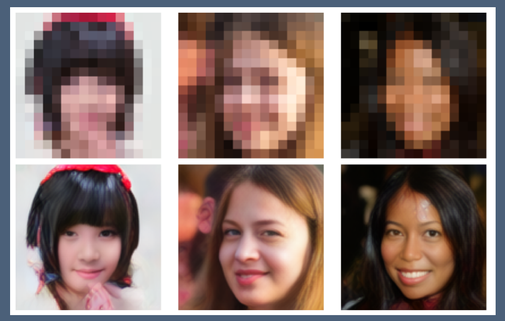
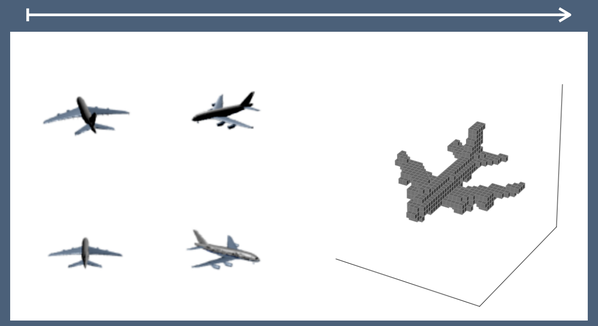




Latent Denoising Deep Diffusion Models (LDDBMs)

Towards General Modality Translation with Contrastive and Predictive Latent Diffusion Bridge

Nimrod Berman*[1, 2], Omkar Joglekar*[1, 3], Eitan Kosman[1], Dotan Di Castro[1], Omri Azencot

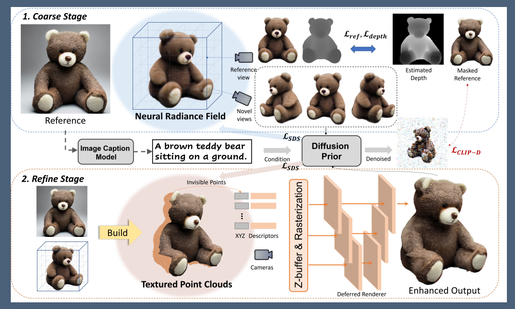
*Equal Contribution, [1]Bosch Center for Artificial Intelligence, [2]Ben-Gurion University, [3]Technical University of Munich

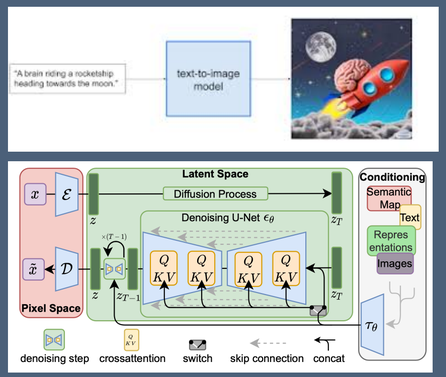
NEURAL INFORMATION PROCESSING SYSTEMS

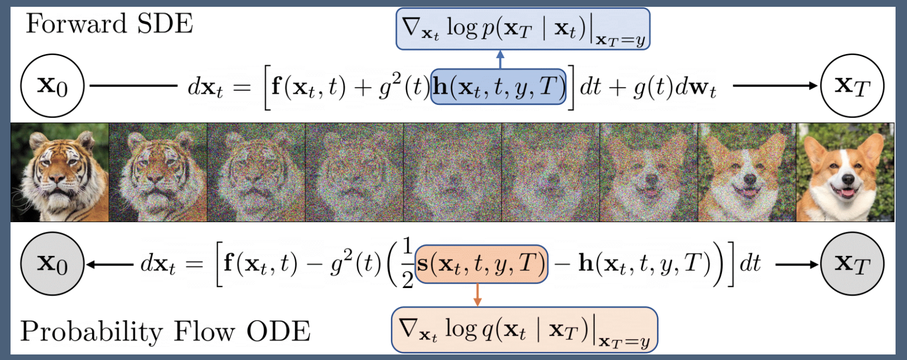
Motivation

Can we design a GENERAL Modality Translation (MT) framework?

