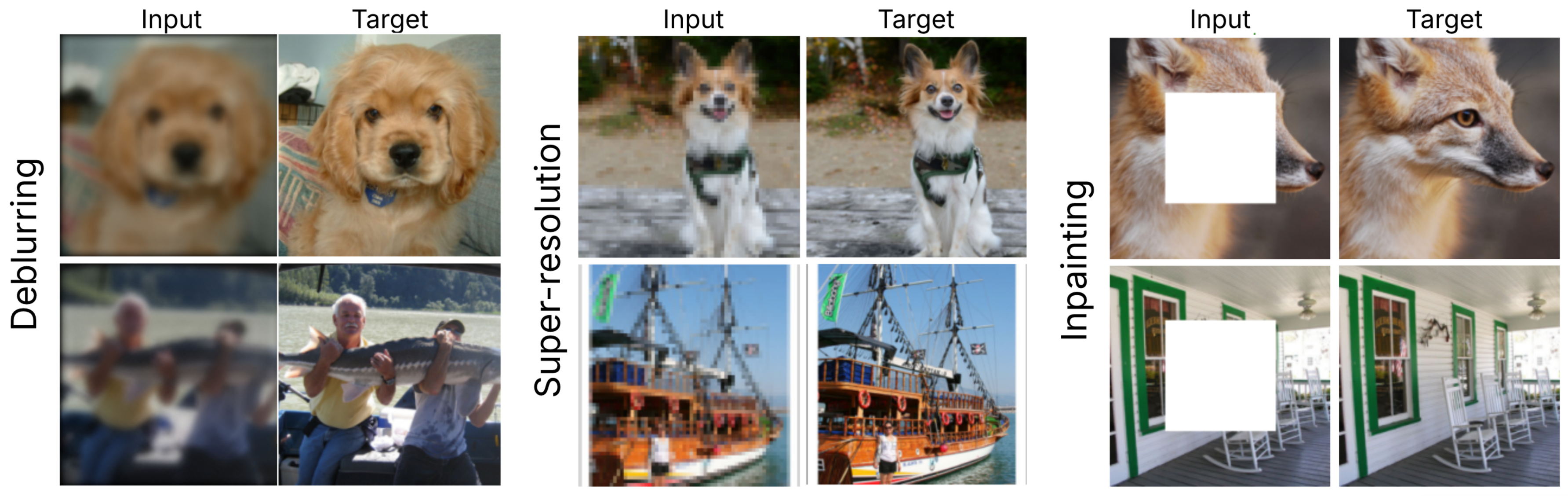
Inverse Bridge Matching Distillation

Nikita Gushchin[1,2] David Li[*,1,3] Daniil Selikhanovych[*,1,4,5]
Evgeny Burnaev[1,2] Dmitry Baranchuk[4], Alexander Korotin[1,2]

*Equal contribution [1]Skoltech, [2]AIRI, [3]MIPT, [4]Yandex Research, [5]HSE University

I Paired Image-to-Image Translation

The paired image-to-image translation problem:
Given a dataset of paired images (corrupted→clean or domain A→domain B), the goal is to train a model that maps each input to its aligned target.



II Diffusion Bridge Models

Consider the prior process $dx_t = f(x_t,t)dt + g(t)dw_t$ and its posterior $q(x_t|x_0,x_T)$, called the bridge.

The diffusion bridge model is provided by the reverse-time SDE from corrupted to clean images:

$$dx_t = \{f(x_t,t) - g^2(t)v^*(x_t,t)\}dt + g(t)d\bar{w}_t, \quad \boxed{x_T \sim p(x_T)} \text{ Corrupted data}$$

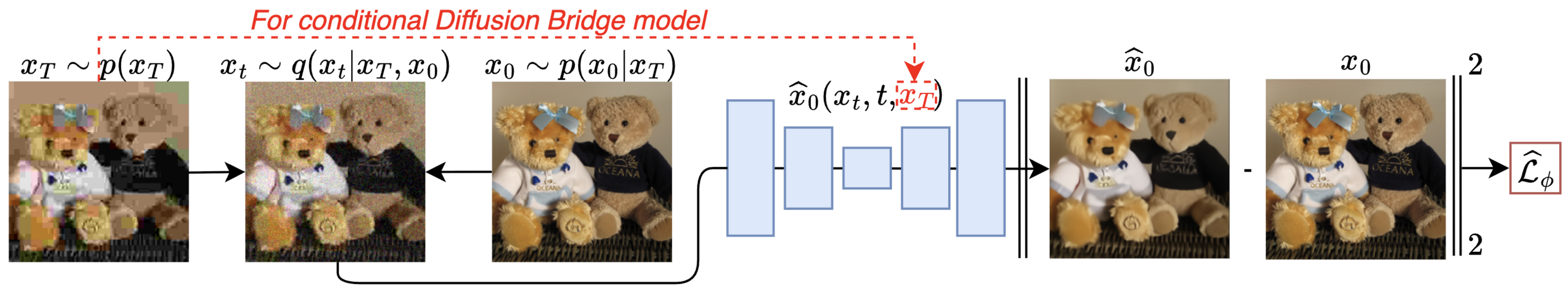
The drift $v^*(x_t,t)$ is learned by solving the Bridge Matching problem (in x_0 parametrization):

