Universal Multi-Domain Translation via Diffusion Routers

Duc Kieu, Kien Do, Tuan Hoang, Thao Minh Le,
Tung Kieu, Dang Nguyen, Thin Nguyen

Email: v.kieu@deakin.edu.au

Motivation

Any-to-any generative models have gained increasing attention

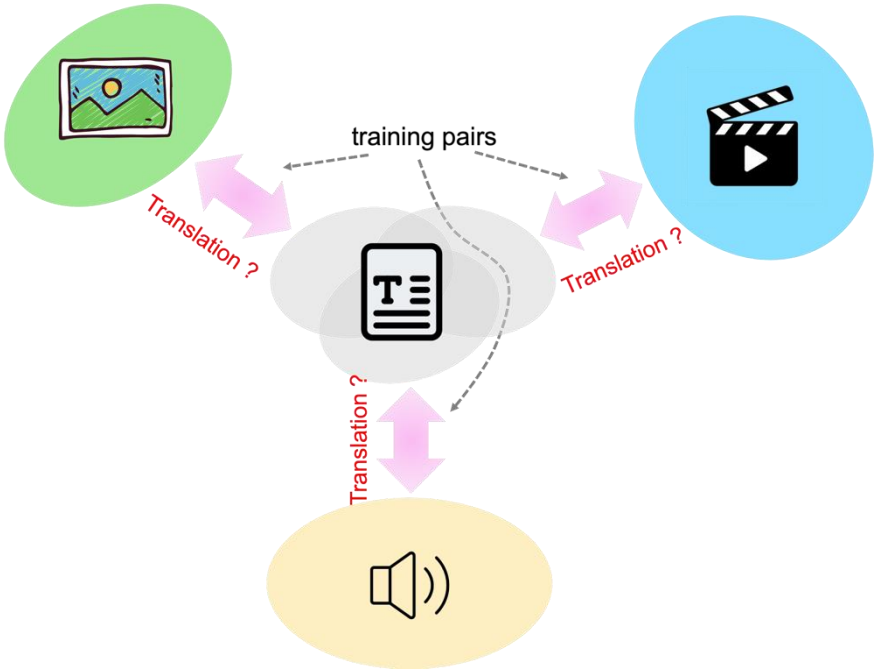
Most existing frameworks rely on one of two strategies:

(i) Fully aligned cross-domain tuples (image, text, audio describing same content)

- ➢ Impractical as the number of domains grows

(ii) Multiple paired datasets

- ➢ Some domain pairs are scarce (audio-image)



Observation:

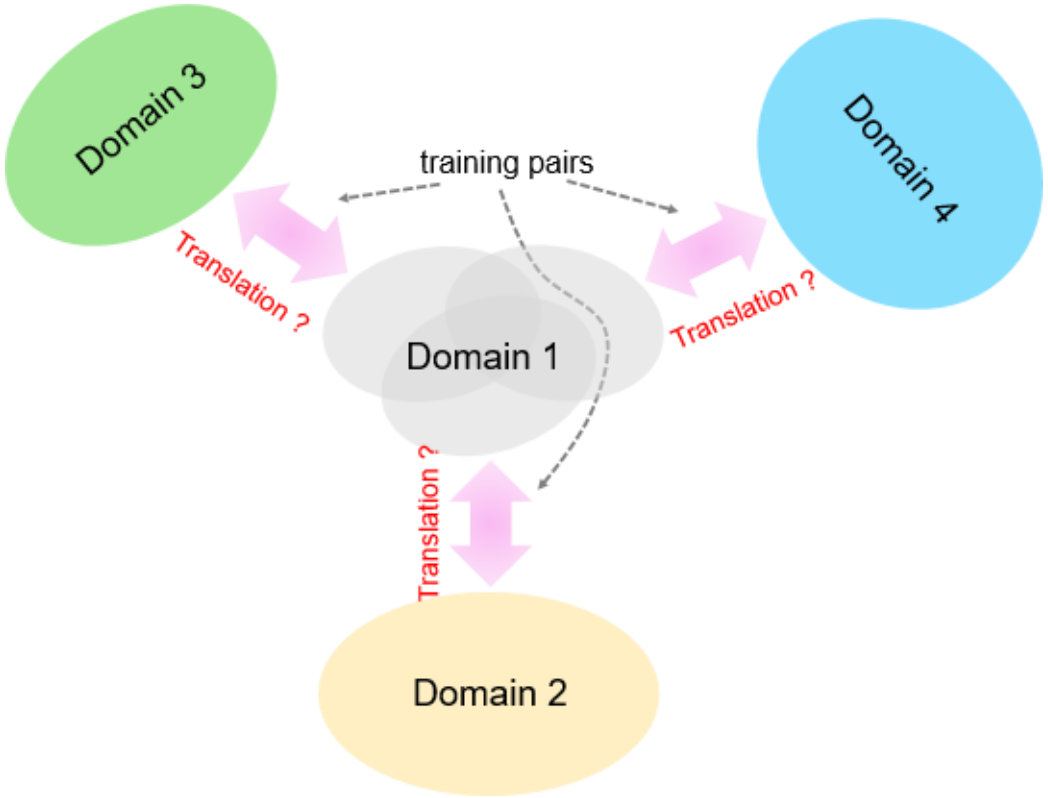
Paired datasets often share a central domain,
e.g., text in image–text and audio–text pairs.