- Diffusion Models excel in both single-modality and modality-translation tasks.
- Denoising Diffusion Bridge Models theoretically support MT and showcase state-of-the-art results.
- However, these models **require a shared space** between modalities.
- U-Nets present **strong inductive biases** towards image data, maintaining modality dependence.



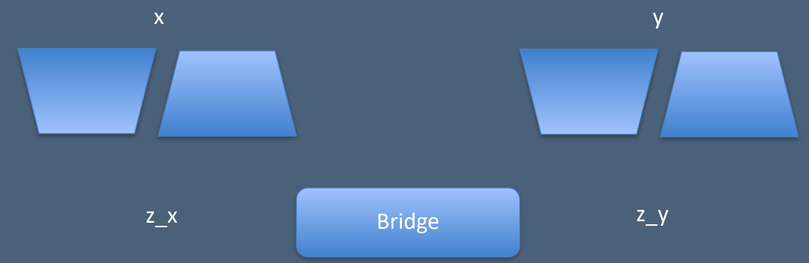




The Simple solution

And it's caveats

- Generate noticeable artefacts and lack high-frequency details in image super-resolution (SR) (Fig. 1, left, "Basic")
- Lack semantic alignment of similar samples across modalities (Fig. 1, center)



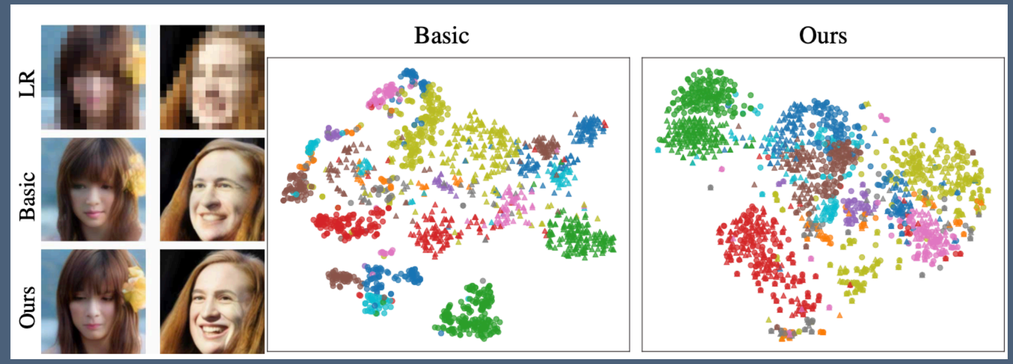
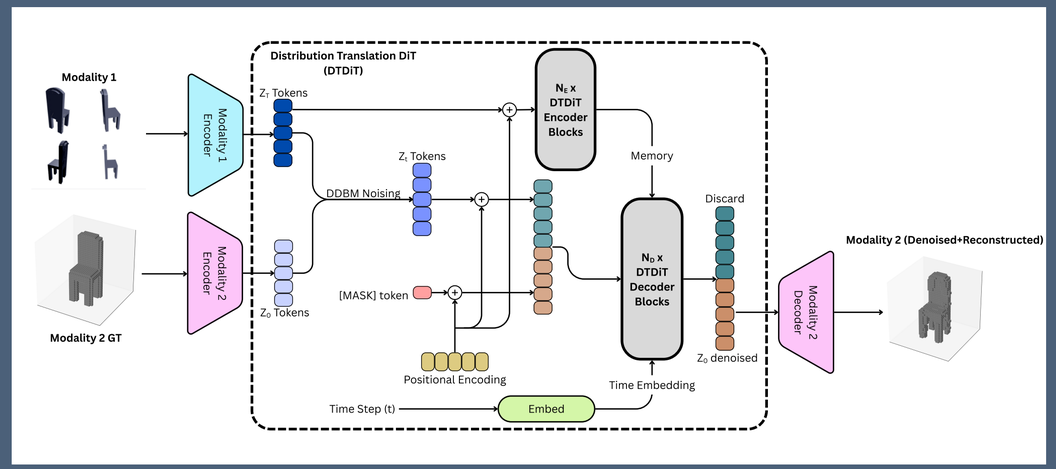


Figure 1: **(Left)** Comparison of our method ("Ours") to the simple solution ("Basic") on a SR task. **(Centre, Right)** t-SNE plots of the "Basic" and "Ours" methods. Circles and triangles denote multi-view images and 3D shapes respectively, coloured by semantic category.