Paired dataset

$$\min_{\phi} \boxed{\mathbb{E}_{x_0,t,x_t,x_T}\left[\lambda(t)\|\widehat{x}_0^{\phi}(x_t,t,x_T) - x_0\|^2\right]}, \boxed{\widehat{\mathcal{L}}_{\phi}}$$
$$\boxed{(x_0,x_T) \sim p(x_0,x_T)}, t \sim U([0,T]),\ x_t \sim q(x_t|x_0,x_T),$$

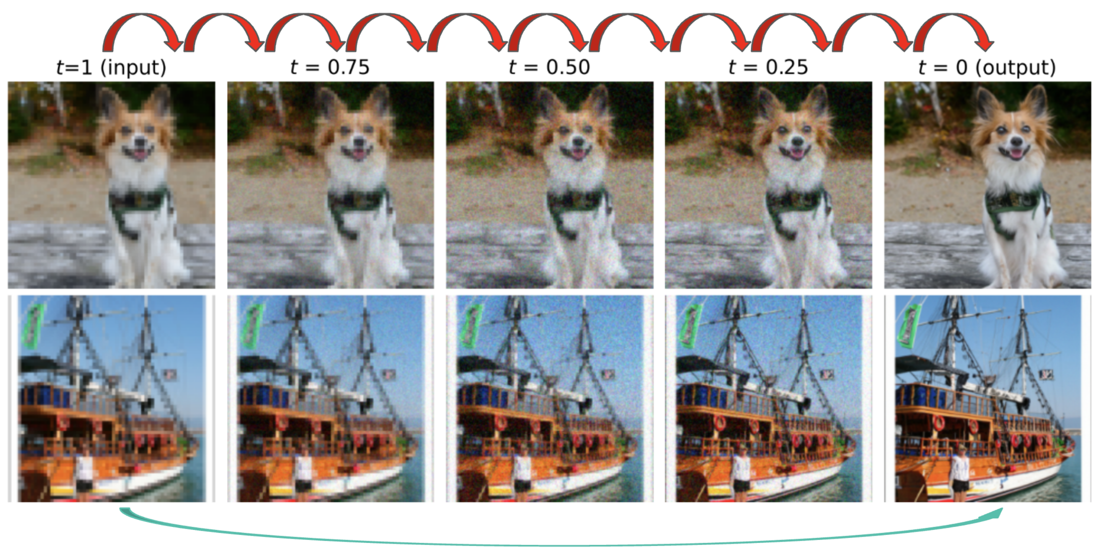
Models with additional conditioning on corrupted images are called the conditional Diffusion Bridge Models.

Training pipeline for Diffusion Bridge Models



III Motivation

The slow inference problem: Diffusion bridge models produce high-quality translations, but they require **10 to 1000** of steps to simulate the reverse SDE.



The goal of distillation is to train a new 1-step or few-step generator that mimics the full diffusion bridge model.

Key idea

The core idea is to learn a one-step generator G_θ of clean images from corrupted images such that the Diffusion Bridge Model for generated data coupling $p_\theta(x_0,x_T)$ matches the teacher Diffusion Bridge Model learned on the ground-truth data pairs $p(x_0,x_T)$:

Teacher model — Model for Gθ

$$\min_{\theta} \mathbb{E}_{x_t,t,x_0,x_T}\left[\lambda(t)\|\boxed{\widehat{x}_0^{*}(x_t,t,x_T)} - \boxed{\widehat{x}_0^{\theta}(x_t,t,x_T)}\|^2\right],$$
$$\text{s.t.} \quad \boxed{\widehat{x}_0^{\theta} = \arg\min_{\widehat{x}_0} \mathbb{E}_{x_t,t,x_0,x_T}\left[\lambda(t)\|\widehat{x}_0(x_t,t,x_T) - x_0\|^2\right],}$$
$$\boxed{(x_0,x_T) \sim p_\theta(x_0,x_T)},\ t \sim U([0,T]),\ x_t \sim q(x_t|x_0,x_T),$$

Data from G_θ

Intractable gradients problem since the objective includes term with argminimum.

Novel Distillation Loss

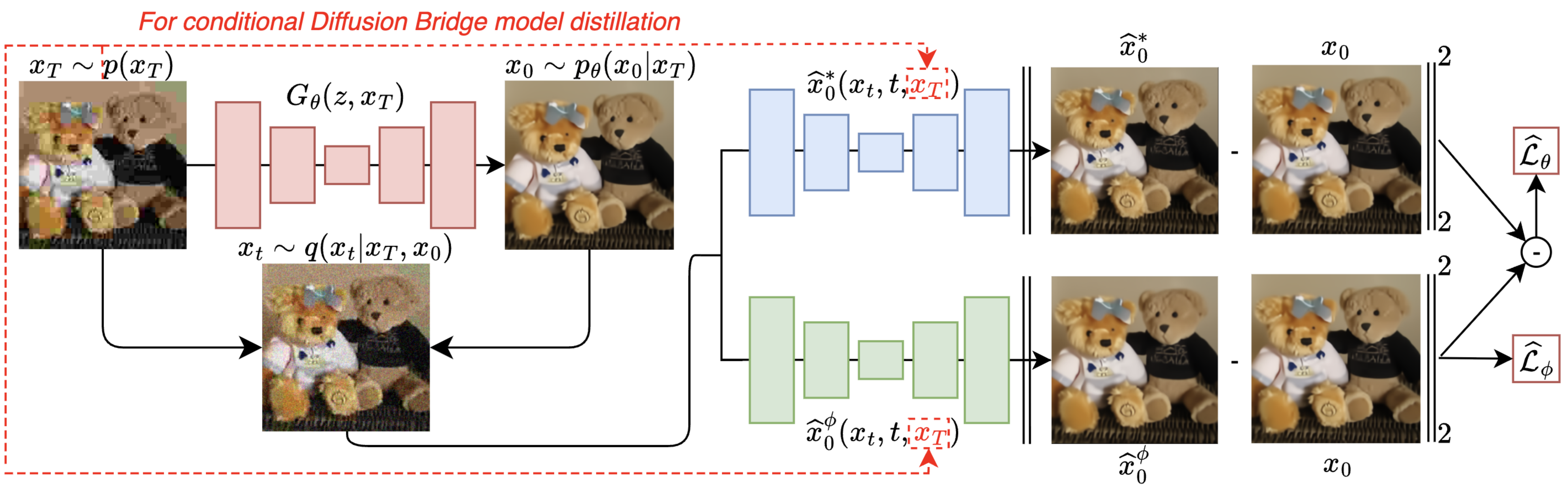
Solution: we show that the original constrained problem can be reformulated in the unconstrained one and used in practice:

$$\min_{\theta}\Big[\mathbb{E}_{x_t,t,x_0,x_T}\left[\lambda(t)\|\boxed{\widehat{x}_0^{*}(x_t,t,x_T)} - x_0\|^2\right] - \quad \text{Teacher model}$$
$$\min_{\phi} \mathbb{E}_{x_t,t,x_0,x_T}\left[\lambda(t)\|\boxed{\widehat{x}_0^{\phi}(x_t,t,x_T)} - x_0\|^2\right]\Big], \quad \text{Model for G}\theta$$
$$(x_0,x_T) \sim \boxed{p_\theta(x_0,x_T)},\ t \sim U([0,T]),\ x_t \sim q(x_t|x_0,x_T).$$

Data from G_θ

Tractable gradients since the objective does not include terms with argminimum.

Training pipeline for the inverse bridge matching distillation (IBMD)



IV Results (Up to 100x Sampling Acceleration)

We show the applicability of our IBMD distillation method for both *unconditional* and *conditional* diffusion bridge models.

We evaluate IBMD on diverse tasks, including super-resolution, JPEG restoration, inpaiting, and sketch-to-image translation.

In all settings, IBMD achieves significant inference speedups (up to 100×) and sometimes improves generation quality over the teacher.

Table 1. Results on the image super-resolution task. Baseline results are taken from I²SB (Liu et al., 2023a).

4× super-resolution (bicubic)	ImageNet (256 × 256)		
	NFE	FID ↓	CA ↑
DDRM (Kawar et al., 2022)	20	21.3	63.2
DDNM (Wang et al., 2023)	100	13.6	65.5
ΠGDM (Song et al., 2023)	100	3.6	72.1
ADM (Dhariwal & Nichol, 2021)	1000	14.8	66.7
CDSB (Shi et al., 2022)	50	13.6	61.0
I²SB (Liu et al., 2023a)	1000	2.8	70.7
IBMD-I²SB (**Ours**)	1	**2.6**	72.3

Table 2. Results on the image JPEG restoration task with QF=5. Baseline results are taken from I²SB (Liu et al., 2023a).