Research question:

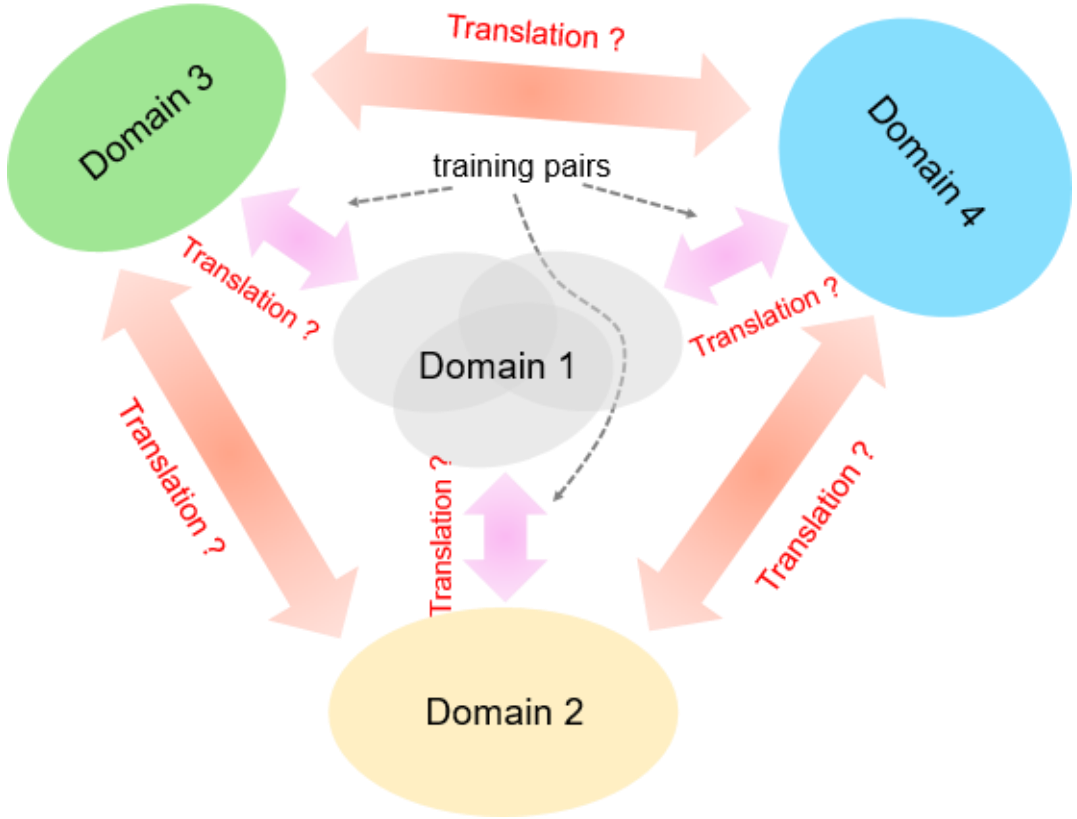
Can we build an any-to-any framework using such paired datasets?

