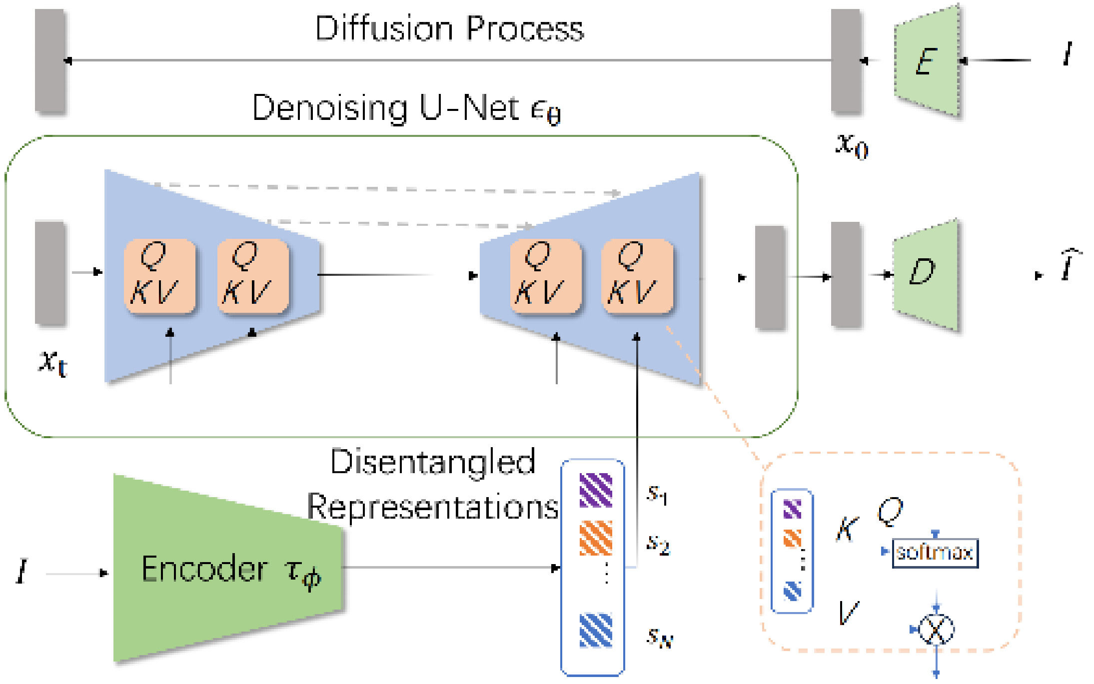
2. Visual Concept Learning by Disentanglement

Network Design

To condition the concept tokens during image generation, cross-attention is used to map these tokens into the representations of the U-Net in the diffusion model. This is accomplished using the cross attention mechanism

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$$

where each spatial feature in the intermediate feature map in diffusion serves as a query, the concept tokens act as keys and values.



(a) Framework of our EncDiff.

2. Visual Concept Learning by Disentanglement

XI'AN JIAOTONG UNIVERSITY

Microsoft

The Information Bottleneck

The diffusion model optimizes a network (e.g., U-Net) ϵ_θ to predict the noise from the noisy input x_t and the condition S (concept tokens), with the loss function defined as