JPEG restoration, QF= 5.	ImageNet (256 × 256)		
	NFE	FID ↓	CA ↑
DDRM (Kawar et al., 2022)	20	28.2	53.9
ΠGDM (Song et al., 2023)	100	8.6	64.1
Palette (Saharia et al., 2022)	1000	8.3	64.2
CDSB (Shi et al., 2022)	50	38.7	45.7
I²SB (Liu et al., 2023a)	1000	**4.6**	67.9
I²SB (Liu et al., 2023a)	100	5.4	67.5
IBMD-I²SB (**Ours**)	1	5.2	66.6

Table 3. Results on the image super-resolution task. Baseline results are taken from I²SB (Liu et al., 2023a).

4× super-resolution (pool)	ImageNet (256 × 256)		
	NFE	FID ↓	CA ↑
DDRM (Kawar et al., 2022)	20	14.8	64.6
DDNM (Wang et al., 2023)	100	9.9	67.1
ΠGDM (Song et al., 2023)	100	3.8	72.3
ADM (Dhariwal & Nichol, 2021)	1000	3.1	73.4
CDSB (Shi et al., 2022)	50	13.0	61.3
I²SB (Liu et al., 2023a)	1000	2.7	71.0
IBMD-I²SB (**Ours**)	1	**2.5**	72.5

Table 4. Results on the image JPEG restoration task with QF=10. Baseline results are taken from I²SB (Liu et al., 2023a).

JPEG restoration, QF= 10.	ImageNet (256 × 256)		
	NFE	FID ↓	CA ↑
DDRM (Kawar et al., 2022)	20	16.7	64.7
ΠGDM (Song et al., 2023)	100	6.0	71.0
Palette (Saharia et al., 2022)	1000	5.4	70.7
CDSB (Shi et al., 2022)	50	18.6	60.0
I²SB (Liu et al., 2023a)	1000	**3.6**	72.1
I²SB (Liu et al., 2023a)	100	4.4	71.6
IBMD-I²SB (**Ours**)	1	3.7	72.4

Table 6. Results on the Image Inpainting Task. Methods are grouped by NFE (> 4, 4, 2, 1), with the best metrics bolded in each group. Baselines results are taken from CDBM.

Inpainting, Center (128 × 128)	ImageNet (256 × 256)		
	NFE	FID ↓	CA ↑
DDRM (Kawar et al., 2022)	20	24.4	62.1
ΠGDM (Song et al., 2023)	100	7.3	**72.6**
DDNM (Wang et al., 2023)	100	15.1	55.9
Palette (Saharia et al., 2022)	1000	6.1	63.0
I²SB (Liu et al., 2023a)	10	5.4	65.97
DBIM (Zheng et al., 2024)	50	3.92	72.4
DBIM (Zheng et al., 2024)	100	**3.88**	**72.6**
CBD (He et al., 2024)	4	5.34	69.6
CBT (He et al., 2024)		4.77	70.3
IBMD-I²SB (**Ours**)		5.1	70.3
IBMD-DDBM (**Ours**)		**4.03**	**72.2**
CBD (He et al., 2024)	2	5.65	69.6
CBT (He et al., 2024)		5.34	69.8
IBMD-I²SB (**Ours**)		5.3	65.7
IBMD-DDBM (**Ours**)		**4.23**	**72.3**
IBMD-I²SB (**Ours**)	1	6.7	65.0
IBMD-DDBM (**Ours**)		**5.87**	**70.6**

Table 5. Results on the Image-to-Image Translation Task (Training Sets). Methods are grouped by NFE (> 2, 2, 1), with the best metrics bolded in each group. Baselines results are taken from CDBM.

	NFE	Edges → Handbags (64 × 64)		DIODE-Outdoor (256 × 256)	
		FID ↓	IS ↑	FID ↓	IS ↑
DDIB (Su et al., 2023)	≥ 40	186.84	2.04	242.3	4.22
SDEdit (Meng et al., 2022)	≥ 40	26.5	3.58	31.14	5.70
Rectified Flow (Liu et al., 2022a)	≥ 40	25.3	2.80	77.18	5.87
I²SB (Liu et al., 2023a)	≥ 40	7.43	3.40	9.34	5.77
DBIM (Zheng et al., 2024)	50	1.14	**3.62**	3.20	**6.08**
DBIM (Zheng et al., 2024)	100	**0.89**	**3.62**	**2.57**	6.06
CBD (He et al., 2024)	2	1.30	3.62	3.66	6.02
CBT (He et al., 2024)		0.80	3.65	**2.93**	**6.06**
IBMD-DDBM (**Ours**)		**0.67**	**3.69**	3.12	5.92
Pix2Pix (Isola et al., 2017)	1	74.8	**4.24**	82.4	4.22
IBMD-DDBM (**Ours**)		**1.26**	3.66	**4.07**	**5.89**