Universal Multi-domain Translation

Consider K distinct domains, $X^1, X^2, ..., X^K$ with training data consisting $K-1$ paired dataset between each domain X^k and a shared central domain X^c



Conventional multi-domain translation



Universal multi-domain translation

Diffusion Router: Indirect translation

Consider 2 domains without paired dataset X^i, X^j, modeling conditional distribution $p\left(x^j|x^i\right)$ and $p\left(x^i|x^j\right)$ indirectly with X^c

$$p\left(x^j|x^i\right) = \int p\left(x^j|x^c\right) p\left(x^c|x^i\right) dx^c, \qquad p\left(x^i|x^j\right) = \int p\left(x^i|x^c\right) p\left(x^c|x^j\right) dx^c$$

- We can model $p\left(x^k|x^c\right)$ and $p\left(x^c|x^k\right)$ using conditional diffusion model or bridge models

We model all bidirectional mappings using a single network $\epsilon_\theta\left(x_t^{\text{tgt}}, t, x^{\text{src}}, \text{tgt}, \text{src}\right)$

The training objective for paired domains is given:

$$\mathcal{L}_{\text{paired}}(\theta) = \mathbb{E}_{(x^k,x^c)\sim\mathcal{D}_{k,c},t,\epsilon,\zeta}\left[\zeta\left\|\epsilon_\theta\left(x_t^k, t, x^c, k, c\right) - \epsilon\right\|_2^2 + (1-\zeta)\left\|\epsilon_\theta\left(x_t^c, t, x^k, c, k\right) - \epsilon\right\|_2^2\right]$$

Diffusion Router: Direct translation

To enable direct translation between X^i, X^j, we model $p_\theta\left(x^j|x^i\right)$ and $p_\theta\left(x^i|x^j\right)$ by minimize the KL divergence with indirect translation distribution:

$$\mathbb{E}_{p(x^i)}\left[D_{KL}\left[p\left(x^j|x^i\right)\|p_\theta\left(x^j|x^i\right)\right]\right]$$

$$\approx \mathbb{E}_{(x^i,x^c)\sim\mathcal{D}_{i,c}}\left[D_{KL}\left(p\left(x^j|x^c\right)\|p_\theta\left(x^j|x^i\right)\right)\right]$$

Intuition: with same-pair (highly correlated) conditions, target-domain samples should be similar.

The training objective for unpaired domains is given:

$$\mathcal{L}_{\text{unpaired}}(\theta) = \mathbb{E}_{(x^i,x^c)\sim\mathcal{D}_{i,c},\boxed{x_t^j\sim p_{\text{ref}}\left(x_t^j|x^c\right)},t,\epsilon}\left[\left\|\epsilon_\theta\left(x_t^j,t,x^i,j,i\right)-\epsilon_{\text{ref}}\left(x_t^j,t,x^c,j,c\right)\right\|_2^2\right],$$

Note: Require backward sampling

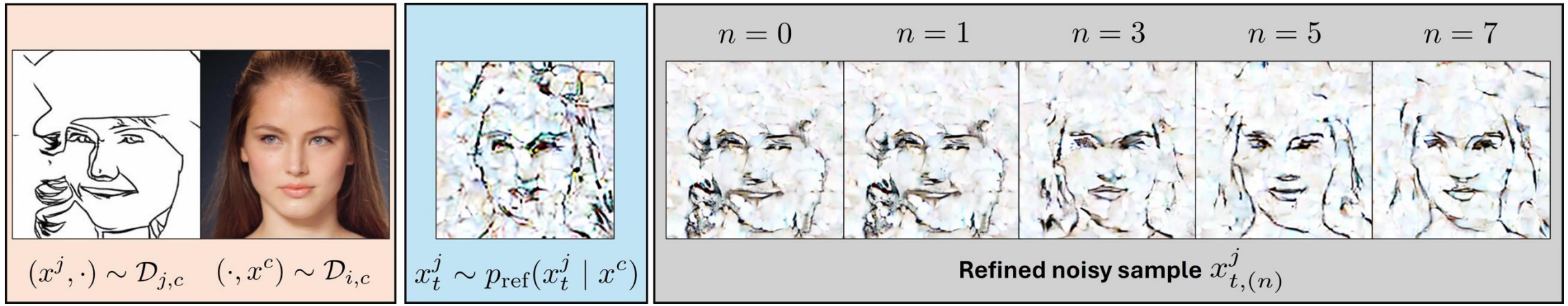
$$\mathcal{L}_{\text{final}}(\theta) = \lambda_1\mathcal{L}_{\text{unpaired}}(\theta)+\lambda_2\mathcal{L}_{\text{paired}}(\theta)$$

Tweedie Refinement

Sampling from $p_{\text{ref}}\left(x_t^j | x^c\right)$ requires backward sampling, costly

- We introduce Tweedie refinement to sample with few steps, iteratively denoise using conditional score and add noise following forward process

$$x_{t,(n+1)}^j = x_{t,(n)}^j + \sigma_t \left(\epsilon - \epsilon_\theta \left(x_{t,(n)}^j, t, x^c, j, c\right)\right)$$