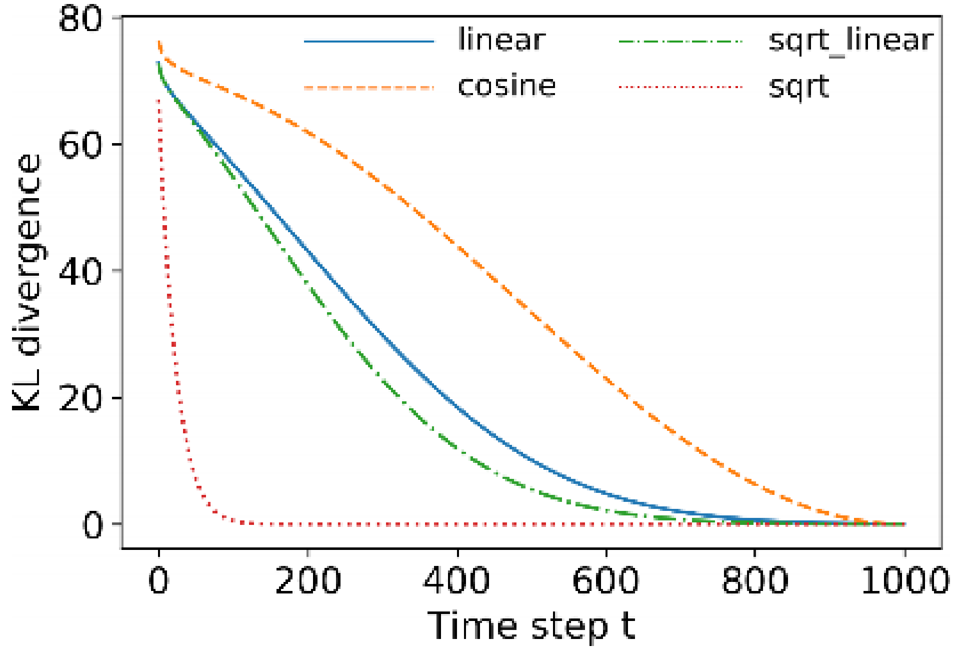
$$\mathcal{L}_r = \mathbb{E}_{x_0,\varepsilon,t}||\varepsilon_\theta(x_t, t, \mathcal{S}) - \varepsilon||.$$

We can interpret the loss function as a reconstruction of latent input x_0. The loss function $\mathcal{L}_r$ can be rewritten as:

$$\mathcal{L}_r = \Sigma_{t\geq 1}|C_t - D(p_\theta(x_{t-1}|x_t, \mathcal{S})||q(x_{t-1}))|,$$

The loss function of AnnealVAE

$$\mathcal{L}(\varphi, \phi) = -E_{q_\varphi(z|x)}[\log p_\phi(x|z)] + \gamma|C - D_{KL}(q_\varphi(z|x)||p(z))|,$$



(b) KL divergence curves.

2. Visual Concept Learning by Disentanglement

Quantitative Analysis

Method	Cars3D		Shapes3D		MPI3D	
	FactorVAE score↑	DCI↑	FactorVAE score↑	DCI↑	FactorVAE score↑	DCI↑
VAE-based:						
FactorVAE [17]	0.906 ± 0.052	0.161 ± 0.019	0.840 ± 0.066	0.611 ± 0.082	0.152 ± 0.025	0.240 ± 0.051
β-TCVAE [4]	0.855 ± 0.082	0.140 ± 0.019	0.873 ± 0.074	0.613 ± 0.114	0.179 ± 0.017	0.237 ± 0.056
GAN-based:						
InfoGAN-CR [21]	0.411 ± 0.013	0.020 ± 0.011	0.587 ± 0.058	0.478 ± 0.055	0.439 ± 0.061	0.241 ± 0.075
Pre-trained GAN-based:						
LD [29]	0.852 ± 0.039	0.216 ± 0.072	0.805 ± 0.064	0.380 ± 0.062	0.391 ± 0.039	0.196 ± 0.038
GS [11]	0.932 ± 0.018	0.209 ± 0.031	0.788 ± 0.091	0.284 ± 0.034	0.465 ± 0.036	0.229 ± 0.042
DisCo [26]	0.855 ± 0.074	0.271 ± 0.037	0.877 ± 0.031	0.708 ± 0.048	0.371 ± 0.030	0.292 ± 0.024
Diffusion-based:						
DisDiff [37]	$\mathbf{0.976} \pm 0.018$	0.232 ± 0.019	0.902 ± 0.043	0.723 ± 0.013	0.617 ± 0.070	0.337 ± 0.057
EncDiff (Ours)	0.773 ± 0.060	$\mathbf{0.279} \pm 0.022$	$\mathbf{0.999} \pm 0.000$	$\mathbf{0.969} \pm 0.030$	$\mathbf{0.872} \pm 0.049$	$\mathbf{0.685} \pm 0.044$

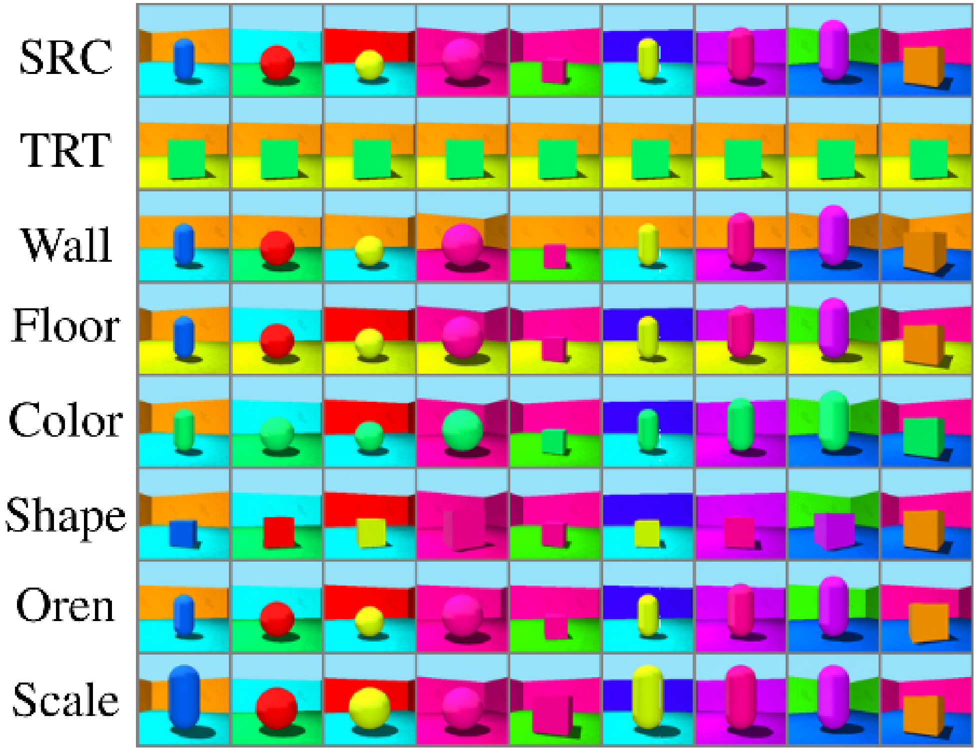
2. Visual Concept Learning by Disentanglement

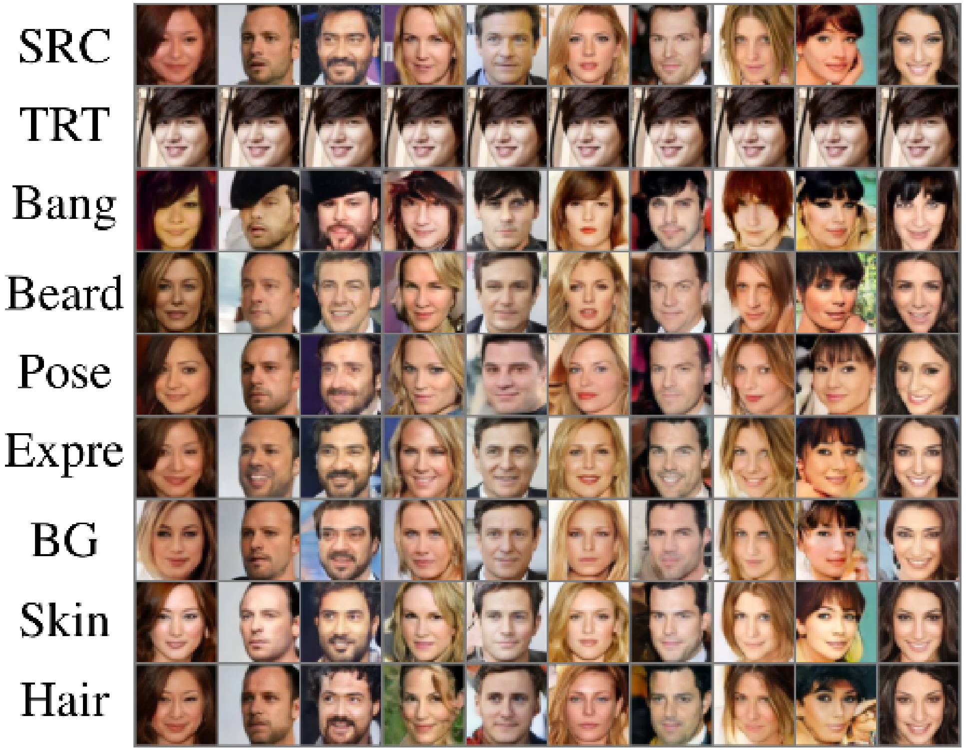
Quantitative Results on CelebA

Model	TAD ↑	FID ↓
β-VAE [14]	0.088 ± 0.043	99.8 ± 2.4
InfoVAE [40]	0.000 ± 0.000	77.8 ± 1.6
Diff-AE [24]	0.155 ± 0.010	22.7 ± 2.1
InfoDiffusion [31]	0.299 ± 0.006	23.6 ± 1.3
DisDiff [37]	0.305 ± 0.010	18.2 ± 2.1
EncDiff	$\mathbf{0.638 \pm 0.008}$	$\mathbf{14.8 \pm 2.3}$