Our Approach

Latent Denoising Diffusion Bridge



- **Core Idea:** Introduce a latent-variable extension of Denoising Diffusion Bridge Models (DDBMs) to connect arbitrarily-dimensioned modalities.
- **Contrastive Alignment Loss:** Semantic alignment of latent codes
- **Predictive Loss:** Regularise the full translation pipeline to accurately reconstruct the target x, improving overall fidelity.
- **Domain Agnostic Architecture:** Introduce a novel Transformer Encoder-Decoder based architecture that has been well-known for its language translation superiority
- **Iterative Training Procedure:** We suggest and empirically prove that training the modality-specific models and the bridge alternately significantly improves performance. This method is inspired by the alternate Generator-Discriminator training in GANs.

$$\mathcal{L}_{\text{bridge}} = \mathbb{E}_{z_t, z_0, z_T, t}\left[w(t)\,|s_\theta(z_t, z_T, t) - \nabla_{z_t}\log q(z_t|z_0, z_T)|^2\right]$$

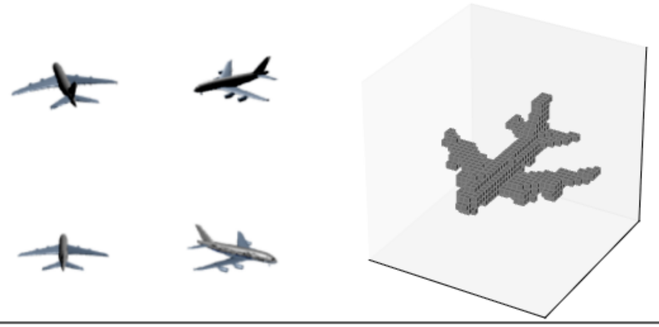
$$\mathcal{L}_{\text{infoNCE}} = \log\frac{\phi(z_0, z_T)}{\phi(z_0, z_T) + \sum_{j=1}^{M}\phi(z_0, z_T^j)}$$

$$\mathcal{L}_{\text{pred}} = d\left(D_x \circ B \circ E_y(y), x\right)$$

$$\mathcal{L} = \mathcal{L}_{\text{bridge}} + \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{infoNCE}}$$

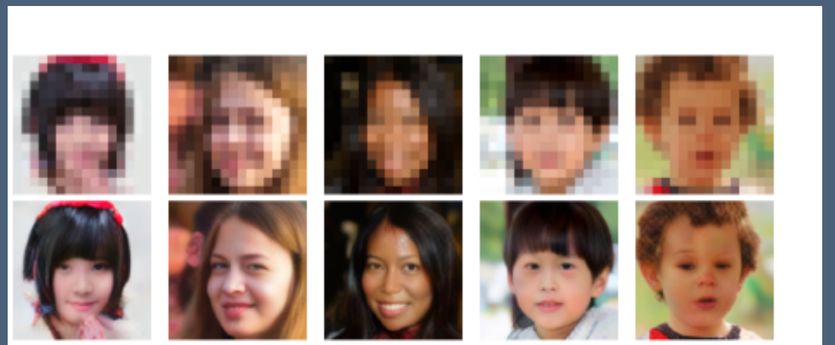
Results

Multi-view to 3D-Shape



	EDM	3D-EDM	DiT	SiT	Ours
1-NNA ↓	.532 ± .013	.575 ± .009	.548 ± .004	.563 ± .007	**.508 ± .005**
IoU ↑	.631 ± .006	.602 ± .003	.613 ± .011	.604 ± .003	**.664 ± .002**