Diffusion Router: Results

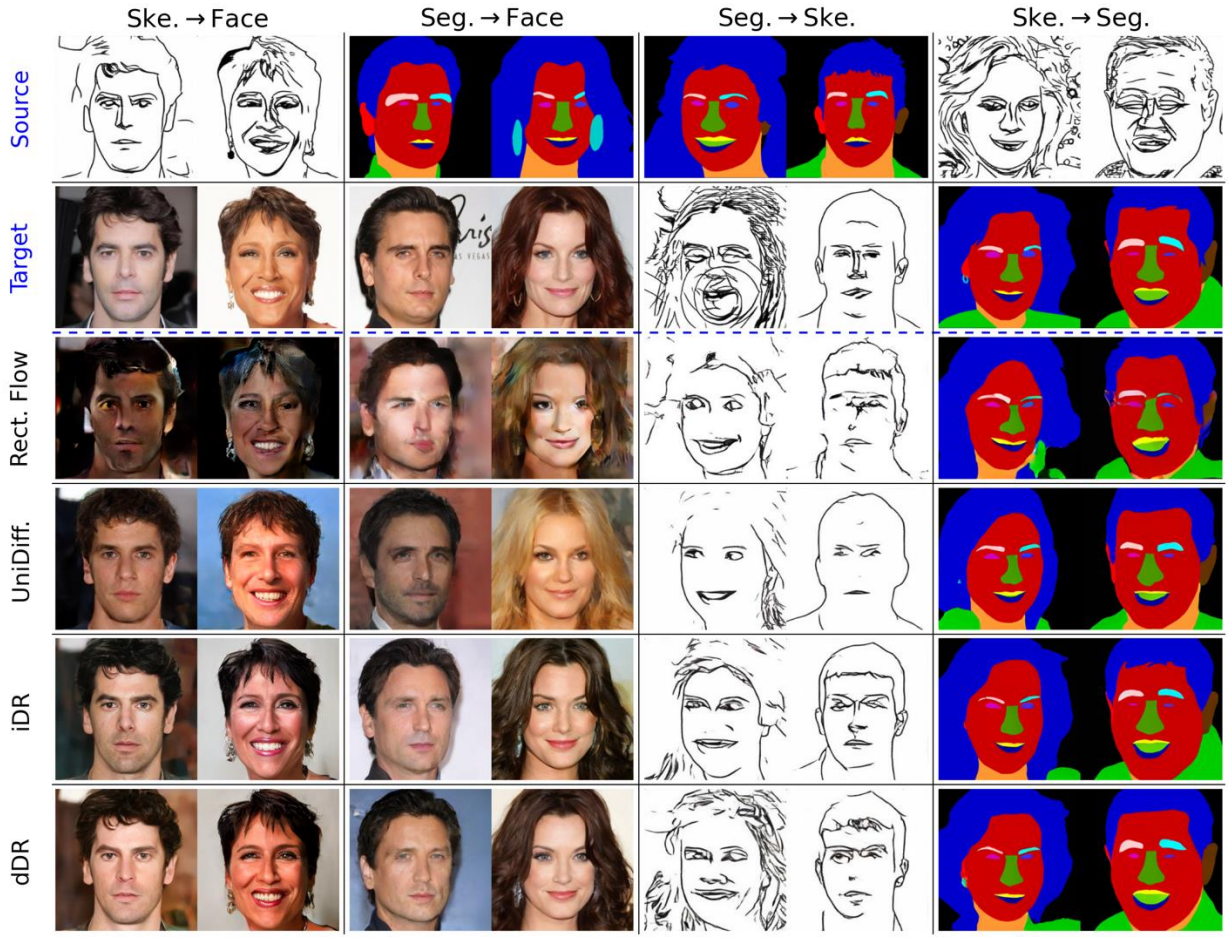
Method	FID↓					
	Shoes-UMDT			Faces-UMDT-Latent		
	Edge↔Shoe	Gray.↔Shoe	Edge↔Gray.	Ske.↔Face	Seg.↔Face	Ske.↔Seg.
StarGAN	9.92/20.18	19.73/42.61	18.64/27.41	-	-	-
Rectified Flow	2.88/30.92	3.75/43.38	20.14/18.83	20.22/97.76	10.85/81.44	50.82/17.31
UniDiffuser	2.98/11.94	2.72/4.40	4.81/12.26	13.13/55.46	11.02/46.04	36.13/12.52
iDR	**1.66/5.15**	**0.53/1.60**	**1.85/5.48**	**9.07/23.88**	6.12/**19.12**	**15.37**/6.15
dDR	2.01/5.76	0.57/1.69	2.74/6.51	9.62/27.09	**3.43**/21.26	19.42/**5.52**

Table 1: FID scores on Shoes-UMDT and Faces-UMDT-Latent. Translations without paired data are marked in brown. The best results are shown in **bold**, and the second-best are underlined.

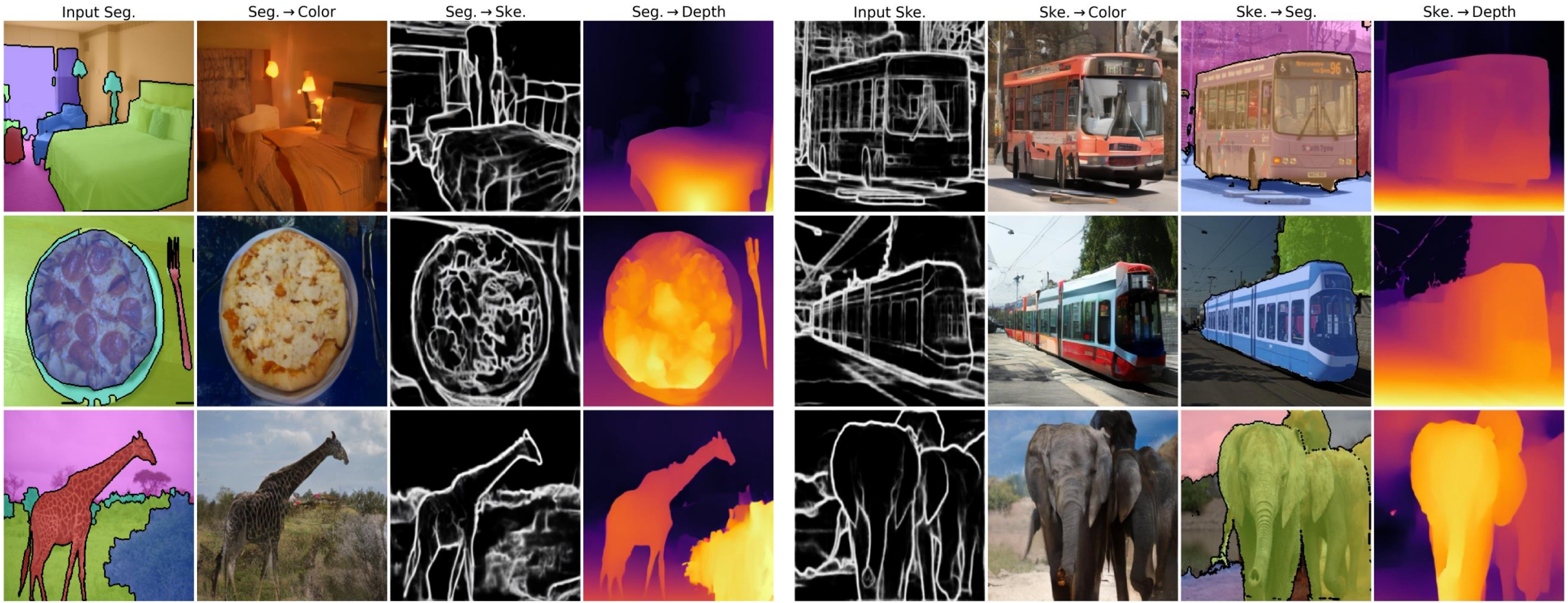
Method	FID↓					
	Ske.↔Color	Seg.↔Color	Depth↔Color	Ske.↔Seg.	Ske.↔Depth	Seg.↔Depth
Rectified Flow	23.18/80.80	54.00/142.15	17.32/112.64	64.47/75.58	78.41/28.69	79.20/35.53
UniDiffuser	15.39/40.93	35.81/89.58	12.64/59.72	39.62/38.44	28.12/15.72	38.39/23.41
iDR	10.72/21.73	21.64/29.28	7.25/24.19	**22.77/22.96**	**17.88/8.63**	**23.19/12.00**
dDR	**10.12/20.94**	**21.23/28.32**	**7.00/23.20**	26.73/23.64	20.75/9.42	24.91/14.87

Table 2: FID scores on COCO-UMDT-Star. Translations without paired data are marked in brown. The best results are shown in **bold**, and the second-best are underlined.

Diffusion Router: Results (cont.)



Diffusion Router: Results (cont.)



Thank you!