2. Visual Concept Learning by Disentanglement

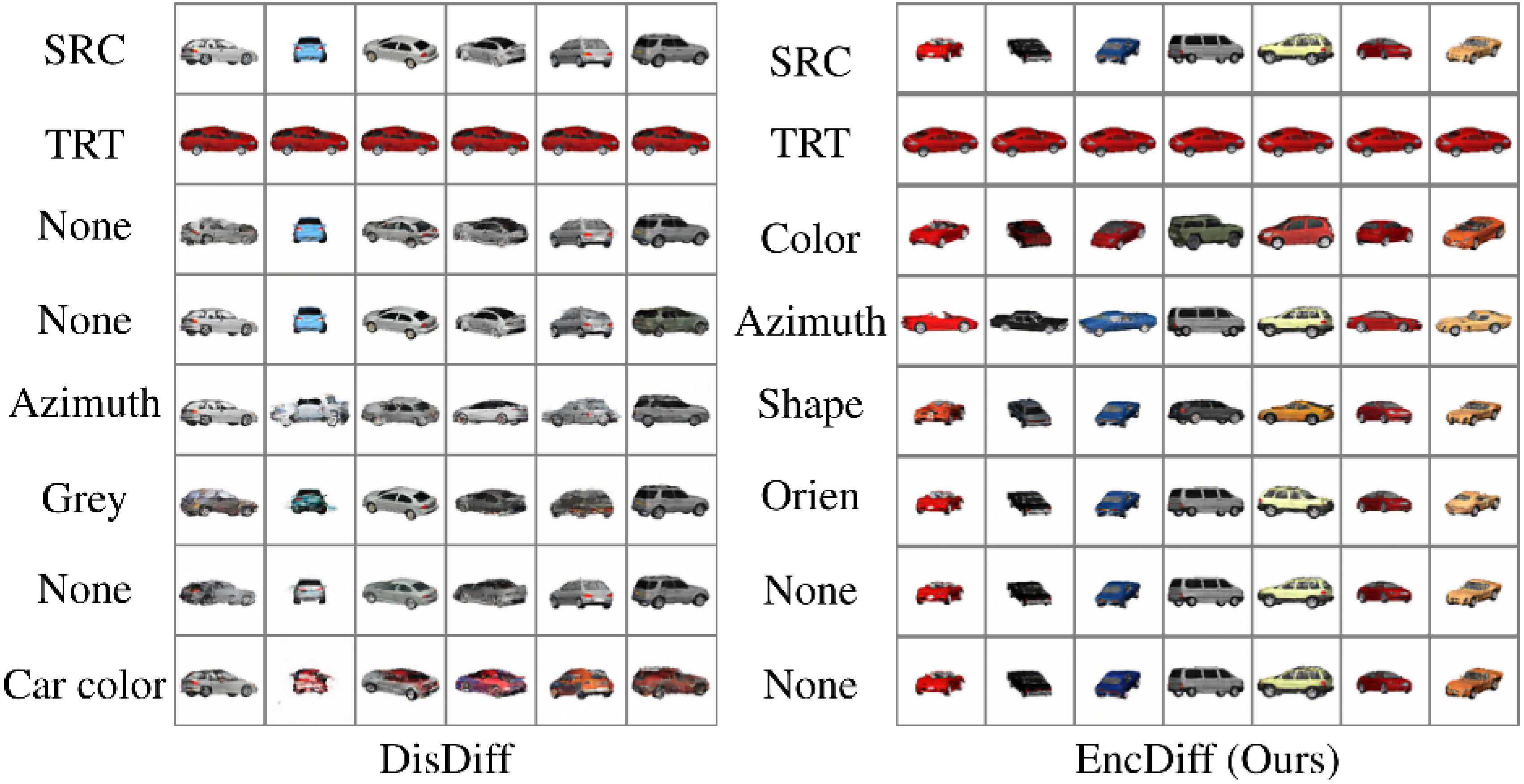
☐ Qualitative Results





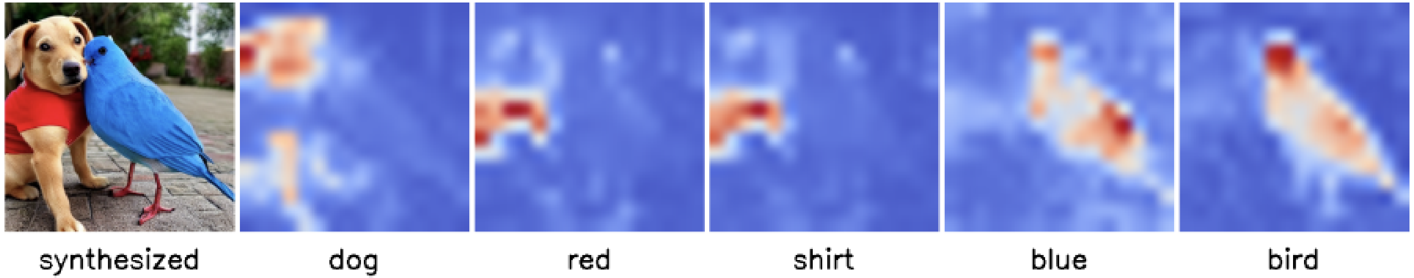
2. Visual Concept Learning by Disentanglement

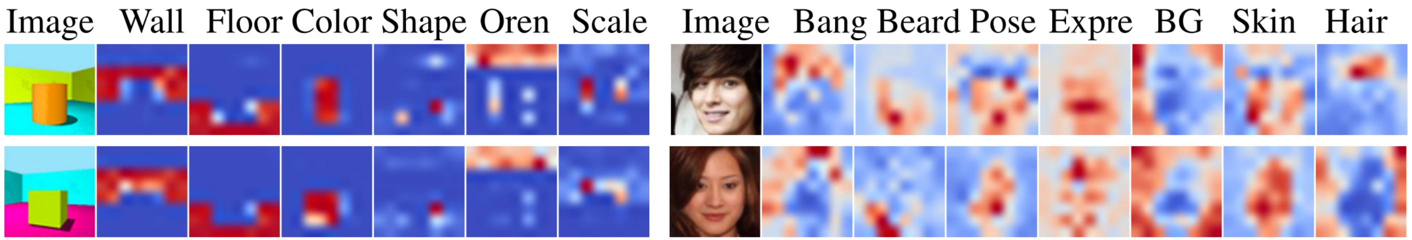
Qualitative Results



2. Visual Concept Learning by Disentanglement

☐ Visualization of Attention Map





2. Visual Concept Learning by Disentanglement

Ablation Study

Method	FactorVAE score↑	DCI↑
EncDiff w/sqrt	0.997 ± 0.011	0.950 ± 0.041
EncDiff w/sqrt linear	0.988 ± 0.026	0.924 ± 0.050
EncDiff w/linear	0.999 ± 0.002	0.930 ± 0.045
EncDiff w/cosine	$\mathbf{0.999} \pm 0.001$	$\mathbf{0.969} \pm 0.030$

Method	FactorVAE score↑	DCI↑
EncDec w/o Diff	0.537 ± 0.074	0.178 ± 0.050
EncDiff w/ AdaGN	0.911 ± 0.101	0.637 ± 0.068
EncDiff	$\mathbf{0.999} \pm 0.000$	$\mathbf{0.969} \pm 0.030$

Method	Params.↓ (M)	FLOPs↓ (M)	Time↓ (s)
Diff-AE [24]	67.8	3955.1	31.0
DisDiff [37]	57.1	5815.8	35.3
EncDiff	42.3	2898.5	11.8

Method	FactorVAE score ↑	DCI↑
EncDiff-V	0.999 ± 0.000	0.900 ± 0.045
EncDiff	$\mathbf{0.999} \pm 0.001$	$\mathbf{0.969} \pm 0.030$

Thank you!