Super Resolution

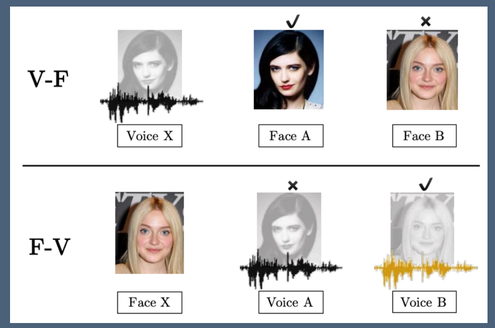


	EDM	DiWa	DiT	SiT	Ours
PSNR ↑	23.1 ± 0.7	23.3 ± –	22.2 ± 1.1	21.5 ± 0.4	**25.6 ± 0.4**
SSIM ↑	0.58 ± .05	0.65 ± –	0.52 ± .07	0.57 ± .02	**0.68 ± .03**
LPIPS ↓	0.41 ± .02	0.39 ± –	0.49 ± .01	0.51 ± .03	**0.32 ± .01**

Edges2Handbag

Method / Metric	FID ↓	Time (s) ↓
LDDBM	4.17	7.8
DDBM	2.93	16.9

Face2Voice, Voice2Face



Method	Face → Voice ↑	Voice → Face ↑
LDDBM	**71.2**	**75.1**
SiT	65.7	68.3
[40]	79.5	81.0

Ablation Studies

Design Choice Ablations

Component	ShapeNet IoU ↑	ShapeNet 1-NNA ↓	nuScene IoU ↑	nuScene 1-NNA ↓	CelebA-HQ PSNR ↑	CelebA-HQ SSIM ↑	CelebA-HQ LPIPS ↓
(1) U-Net	.635	.518	.216	.818	23.2	0.57	0.42
(2) DiT	.613	.548	.208	.825	22.2	0.52	0.49
(3) + Encoder-Decoder	.651	.518	.217	.821	23.4	0.53	0.38
(4) + Spatial Embedding	.658	.522	.224	.812	22.9	0.56	0.41
(5) + $[MASK]$ (Ours)	**.664**	**.508**	**.233**	**.807**	**25.6**	**0.68**	**0.32**

Losses Ablations

		$\mathcal{L}_{\text{REC}}$	$\mathcal{L}_{\text{PRED}}$	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{INFONCE}}$	$\mathcal{L}_{\text{PRED}} + \mathcal{L}_{\text{INFONCE}}$
SHAPENET	1-NNA ↓	.625 ± .003	.522 ± .002	.578 ± .004	**.508 ± .005**
	IoU ↑	.609 ± .007	.643 ± .005	.627 ± .007	**.664 ± .002**
CELEBA-HQ	PSNR ↑	20.5 ± 0.4	23.7 ± 0.3	21.4 ± 0.1	**25.6 ± 0.4**
	SSIM ↑	0.49 ± .03	0.64 ± .02	0.51 ± .05	**0.68 ± .03**
	LPIPS ↓	0.62 ± .04	0.41 ± .02	0.63 ± .03	**0.32 ± .01**

Training Ablations

		TWO-STEP	END-TO-END	ITERATIVE
SHAPENET	1-NNA ↓	.522 ±.006	.517 ±.003	**.508 ± .005**
	IoU ↑	.637 ±.003	.642 ±.005	**.664 ± .002**
CELEBA-HQ	PSNR ↑	23.3 ±0.6	23.4 ±0.3	**25.6 ± 0.4**
	SSIM ↑	0.58 ±.07	0.57 ±.05	**0.68 ± .03**
	LPIPS ↓	0.40 ±.02	0.39 ±.01	**0.32 ± .01**

Conclusions

- We introduce a method that can achieve better **general** modality translation
- Our method supports **arbitrary modality pairs**
- **Easy-to-use** method
- **Strong empirical results**

Thank you!

See you at our poster!

Our Project Page: